ОГЛАВЛЕНИЕ

Часть 1. Тематический выпуск по материалам XXVI Всероссийской научной конференции «НАУЧНЫЙ СЕРВИС В СЕТИ ИНТЕРНЕТ», I

Часть 2. Оригинальные статьи

УДК 004.738.52

МЕТОД АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ПОЛНОТЕКСТОВЫХ ОПИСАНИЙ КЕРНОВ С ИСПОЛЬЗОВАНИЕМ СЛОВАРЕЙ

А. П. Антонов[1] [0009-0007-3642-7734], С. А. Афонин[2] [0000-0003-3058-9269],
А. С. Козицын[3] [0000-0002-8065-9061], В. М. Староверов[4] [0000-0001-8289-2273]

[1, 4]*Московский государственный университет им. М. В. Ломоносова, г. Москва, Россия*

[2, 3]*Научно-исследовательский институт механики МГУ им. М. В. Ломоносова, г. Москва, Россия*

[1]alexey.p.antonov@gmail.com, [2]serg@msu.ru, [3]alexanderkz@mail.ru, [4]staroverovvl@yandex.ru

Аннотация

Использование методов автоматической обработки текстов, в том числе методов классификации полнотекстовых описаний, позволяет достичь существенного снижения трудозатрат при обработке экспериментальных данных. В настоящей работе рассмотрено применение метода автоматической классификации текстов в области обработки и классификации элементов керна и определения литофаций. Литофациями называют одновозрастные геологические тела (отложения), которые по своему составу или строению отличаются от соседних слоев.

При проведении оценки нефтегазового потенциала месторождений требуется выполнять построение карт и схем распространения литофаций. Для этого необходимо осуществить классификацию большого количества полнотекстовых описаний участков керна, выполненных специалистами. Алгоритм, представленный в статье, позволяет на основе заданных правил и словарей провести классификацию с учетом порядка и значимости ключевых слов в предложениях. Преимуществами такого подхода являются возможность различать близкие литофации, возможность использования архивных данных, простота настройки на новые классы, адаптация к русскоязычным описаниям кернов и возможность локального использования без необходимости передавать описания кернов сторонним приложениям.

***Ключевые слова**: классификация текстов, литофации, словари, информационные системы.*

ВВЕДЕНИЕ

Опыт крупных мировых компаний показывает, что для увеличения эффективности разведки и разработки нефтяных и газовых месторождений необходимо внедрять в производственный процесс методы машинного обучения, в том числе при разведке и оценке месторождений [1]. Одной из важных задач при проведении оценки нефтегазового потенциала месторождений является построение карт и схем распространения литофаций [2]. Литофациями называют одновозрастные геологические тела (отложения), которые по своему составу или строению отличаются от соседних слоев. Классификация литофаций является двухуревневой. На первом уровне определяется название фации («фация морен», «фация аллювиальных конусов выноса», «фация речной поймы» и др.), на втором уровне – компонента лито- («алевро-глинистая», «углисто-алевролитовая», «битуминозно-кремнисто-глинистая» и др.). Одним из часто используемых методов построения карт распространения литофаций является анализ результатов бурения. Полученные в процессе бурения керны исследуются специалистами и описываются в свободном текстовом формате с разделением всей глубины керна на отдельные однородные участки. Результатом такого анализа является набор данных, включающий координаты и параметры скважины, глубину и полнотекстовое описание состава породы, полученной в результате бурения пробы. На основе выполненного текстового описания (литологического описания керна, содержащего, обычно, от одного до десяти предложений) необходимо сопоставить каждому участку керна один из заданных классов литофаций. Следует отметить, что в разных исследовательских группах и организациях существуют различные стандарты на классификаторы литофаций. В зависимости от предъявляемых требований и поставленных задач могут использоваться классификаторы, содержащих от 5–8 классов до сотни классов.

Автоматизация процесса проведения такой классификации позволяет значительно упростить и унифицировать обработку полнотекстовой информации, полученной от различных специалистов по десяткам тысяч кернов за последние

десятилетия, при анализе описаний кернов со скважин исследуемого региона. Возможные ошибки в работе впоследствии могут корректироваться при построении карт за счет сглаживания данных на соседних участках.

ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

Использование методов машинного обучения и автоматизация процессов определения литофаций керна в настоящий момент развиваются по трем основным направлениям. Наибольшее количество работ посвящено автоматизации распознавании изображений шлифов. Здесь следует отметить такие инструментальные средства, как ИС ABAI [3], сервис DeepCore копании Digital Petrelium [4], комплекс DHD [5], программный комплекс «Цифровой керн» [6] и «Нейросетевое распознавание текстурных особенностей графических керновых данных» [7].

В составе ИС ABAI (Advanced Base Artificial Intelligence) модуль «Автоматическая интерпретация керна» позволяет проводить автоматическое определение литофации по фотографии керна с использованием методов распознавания изображения. Обучение системы проводилось по 1300 скважинам. Классификатор содержит пять типов фаций и отдельный шестой тип для неопределенных изображений, может обрабатывать как фотографии, так и видео. Разрабатывается и эксплуатируется казахской компанией «КазМунайГаз».

Сервис DeepCore компании ООО «Диджитал Петролеум» производит обработку изображений кернов с использованием сверточных нейронных сетей. Обученная модель производит классификацию изображений кернов по 14 классам с точностью 70%. Сервис не заменяет полностью эксперта-геолога, однако, согласно данным, приведенным в статье [4], использование сервиса позволяет ускорить работу по описанию литофаций керна в 7 раз.

В программном комплексе DHD [5] компании Шлюмберже реализован модуль цифрового анализа керна (ЦАК), который по данным рентгеновских микрофотографий строит трехмерную цифровую модель керна с возможностью оценки ее физических свойств и сегментации.

Программный модуль «Цифровой керн» [6] компании Норникель позволяет на основе изображения керна в видимом диапазоне описывать характеристики керна и проводить классификацию участков керна по процентному содержанию сульфидов в исследуемом образце. Это позволяет в режиме online находить

и анализировать наличие рудного материала в керне по фотографии и с высокой вероятностью определять процент рудной минерализации.

Подобные решения позволяют существенно сократить трудозатраты при анализе данных бурения, поскольку не требуют или значительно уменьшают ручную работу специалистов по составлению описания керна. Однако такая технология неприменима для уже накопленного экспериментального материала. Кроме того, количество выделяемых классов фиксировано для каждой системы и оказывается существенно меньшим, чем при традиционных методах обработки текстовых данных.

В ряде работ предложено проводить классификацию литофаций по числовым характеристикам и измеряемым физическим свойствам пород, например, в работе [8] сравнены результаты классификации пород на 4 класса в соответствии со значением функций давления и внутреннего трения между частицами породы. Подобные методы также неприменимы для распределения по большим классификаторам, содержащим десятки классов.

В работе [9] исследованы методы определения физических характеристик керна на основе изображений, в том числе с использованием компьютерной томографии, что позволяет упростить получение данных о характеристиках кернового материала, а также дополнить результаты лабораторных и натурных исследований свойств пластов.

В ряде работ для распознавания и классификации полнотекстовых описаний предложено использовать современные системы с искусственным интеллектом. Например, в работе [10] описан метод классификации описаний с использованием векторных моделей текстов и нейронных сетей. Для проведения анализа полнотекстовых описаний кернов авторы преобразуют текст в векторное описание в модели GeoVec [11]. Эта модель построена на основе модели GloVe[12], обученной на текстах 280 тыс. англоязычных статей по геологии, доступных для скачивания через Elsevier ScienceDirect APIs, и отобранных вручную страницах Википедии "List_of_rock_types", "List_of_minerals", "List_of_landforms", "Rock_(geology)", "USDA_soil_taxonomy", "FAO_soil_classification" и др. Обученная модель позволяет автоматически определять близость отношений геологических

понятий, например, оказывается близкой векторная разность таких пар, как «гранит» – «магматический», «гнейс» – «метаморфический», «известняк» – «осадочный», «туф» – «вулканический», или пар «песок» – «песчаник», «гравий» – «конгломерат». На основе построенной модели авторы вычисляют для каждого предложения усреднение вектора его слов в модели GeoVec, которые подают на вход обученной нейронной сети. Классификация текстовых описаний кернов производилась по 18-ти классам литофаций.

Основным недостатком подобного подхода является отсутствие учета порядка слов в предложении. Поскольку для обучения и анализа в качестве входных данных используется усредненный вектор слов, элементы описания «глины с вкраплениями песка» и «песок с вкраплением глин» становятся неразличимыми. Соответственно, такой подход неприменим для проведения анализа с целью распределения текстов по детализированным классификаторам литофаций, включающим такие классы, как «глинисто-песчанные» и «песчанно-глинистые». Дополнительным ограничением применения машинного обучения на основе векторных моделей представления текстов является отсутствие предобученных моделей на русском языке, аналогичных GeoVec. Использование пословного перевода обученных моделей не дает удовлетворительного результата ввиду многозначности значительной части слов, используемых в английском и русском языках.

В работе [13] рассмотрена задача классификации полнотекстовых описаний кернов по 9 классам с использованием подходов, основных на сверточных нейронных сетях для классификации текста (TextCNN), сетей двунаправленной длительной-кратковременной памяти (BiLSTM) и сетей представлений двунаправленного кодера (BERT). Процент правильного распознавания существенно зависит от класса. Для трех классов вероятность правильного распознавания составила 99%, для оставшихся шести классов – от 24% до 32%. Для токенизации была использована модель RuBERT, обученная на русскоязычном варианте Википедии и новостных лентах, поскольку специализированные геологические модели для русского языка отсутствуют.

В этой связи для построения систем более точной классификации необходимо использование моделей и алгоритмов, учитывающих порядок слов в предложениях и адаптированных к русскому языку. Для этого можно применить методы, которые используются в классических задачах тематического анализа по ключевым словам [14, 15].

ОПИСАНИЕ АЛГОРИТМА

Разработанный нами алгоритм опирается на использование словарей, составленных геологами, с описанием характеристик различных литофаций. На вход алгоритму поступает полнотекстовое описание участка керна, выполненное специалистом при анализе результатов бурения. Результатом работы алгоритма является ранжированный список возможных литофаций, которые соответствуют заданному текстовому описанию. Полнотекстовое описание керна дается в свободном формате, но, как правило, содержит строгие формулировки. Например, «Переслаивание мощных пачек чередования хорошо сортированных песчаников, алевролитов с преимущественно песчаными прослоями в верхней части разреза и алевролитовыми – в нижней. Текстура пород преимущественно линзовидная, волнистослоистая».

Настройка алгоритма на специфику предметной области производится при помощи формирования словаря описаний фаций на основе возможных признаков фации, которые должны встречаться в описании (характерные для данной фации) или, наоборот, не могут встречаться в ее описании. Например, в описании фации русла рек не могут встречаться морские организмы, а в описании морен не может встречаться текстура воздушной ряби и орагнические включения в виде кораллов и трилобитов. Каждый признак является словом или словосочетанием и относится к определенному типу. В текущей программной реализации были рассмотрены следующие характеристики: название породы (например, «песчаник», «глина», «аргиллит», «песок»), ее цвет (например, «серого», «бежевого», «бурого», «рыжего»), структура (например, «псаммитовая», «алевритовая», «мелкозернистые»), текстура (например, «горизонтальная», «линзовидная», «массивная»), включения флоры и фауны (например, «трилобиты», «кораллы»,

«криноидеи», «радиолярии»), окатность (например, «неокатанные», «угловатые», «плохо окатанные»), сортировка (например, «сортировка плохая», «сортировка средняя»), границы (например, «волнистые», «ровные», «нечеткие»). При необходимости список анализируемых характеристик может пополняться. В качестве словосочетаний рекомендуется использовать пары вида «существительное прилагательное» или «наречие причастие», например, «граница ровная» или «хорошо окатанные». Такой подход позволяет получить достаточно точные формализованные критерии, которыми пользуются геологи при решении аналогичной задачи в ручном режиме.

Для учета специфики формирования полнотекстовых описаний используют вспомогательные словари, которые не несут в себе информации о предметной области, но позволяют правильно расставлять акценты и определять значимость ключевых элементов описания. Словарь ослабляющих слов и выражений с глаголами (например, «изредка встречаются») определяет слова и выражения, после которых значимость всех ключевых терминов уменьшается до конца предложения или глагола. Словарь усиливающих слов и выражений с глаголами (например, «основной», «превалирует», «обильно», «часто», «значительно», «многочисленный») позволяет задавать усиление значимости всех ключевых терминов до конца предложения или глагола. Кроме того, используется словарь глаголов-исключений, которые не прерывают действие ослабляющих и повышающих слов (например, «обладать», «содержать», «содержаться», «слагать»), словари синонимов (например, «аргиллит – глина – глинистые») и гиперонимов (например, «органогенный детрит: шлам, раковинный детрит, обломки раковин»). Следует отметить различие в обработке синонимов и гиперонимов. При сравнениях все синонимы считаются совпадающими терминами с учетом транзитивных зависимостей. Гипонимы и гиперонимы при обработке текстовых описаний считаются совпадающими терминами, но без учета транзитивных зависимостей между терминами. Словарь «лито» содержит возможные определения для каждой составляющей возможных описаний каждой компоненты лито (например, «глинистая, глина, глинисто, аргиллит, аргиллитовый, аргиллитовая»).

Разработанный алгоритм состоит из четырех этапов. Схема алгоритма представлена на рис. 1.

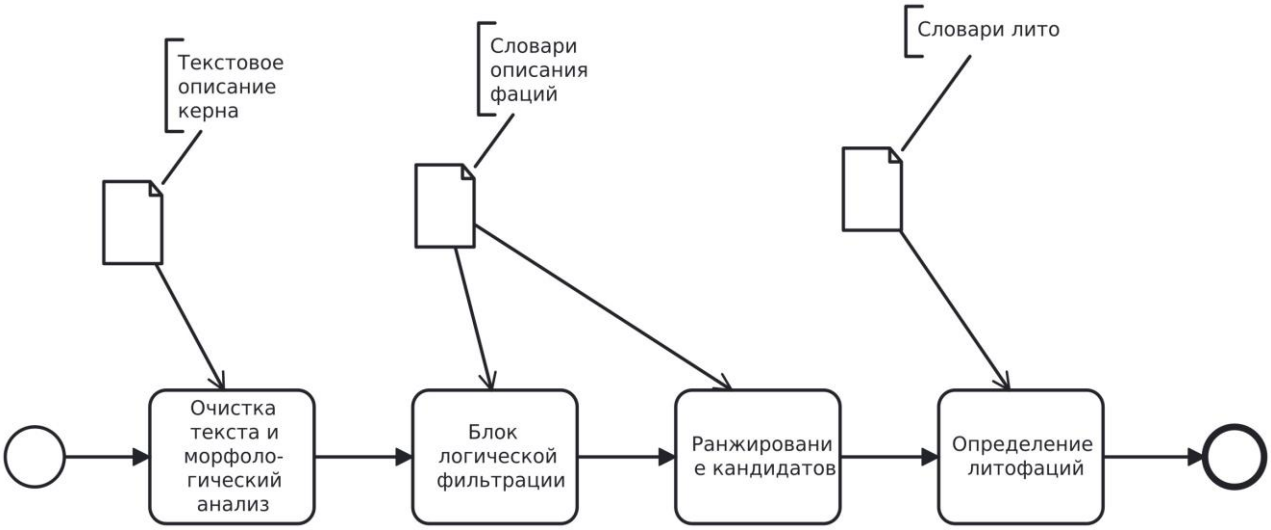


Рис. 1. Схема работы алгоритма классификации

На первом этапе работы алгоритма производятся подготовка текста для анализа, включая очистку текста от лишних символов, разметка предложений, разметка областей, заключенных в скобочки, а также проводится морфологический анализ. Для морфологического анализа используется свободно распространяемая библиотека pymorphy3 для языка Python, которая позволяет получать по заданной словоформе возможные варианты нормальной формы слова. В случае наличия нескольких вариантов алгоритм рассматривает все возможные случаи и для каждого ищет совпадения в используемых словарях. В процессе анализа для каждого найденного термина (слова или словосочетания) указывается его относительная позиция в тексте, принадлежность словарям характеристик фаций, наличие повышения или понижения его значимости в соответствии с наличием повышающих («преобладает», «преимущественно») и понижающих («иногда бывать», «изредка», «иногда», «прослои») слов. После повышающего или ослабляющего выражения до конца предложения, другого повышающего слова или глагола не из списка глаголов исключений все термины помечаются соответственно повышенными или ослабленными. Такой подход позволяет корректно обрабатывать описания вида «ниже залегает толща доломиты серых с прослоями мергелей и ангидритов». Для данного случая будет отмечено, что основной породой являются доломиты, а мергели и ангидриты имеют второстепенное значение, поскольку встречаются только в виде прослоек. Результатом работы первого этапа

алгоритма является структурированное описание анализируемого участка керна с разметкой позиций и значимости терминов.

В блоке логической фильтрации осуществляется фильтрация фаций-кандидатов по запрещающим правилам. Для каждой фации в словарях указывается, каких терминов из различных типов описаний в них не может встречаться. В случае, если хотя бы один из запрещенных для фации терминов встретился в анализируемом описании керна, указанная фация исключается из списка возможных кандидатов. Кроме того, фация удаляется из списка кандидатов, если в описании керна не встретилось ни одного термина из какого-либо значимого (непустого и с коэффициентом значимости «1») описания данной фации. При анализе встречаемости терминов в описании фаций используются словари синонимов («песок, песчанник») и гипонимов («красный: светло-красный, темно-красный, алый»). Результатом работы второго этапа алгоритма является список фаций, характеристики которых не противоречат анализируемому описанию участка керна.

В блоке ранжирования производится анализ списка всех терминов, найденных в описании керна, отдельно для каждого типа характеристики. Итоговый ранг фации вычисляется как сумма баллов, набранных фацией за соответствие каждого типа характеристики рассматриваемому описанию керна по следующей формуле:

$$css \sum_{i \in H} \sum_{w \in W} \begin{cases} dsk_i ss_{p_i} \cdot t_{iw}, & w \in D_i, \\ -k_2 \cdot dsk_i \cdot ssn_{p_i} \cdot t_{iw}, & w \notin D_i, \end{cases}$$

где CSS– нормирующий коэффициент фации, зависящий от количества описанных типов характеристик для рассматриваемой фации, dsk_i– коэффициент значимости типа признака *i* в расчете ранга, ss_j – коэффициент учета в ранге найденных в описании слов из признака со значимостью *j*, ssn_j – коэффициент учета в ранге не найденных в описании слов из признака со значимостью *j*, p_i – заданная экспертом значимость признака *i* для рассматриваемой фации (принимает значение 1, 2, 3 (1 – наиболее важное, 3 – наименее важное)); k_2 – размер штрафа за ненайденное слово, $t_{iw} = clf_i \cdot t_{wi}(w)$ – вес термина w_i; clf_i – способ учета длины описания типа признака *i* в рассматриваемой фации, в зависимости от значения параметра обучения либо 1, либо $1/n_i$; n_i – количество терминов в описании типа

признака *i* для фации, $t_{wi}(w)$ – функция учета idf для слова w, H – множество признаков фации, W – множество слов в описании, D_i – описание признака фации.

Конкретные значения коэффициентов определяются на этапе обучения модели. Результатом работы третьего этапа алгоритма является ранжированный список наиболее вероятных фаций, соответвующих текстовому описанию участка керна.

На последнем этапе для каждой найденной фации выбирается название лито из списка возможных значений для каждой фации, заданного экспертом в словаре. Например, для фации кам возможными лито будут «глинисто-песчаная» и «песчаная», для фации аллювиальных конусов выноса возможными фациями будут «алевро-песчаная», «мелко-грубообломочная» и «грубо-мелкообломочная».

Основным принципом построения названия лито является повышение приоритета слова от начала к концу термина. Например, название лито «углисто-алевро-глинистая» означает, что его основу составляют глины, в меньшей степени встречаются алевролитовые слои и в совсем небольшом количестве в образце присутствуют угольные вкрапления. Для поиска подходящего названия лито из описания керна выделяются все термины, сортируются с учетом позиции текста и значимости (значимые передвигаются вперед, незначимые – назад), после этого все термины с дефисом переворачиваются. Для каждого возможного лито определяется порядок встречаемости его терминов в получившемся списке. Описания лито, термины которых не встретились в списке или встретились не в том порядке, исключаются из рассмотрения. Среди оставшихся названий лито выделяется название, термины которого оказались ближе к началу списка. Результатом работы данного шага алгоритма является наиболее вероятное название лито для каждой рассматриваемой фации. Дополнительно в случае нахождения единственного возможного названия лито или нескольких возможных вариантов лито к рангу фации, вычисленному на предыдущем шаге, добавляется дополнительный коэффициент, поднимающий ее выше в ранжированном списке возможных вариантов фаций.

Результатом работы алгоритма является ранжированный список фаций с указанием лито для каждой фации. Этот список может использоваться для автоматической или автоматизированной классификации. В случае автоматической обработки описанию керна сопоставляется литофация с наибольшим рангом. В случае автоматизированной работы пользователю вместе с описанием предлагается выбор из нескольких фаций, с наиболее высоким рангом, что позволяет значительно сократить время поиска среди большого списка позиций.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И РЕЗУЛЬТАТЫ

Программная реализация алгоритма выполнена на языке Python 3.11. Морфологический анализ осуществляется с использованием библиотеки pymorphy3. Для работы также используются библиотеки dataclass, math и numpy. Для описания словарей использовались файлы в текстовом формате. Такой подход позволяет легко редактировать словари в любом текстовом редакторе, например, создавать описания новых фаций или корректировать существующие. Для ускорения работы по исходным словарям в процессе работы в автоматическом режиме строятся дополнительные индексные файлы.

Процесс обучения разработанной программной реализации алгоритма заключался в подборе значений параметров алгоритма, указанных в табл. 1. Обучение проводилось на основе обучающей выборки из 66 примеров, для каждого из которых экспертом был выбран правильный вариант ответа. Подбор параметров производился с использованием метода градиентного спуска. Метрика качества L для проведения обучения рассчитывалась по следующей формуле

$$L = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k_i},$$

где N – количество тестовых примеров, k_i – порядковый номер правильного ответа в полученном ранжированном списке для примера i.

В случае наличия в ранжированном списке нескольких примеров с одинаковым рангом порядковый номер для них усредняется. Например, для ранжированного (с указанием ранга) списка A(5.2), B(4.7), C(4.7), D(4.7)), E(1.6) при правильном ответе B значение k_i будет равняться 3 (средняя позиция B, C и D). В результате обучения на обучающей выборке было достигнуто значение метрики качества L = 0.681.

Табл. 1. Параметры обучения алгоритма.

Параметр	Возможные значения	Оптимальное значение
Учитывать ли длину описания фации	0 – нет, 1 – да	0
Нормирующий коэффициент фации, зависящий от количества описанных типов характеристик для рассматриваемой фации	0 – не учитывается, 1 – учитывается линейно; иначе с заданным основанием логарифма (например, 10)	10
Коэффициент учета в ранге найденных в описании слов признака со значимостью 1	Число	1
Коэффициент учета в ранге найденных в описании слов признака со значимостью 2	Число	0.9
Коэффициент учета в ранге найденных в описании слов признака со значимостью 3	Число	0.8
Коэффициент учета в ранге не найденных в описании слов со значимостью 1	Число	1
Коэффициент учета в ранге не найденных в описании слов со значимостью 2	Число	2/3
Коэффициент учета в ранге не найденных в описании слов со значимостью 3	Число	4/9
Коэффициент понижения значимости для «незначимых» слов	Число	0

Размер штрафа за ненайденное слово	Число	1
Тип учета idf	Целое число: 0 – не учитывается, 1 – учитывается линейно, иначе учитывается как логарифм с указанным основанием	0
Коэффициент значимости типа признака в расчете ранга	Список значений коэффициентов для каждого типа признака	{"rock":2, "Color":0.6, "structures":1, "texture":1, "inclusion":1, "inclusionorg":1, "roundness":1, "sort":1, "border":1}
Добавочный коэффициент за единственное найденное название лито	Число	3
Добавочный коэффициент за несколько возможных найденное названий лито	Число	0.5

Полученные в результате процесса обучения значения параметров использовались для проведения тестирования результатов работы на тестовой выборке описаний кернов.

Тестирование проводилось на 58 примерах. Результаты тестирования: L = 0.576, на первом месте нужная фация при выборе из 42 фаций оказалась в 25 примерах, на втором – в 7 примерах, на третьем – в 8 примерах. Компонента лито была определена правильно в 46 примерах.

ЗАКЛЮЧЕНИЕ

Представленный алгоритм классификации полнотекстовых описаний кернов может использоваться для автоматизации процесса определения классов литофаций при построении литофационных карт, в том числе в разрабатываемых в настоящее время системах, которые должны заменить Petromod в национальных корпорациях. При обработке специалистом описаний кернов алгоритм подбирает наиболее вероятные классы, сокращая время разметки исходного материала. Преимуществами алгоритма являются возможность обработки архивных данных и данных сторонних исследований, адаптация к русскому языку, возможность локального использования, а также возможность учета порядка слов в описаниях.

СПИСОК ЛИТЕРАТУРЫ

1. Искусственный интеллект в нефтегазовой индустрии Китая. URL: https://nntc.pro/tpost/h2hoet4se1-iskusstvennii-intellekt-v-neftegazovoi-i (дата обращения: 11.12.2025)

2. *Антонов А.П., Афонин С.А., Козицын А.С. и др*. Автоматизированное построение реалистичных литофациальных карт методами комбинаторной оптимизации // Интеллектуальные системы. Теория и приложения. 2024. Т. 28, № 4. С. 5–20.

3. Информационная система ABAI. URL: https://kmge.kz/abai/ (дата обращения: 11.12.2025)

4. *Барабошкин Е.Е., Панченко Е.А., Демидов А.Е. и др.* Система автоматического описания керна в производственном процессе. Опыт применения // Пути реализации нефтегазового потенциала Западной Сибири: Материалы XXV научно-практической конференции, Ханты-Мансийск, 23–26 ноября 2021 года / Под редакцией Э.А. Вторушиной, Е.Е. Оксенойд, С.А. Алёшина, Н.Н. Захарченко, Е.В. Олейник, Т.Н. Печёрина. Ханты-Мансийск: Автономное учреждение Ханты-Мансийского автономного округа – Югры. Научно-аналитический центр рационального недропользования им. В.И. Шпильмана, 2022. С. 293–299.

5. Комплекс DHD. URL: https://magazine.neftegaz.ru/articles/tsifrovizatsiya/682038-tsifrovoy-analiz-

kerna-v-zadachakh-proektirovaniya-razrabotki-neftyanykh-i-gazovykh-mestorozhdeniy-/ (дата обращения: 11.12.2025)

6. Программный комплекс «Цифровой керн».
URL: https://globalcio.ru/projects/10448/ (дата обращения: 11.12.2025)

7. *Аристов А.И., Зеленин А.В., Катанов Ю.Е.* Нейросетевое распознавание текстурных особенностей графических керновых данных. Свидетельство о регистрации программы для ЭВМ RU 2024615647, 11.03.2024. Заявка № 2024614650 от 11.03.2024.

8. *Li H, Wan B, Chu D, Wang R, Ma G, Fu J, Xiao Z.* Progressive Geological Modeling and Uncertainty Analysis Using Machine Learning // ISPRS International Journal of Geo-Information. 2023. Vol. 12 (3). 97.
https://doi.org/10.3390/ijgi12030097

9. *Химуля В.В.* Применение технологии цифрового анализа керна для изучения фильтрационно-емкостных свойств и структуры высокопроницаемых пород подземных хранилищ газа // RJES. 2024. №5. С. 1–15.
URL: https://rjes.ru/temp/fddc89c0f81314f3d14bad3446565446.pdf (дата обращения: 11.12.2025).

10. *Fuentes I., Padarian J., Iwanaga T.,* Vervoort R.W. 3D lithological mapping of borehole descriptions using word embeddings // Computers & Geosciences. 2020. Vol. 141. 104516. https://doi.org/10.1016/j.cageo.2020.104516.
URL: https://www.sciencedirect.com/science/article/pii/S0098300419306533

11. *Padarian J., Fuentes I.* Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts // SOIL. 2019. Vol. 5. P. 177–187. https://doi.org/10.5194/soil-5-177-2019.
URL: https://soil.copernicus.org/articles/5/177/2019/

12. *Pennington J., Socher R., Manning C.* Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014. P. 1532–1543.

13. *Катанов Ю.Е., Аристов А.И., Ягафаров А.К., Новрузов О.Д.* Цифровой керн: нейросетевое распознавание текстовой геолого-геофизической информации // Известия высших учебных заведений. Нефть и газ. 2023. № 3 (159). С. 35–54.

14. *Денисов Д.В.* Анализ методов машинного обучения для тематической классификации текстов// Международный журнал информационных технологий и энергоэффективности. 2024. Т. 9, № 4 (42). С. 5–11.

15. *Козицын А.С.* Алгоритмы тематического поиска данных в наукометрических системах // Программная инженерия. 2022. Т. 13. № 6. С. 291–300.

METHOD FOR AUTOMATIC CLASSIFICATION OF FULL-TEXT DESCRIPTIONS OF CORES USING DICTIONARIES

A. P. Antonov[1] [0009-0007-3642-7734], S. A. Afonin[2] [0000-0003-3058-9269],
A. S. Kozitsin[3] [0000-0002-8065-9061], V. M. Staroverov[4] [0000-0001-8289-2273]

[1, 4]*Lomonosov Moscow State University, Moscow, Russia*

[2, 3]*Institute of Mechanics, Lomonosov Moscow State University, Moscow, Russia*

[1]alexey.p.antonov@gmail.com, [2]serg@msu.ru, [3]alexanderkz@mail.ru, [4]staroverovvl@yandex.ru

Abstract

The use of automatic text processing methods, including full-text description classification methods, allows achieving a significant reduction in labor costs when processing experimental data. This paper discusses the use of the automatic text classification method in the field of processing and classifying core elements and determining lithofacies. Lithofacies are coeval geological bodies (deposits) that differ in composition or structure from adjacent layers. When assessing the oil and gas potential of fields, it is necessary to construct maps and diagrams of lithofacies distribution. This requires classifying a large number of full-text descriptions of core sections prepared by specialists. The algorithm presented in the article allows, based on specified rules and dictionaries, to conduct classification taking into account the order and significance of keywords in sentences. The advantages of this approach are: the ability to distinguish between close lithofacies, the ability to use archival data, ease of adjustment to new classes, adaptation to Russian-language core descriptions and the possibility of local use without the need to transfer core descriptions to third-party applications.

Keywords*: text classification, lithofacies, dictionaries, information systems.*

REFERENCES

1. Iskusstvennyi intellekt v neftegazovoi industrii Kitaia. URL: https://nntc.pro/tpost/h2hoet4se1-iskusstvennii-intellekt-v-neftegazovoi-i

2. *Antonov A.P., Afonin S.A., Kozitsyn A.S. i dr.* Avtomatizirovannoe postroenie realistichnykh litofatsialnykh kart metodami kombinatornoi optimizatsii // Intellektualnye sistemy. Teoriia i prilozheniia. 2024. Vol. 28, № 4. S. 5–20.

3. Informatsionnaia sistema ABAI. URL: https://kmge.kz/abai/

4. *Baraboshkin E.E., Panchenko E.A., Demidov A.E. i dr.* Sistema avtomatich-eskogo opisaniia kerna v proizvodstvennom protsesse. Opyt primeneniia // Puti real-izatsii neftegazovogo potentsiala Zapadnoi Sibiri: Materialy XXV nauchno-prakticheskoi konferentsii, Khanty-Mansiisk, 23–26 noiabria 2021 goda / Pod redaktsiei E.A. Vtorushinoi, E.E. Oksenoid, S.A. Aleshina, N.N. Zakharchenko, E.V. Oleinik, T.N. Pecher-ina. Khanty-Mansiisk: Avtonomnoe uchrezhdenie Khanty-Mansiiskogo avtonomnogo okruga – Iugry "Nauchno-analiticheskii tsentr ratsionalnogo nedropolzovaniia im.V.I.Shpilmana", 2022. S. 293–299.

5. Kompleks DHD. URL: https://magazine.neftegaz.ru/articles/tsifrovizatsiya/682038-tsifrovoy-analiz-kerna-v-zadachakh-proektirovaniya-razrabotki-neftyanykh-i-gazovykh-mestorozh-deniy-/ (11.12.2025)

6. Programmnyi kompleks "Tsifrovoi kern". URL: https://globalcio.ru/projects/10448/

7. *Aristov A.I., Zelenin A.V., Katanov Iu.E.* Neirosetevoe raspoznavanie teksturnykh osobennostei graficheskikh kernovykh dannykh. Svidetelstvo o registratsii programmy dlia EVM RU 2024615647, 11.03.2024. Zaiavka № 2024614650 11.03.2024.

8. *Li H, Wan B, Chu D, Wang R, Ma G, Fu J, Xiao Z.* Progressive Geological Modeling and Uncertainty Analysis Using Machine Learning // ISPRS International Jour-nal of Geo-Information. 2023. Vol. 12(3). 97. https://doi.org/10.3390/ijgi12030097

9. *Khimulia V.V.* Primenenie tekhnologii tsifrovogo analiza kerna dlia izuche-niia filtratsionno-emkostnykh svoistv i struktury vysokopronitsaemykh porod pod-zemnykh khranilishch gaza // RJES. 2024. №5. S. 1–15. URL: https://rjes.ru/temp/fddc89c0f81314f3d14bad3446565446.pdf

10. *Fuentes I., Padarian J., Iwanaga T., Vervoort R.W.,* 3D Lithological mapping of borehole descriptions using word embeddings // Computers & Geosciences. 2020. Vol. 141. 104516. https://doi.org/10.1016/j.cageo.2020.104516
URL: https://www.sciencedirect.com/science/article/pii/S0098300419306533

11. *Padarian J., Fuentes I.* Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts // SOIL. 2019. Vol. 5. P. 177–187. https://doi.org/10.5194/soil-5-177-2019, 2019.
URL: https://soil.copernicus.org/articles/5/177/2019/

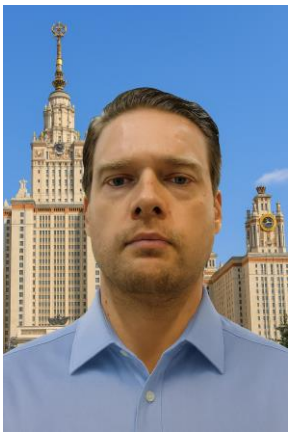
12. *Pennington J., Socher R., Manning C.* Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. P. 1532–1543.

13. *Katanov Iu.E., Aristov A.I., Iagafarov A.K., Novruzov O.D.* Tsifrovoi kern: neirosetevoe raspoznavanie tekstovoi geologo-geofizicheskoi informatsii // Izvestiia vysshikh uchebnykh zavedenii. Neft i gaz. 2023. № 3 (159). S. 35–54.

14. *Denisov D.V.* Analiz metodov mashinnogo obucheniia dlia tematicheskoi klassifikatsii tekstov // Mezhdunarodnyi zhurnal informatsionnykh tekhnologii i energoeffektivnosti. 2024. Vol. 9, № 4(42). S. 5–11.

15. *Kozitsyn A.S.* Algoritmy tematicheskogo poiska dannykh v naukometricheskikh sistemakh // Programmnaia inzheneriia. 2022. Vol. 13, № 6. S. 291–300.

СВЕДЕНИЯ ОБ АВТОРАХ



АНТОНОВ Алексей Петрович – доцент, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области гармонического анализа.

Alexey Petrovich ANTONOV – Associate Professor, Ph.D., graduated from Lomonosov Moscow State University. Specialist in the field of harmonic analysis.

email: alexey.p.antonov@gmail.com

ORCID:0009-0007-3642-7734

АФОНИН Сергей Александрович – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М. В. Ломоносова. Специалист в области регулярных языков и информационных систем.

Sergey Alexandrovich AFONIN – Leading Researcher, Ph.D., graduated from Lomonosov Moscow State University. Specialist in the field of regular languages and information systems.

email: serg@msu.ru

ORCID:0000-0003-3058-9269

КОЗИЦЫН Александр Сергеевич – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М. В. Ломоносова. Специалист в области информационного поиска и баз данных.

Alexander Sergeevich KOZITSYN – Leading Researcher, Ph.D., graduated from Lomonosov Moscow State University. Specialist in the field of information retrieval and database.

email: alexanderkz@mail.ru

ORCID: 0000-0002-8065-9061

СТАРОВЕРОВ Владимир Михайлович – доцент, к. ф.-м. н., окончил мехмат МГУ им. М. В. Ломоносова. Специалист в области геологического моделирования.

Vladimir Mikhailovich STAROVEROV – Associate Professor, Ph.D., graduated from Lomonosov Moscow State University. Specialist in the field of geological modelling.

email: staroverovvl @yandex.ru

ORCID: 0000-0001-8289-2273

Материал поступил в редакцию 18 декабря 2025 года

УДК 004.43(042.4)

ФОРМЫ ДЛЯ ПОКАЗА РЕЗУЛЬТАТОВ СРАВНЕНИЯ ЯЗЫКОВ ПРОГРАММИРОВАНИЯ НА ПРИМЕРЕ ДИАЛЕКТОВ ЯЗЫКА LISP

Л. В. Городняя[0000-0002-4639-9032]

Институт систем информатики им. А. П. Ершова СО РАН,
г. Новосибирск, Россия
Новосибирский государственный университет, г. Новосибирск, Россия

lidvas@gmail.com

Аннотация

Статья посвящена выработке форм для показа результатов анализа и сравнения особенностей языков, систем и парадигм программирования. Предлагаемая форма продемонстрирована на примере результатов сравнения языка Lisp, наиболее успешных его диалектов (Scheme, Common Lisp, Racket, Clojure) и парадигмы функционального программирования на разных уровнях определения языков и систем программирования. Форма позволила лаконично показать наследование ряда особенностей языка Lisp и их развитие в диалектах на уровне конкретного синтаксиса, абстрактной семантики и системной прагматики.

Ключевые слова*: язык программирования, Lisp, Scheme, Common Lisp, Racket, Clojure, функциональное программирование, сравнение языков программирования, конкретный синтаксис, абстрактная семантика, системная прагматика.*

ВВЕДЕНИЕ

Очередной этап разработки методики анализа и сравнения языков, систем и парадигм программирования потребовал специальных форм для лаконичного представления и показа результатов применения этой методики. Статья посвящена текущим исследованиям, продолжающимся в Лаборатории информационных систем Института систем информатики им. А. П. Ершова Сибирского отделения Российской академии наук (СО РАН) в рамках тематики, связанной с методами преподавания программирования, требующими для контроля успехов в обучении оценки продуктивности программирования и производительности

программ. Ранее была выработана визуально-табличная форма показа категорий семантических систем языка программирования, затрагивающая кроме абстрактной семантики механизмы системной прагматики [1]. Теперь начинается проверка разработанной парадигмально-семантической методики на конкретных долгоживущих языках программирования.

Изложение начинается с описания форм и обозначений для демонстрации разноуровневых различий между диалектами языка программирования. Затем дана краткая справка о языке Lisp и его диалектах Pure Lisp (1962), Scheme (1976), Common Lisp (1984), Racket (1994) и Clojure (2007) в порядке их появления. Далее сформулированы выводы о замеченных особенностях создания диалектов языка Lisp.

Гомоиконный конкретный синтаксис программы представляет программу в виде ее абстрактного синтаксического дерева (abstract syntax tree – AST), позволяющего применять автоматизированную генерацию распознавателей принадлежности программы языку программирования. Трансформационная абстрактная семантика отражает эквивалентность разных форм представления программ и данных, дающую основания для оптимизирующих преобразований программ. Приаппаратная системная прагматика вычислений подчинена требованиям эффективности и производительности кода программ, включая проблемы безопасности и надежности. Парадигма программирования отражает стиль мышления в процессе постановки задачи, способствующий продуктивному программированию ее решения. Взаимодействие синтаксиса, семантики, прагматики и парадигм можно рассматривать как логическую интерпретацию диалектных абстракций языка программирования, представление которой требует специальных форм, показывающих архитектуру языка.

Методика учитывает, что термин «язык программирования» в речевой практике понимается как «входной язык системы программирования, обеспечивающей доступ к определенным аппаратным средствам». Такое понимание потребовало специальных форм показа результатов анализа и сравнения, отражающих перемещение сквозных понятий на разные уровни реализации языка программирования. Поэтому, кроме сравнения конструкций языка программирования на уровне конкретного синтаксиса и абстрактной семантики, проанализированы структуры данных, пространства доступных процессов обработки данных

и дисциплины доступа к памяти в типовых языках программирования на уровне системной прагматики. Учтено, что понимание языка программирования всегда опирается на ряд известных, возможно неявных конструкций, необходимых для его реализации в системе программирования и воспринимаемых в практике как неотъемлемая часть языка, в реальности существующего как целостный комплекс, составляющие которого взаимосвязаны.

Долгоживущие языки программирования обычно расширяют ряд вычислительных возможностей и парадигм программирования подключением стандартных библиотек, пакетов, монад или выделением диалектов, повышающих продуктивность программирования. Диалект становится самостоятельным языком, наследуя особенности исходного языка программирования, слегка изменяя и дополняя их. Цель выполненного эксперимента – изучить особенности изменения конкретного синтаксиса, абстрактной семантики и системной прагматики языка программирования в диалектах и наследниках, показать особенности наследования конструкций уровня и синтаксиса, и семантики, и прагматики. При сравнении выделяются диалектные абстракции, работающие как метапонятия, смысл которых немного варьируется в диалектах при сохранении архитектуры языка и поддержанных им парадигм.

В статье приведены результаты сравнения языка Lisp с его успешными диалектами. Результаты представлены в форме, показывающей, что унаследовано, что отвергнуто, что изменено и чем дополнено. Выбор языка Lisp для первого эксперимента обусловлен не только четкостью и лаконизмом его описания [2], но и ростом интереса к функциональному программированию, регулярно происходящим при смене элементной базы и расширении сферы применения информационных технологий. Кроме того, Lisp можно характеризовать как язык программирования одновременно и низкоуровневый, и сверхвысокого уровня в зависимости от уровня решаемых задач. Lisp легко адаптируется к решению новых задач и изобретению лаконичных и эффективных конструкций, не всегда соответствующих шаблонам, навязываемым более популярными языками программирования. Такой спектр возможностей языка позволяет определять логическую интерпретацию сквозных понятий, показывающую общность решений, связанных с языком, – архитектуру языка.

Эксперимент выполнен на материале диалектов Pure Lisp, Scheme, Common Lisp, Racket и Clojure [2–8], появившихся с шагом в 10 лет. Анализ этих диалектов показал различие целей их создания и механизмов достижения целей при минимальных изменениях семантики и прагматики исходного языка Lisp. Более заметны изменения на уровне лексикона[1] и конкретного синтаксиса, что удобно для выделения диалектных абстракций. При сравнении языка Lisp с его диалектами уделено внимание своду принципов функционального программирования и расслоению языка программирования на базис, расширение, средства диагностики границ вычислимости и отладки программ, а также средства связи процесса вычислений с внешним миром.

Для оценки особенностей диалектов были использованы примеры реализации отдельных конструкций языка Lisp и его диалектов, проверенные на системе HomeLisp[2], платформе jdoodle.com[3] и других онлайн-компиляторах.

ФОРМЫ ПОКАЗА РЕЗУЛЬТАТОВ СРАВНЕНИЯ ЯЗЫКОВ ПРОГРАММИРОВАНИЯ

Традиция описывать языки программирования и представлять их формальные определения сложилась во времена, когда каждый язык реализовывали автономно, начиная с прагматики решений уровня аппаратуры, полной реализации анализаторов синтаксиса и семантики языка программирования, возможно с разработкой своего формализма, типа расширенных форм Бэкуса – Наура (БНФ). Такие формализмы не претендовали на описание контекстнозависимых особенностей абстрактной семантики и системной прагматики из-за чрезмерного разнообразия новых архитектур и развития методов реализации языков программирования.

С тех пор в практике реализации новых языков программирования сложилась тенденция ограничиваться синтаксической надстройкой над существующими языками без пересмотра или уточнения решений уровня прагматики, ограниченной RISC-архитектурой[4], с небольшими вариациями семантики, что

[1] Под термином «лексикон» понимается множество имен доступных функций.

[2] http://homelisp.ru/

[3] https://www.jdoodle.com/

[4] RISC-архитектура – reduced instruction set computer

означает выделение небольшого числа известных стереотипов в этой сфере, почти не подверженных вариациям. Это позволяет представлять наследование конструкций между языками и выполнять логическую интерпретацию различий в языке программирования в терминах диалектной абстракции.

Основные трудности представления логической интерпретации связаны с тем, что описания языков программирования и их стандартов обладают слишком большим объемом (от 700 до 1500 страниц). Формализмы, используемые в описаниях языка программирования, представляют собственно конкретный синтаксис языка с неформальными пояснениями без показа границ наследования конструкций предшествующих языков и решений уровня абстрактной семантики и системной прагматики, описываемых средствами естественного языка. Сведения, полученные из таких источников, требуют проверки на реальных языках программирования, подверженных развитию.

В порядке эксперимента для показа деталей наследования на уровне конкретного синтаксиса между диалектами языка программирования предложено использовать специальное расширение форм Бэкуса – Наура [9], отражающее отношение наследования между понятиями разных диалектов и их одноименными определениями в предшественниках, частично дополненное показом особенностей некоторых понятий на уровне абстрактной семантики и системной прагматики (табл. 1). Примеры такого представления конструкций «S-выражение», «Форма», «Функция», «ленивые вычисления», структуры данных и особенностей «REPL-цикла»[5], полученных при сравнении диалектов языка Lisp, приведены в препринтах [22, 23]. В эксперименте удалось показать с помощью таких обозначений некоторые особенности абстрактной семантики и системной прагматики с небольшими комментариями.

Для более полного показа лаконичной формы пока не нашлось, одни и те же конструкции могут перемещаться из прагматики в семантику, из семантики в синтаксис – граница между синтаксисом, семантикой и прагматикой условна.

[5] REPL-цикл – название основного рабочего цикла, определяющего обработку программ, сокращение от Read Eval Print Loop.

Табл. 1. Расширение БНФ для показа результатов сравнения языков программирования

Формула	Примечание
Диалект: понятие	Объявление понятия в диалекте
Старое_понятие**.**Предшественник	Используется определение из предшествующего языка
Старое_понятие**!~**Шаблон**.**Предшественник	Из определения предшествующего языка исключаются фрагменты, соответствующие шаблону
Одноименное_понятие	Используется новое определение, полностью замещающее старое
Элемент …	Произвольное число вхождений элемента, возможно ни одного
_* «Symbol» [понятие]	Последовательность литер, кроме Symbol – продолжение после ошибки
⊥! «строка-диагноз» **[** понятие **]**	Сообщение диагноза с приемом дополнения в диалоге
Синтаксис Семантика Прагматика // Комментарий	Уровни определения (табл. 2)
[[Формулы над множествами]] // скобки для наглядности перехода с семантике вычислений	Семантические системы. Семантика вычислений в конкретном пространстве (табл. 3)
((Операции над состояниями памяти)) // скобки для наглядности перехода к прагматике изменения памяти	Системная прагматика. Дисциплина изменения состояний памяти (табл. 4)

Кроме показа наследования такие формулы позволяют определять продолжающие и диагностические грамматики. Понятия разных уровней могут входить в общую формулу (табл. 2).

Табл. 2. Формы для показа границ между уровнями понятий языка программирования

Формула	*Примечание*
Элемент ::= { **Синтаксис** ***Семантика*** Прагматика *// Комментарий* }	*Разные шрифты*
Элемент ::= **{** **Синтаксис** ***Семантика*** Прагматика *// Комментарий* **}**	*Разные уровни*

Строки табл. 2 выражают разноуровневые составляющие определения языка программирования привлечением разных шрифтов для понятий уровня **синтаксиса**, ***семантики*** и прагматики. **Синтаксис** – жирный шрифт, ***семантика*** – жирный курсив, прагматика – обычный подчеркнутый шрифт, *комментарий* – курсив. Более наглядно использовать индексы ($^{\text{верхний}}$, обычный, $_{\text{нижний}}$). Примеры так представленных различий в уровнях и границах вхождения в языки программирования структур данных диалектов языка Lisp приведены в препринте [23].

Не все важные особенности языка программирования удалось выразить такими компактными формами. При поиске форм для показа результатов сравнения языков программирования, удобных для оценки выразительной силы языка программирования, а также трудоемкости и продуктивности его реализации, эффективности и производительности программ, создаваемых на базе языка программирования, учтена зависимость от последовательности критериев принятия решений по декомпозиции программ, что не является однозначным, зависит от парадигм программирования и классов решаемых задач. Для диалектов языка Lisp последовательность критериев зависит от принципов функционального программирования (таких как универсальность данных, самоприменимость определений, равноправие и независимость параметров и единственность результатов функций, гибкость границ блоков памяти и неизменяемость хранимых значений). Технически в качестве основного критерия выбрана

семантическая декомпозиция определений языков программирования, позволяющая показывать различия и дистанцию в понятийной сложности между похожими семантическими системами[6], образующими язык программирования. Для лаконичного показа различий абстрактной семантики и системной прагматики предложенные обозначения немного различаются для таких категорий семантических систем, как вычисления, структуры данных, управление вычислениями и обработка памяти – они обладают разными шаблонами определения функций.

Семантическая система – это тройка **[[V, F, R]]**, где:

V – основное множество данных, возможно бесконечное;

F – набор операций, возможно принадлежащих множеству **V,** расширяемый программируемыми функциями;

R – варианты правил применения операций **F** к данным из **V**, возможно входящих в **F**, представимые как данные из **V**, возможно программируемые как функции.

Компактная форма взаимосвязей составляющих семантических систем функционального программирования выражается формулой

$$[[V, F, R]] \mid R \subset F \subset V$$

где «**R** является подмножеством **F** и F является подмножеством **V**» или «**R** включено в **F**, а **F** включено в **V**».

Такой формат семантики языка программирования, присущий функциональному программированию, поддерживает передачу опыта программирования в форме диалектов и пакетов со своими правилами их интерпретации. В процессе программирования новых функций, расширяющих **F**, возможна разработка новых вариантов **R** и новых семантических систем. Такие взаимосвязи между понятиями позволяют формализовать и развивать правила **R** применения операций **F** к данным **V**, включая кумулятивные (накопительные) эффекты между составляющими системы программирования. Представления операций и программируемых функций **F** включаются в основное множество **V**, а правило применения **R** операций **F** к данным **V** – не более чем одна из функций, возможно

[6] С. С. Лавров предложил понятие «семантическая система» как расширение понятия «алгебраическая система» заданием явного правила R применения операций к данным.

программируемая. Различные категории семантических систем (вычисления, структуры данных, управление вычислениями и обработка памяти) могут быть подчинены разным правилам применения **R**, требующим различных систем обозначений (табл. 3 и 4).

Эти обозначения позволяют показать различие в пространствах допустимых значений и особенности ограничений на представление и выполнение функций в разных семантических системах языка. Например, формула [[∃ { BSD[7] … } ∀ { etd[8] …]] задает пространство допустимых данных, устроенное как кумулятивная иерархия структур данных BSD над любыми значениями из множеств элементарных типов значений etd. На уровне абстрактной семантики языка программирования определение etd представляется как набор предикатов, распознающих принадлежность значения к конкретному типу или виду элементарных значений. Определение BSD кроме предикатов содержит набор конструкторов и деструкторов, связанных с неявными функциями доступа к памяти и с диагностикой ошибочных значений.

Табл. 3. Обозначения для показа различий в определении **V** – основного множества структур, видов и типов данных языка программирования.

[[Формула]]	*Пояснение: Формула для семантических систем заключается в двойные квадратные скобки*
∀etd	Область определения всех функций опирается на любые элементы множества типов данных etd[9]
∃ BSD etd	Область определения всех функций может использовать любые элементы кумулятивной иерархии[10] базовых структур данных из множества BSD[11] над элементами типов данных из etd
V ‖ S V ∩ S	Фильтрация, пересечение множеств для выделения подходящих
⊂ ⊃ U ∩ ∈ ∉	Операции над множествами

[7] Базовые структуры данных.

[8] Элементарные типы значений.

[9] Примеры etd: атом, number, string, metaD (метаданные) и др.

[10] Универсум фон Неймана**,** кумулятивная иерархия множеств.

[11] Примеры BSD: list, array, hash, set, structure и др.

[[Формула]]	*Пояснение: Формула для семантических систем заключается в двойные квадратные скобки*
: => λ (x y)	Отображение, переход, формат представления схемы функции
^S ’S ©S	Методы обработки форм: eval quote compile
≈ $\equiv_s$ $\equiv_w$	Эквивалентно
∉!	Ошибочное значение
⊥!	Неопределенное значение

Такие обозначения позволили выразить различие между пространствами допустимых данных, создания и обработки элементарных, встроенных и программируемых структур данных в соответствии с принципами функционального программирования [22, 23]. Предложенные формулы могут использоваться как уточнение анализатора текстов программ для оптимизирующей компиляции и проверки условий семантической корректности программ и компиляторов.

Табл. 4. Обозначения для показа механизмов обработки памяти в языке программирования

((Действие))	*Пояснение: Действия заключаются в двойные круглые скобки*
< U ; …>	Структура блоков памяти
V ↕ S V ‖ S	Из V выбираются такие, как S
Δ S	За исключением **S**
→ ↓ ↑ ← ↔	Операции над элементами памяти: → инициированные, ↓ запись, ↑ чтение[12], ← удаление, ↔ обмен данными.
!@	Чтение произвольного элемента с удалением из структуры данных
@	Адрес произвольного элемента памяти из структуры данных
from ! из	Выбор произвольного элемента памяти из структуры данных
+=	Пополнение блока памяти
±	Переход к другой дисциплине функционирования

[12] ↑: переменная **=>** пара: данное с адресом

((Действие))	*Пояснение: Действия заключаются в двойные круглые скобки*
∅	Пустое множество
#	Число элементов блока памяти или структуры данных
*****	Многократное повторение операции, может ни одного
~◊	Недостижимый из программы элемент памяти
(GC …)	Вызов мусорщика из программы
H VM GC	Блок памяти, виртуальная машина, мусорщик
Prog ⊂ Heap	Включение одного блока памяти в другой
Var ∉ call	Вхождение элемента в выражение или категорию функций

Эти обозначения могут соответствовать неявным действиям, сопровождающим вычисления над структурами данных. Например, дисциплина обработки памяти, определяемая формулой ((→ ↓ ↑* ←))**,** задает действия, сопровождающие применение локальной переменной, как рассредоточенную последовательность[13] неявных операций над памятью. Сначала разрешено завести элемент памяти (**→**), ссылка на который связывается с переменной. Потом в элемент памяти разрешается записать значение (**↓**). После этого записанное значение можно читать из памяти произвольное число раз (**↑***) в пределах локальной области видимости. После выхода из этой области следует удалить ссылку на элемент памяти (**←**), связь переменной с элементом памяти исчезнет, он становится недостижимым из программы. Перед выполнением операции над памятью возможна проверка, является ли операция допустимой в последовательности, определяющей дисциплину обработки памяти. При определении языка программирования на уровне системной прагматики в таких обозначениях можно задавать и другие условия корректности работы с памятью, неявно сопровождающей семантические функции. В результате общее определение языка программирования можно выразить как комплект проекций – схем, выражающих возможность присоединять ассоциированные определения абстрактной семантики и системной прагматики к определению конкретного синтаксиса для автоматической генерации компилятора.

[13] Рассредоточенная последовательность задает порядок выполнения операций, между выполнением которых происходят вычисления, определенные независимо.

Такие обозначения позволили при сравнении диалектов языка Lisp выразить различие между семантическими системами поддержки обработки памяти и хранимых в ней значений в соответствии с принципами гибкости границ блоков памяти, использующей функцию GC – вызов мусорщика, и неизменяемости хранимых значений, адреса которых достижимы из программы [23]. Общая схема определения приобретает вид

Синтаксис [⟦ ***Семантика*** ⟧] [**((** Прагматика **))**]

Конструкциям уровня абстрактной семантики и системной прагматики могут соответствовать разные шаблоны кодогенерации, включая функционально эквивалентные шаблоны для отладчика, компилятора и интерпретатора, представления которых требуют другой, специальной макротехники.

Такие формулы можно использовать как представление дисциплины работы с памятью, контролируемой и на этапе компиляции, и в процессе исполнения программы, что может способствовать обеспечению надежности и безопасности программ.

НЕМНОГО ИСТОРИИ ЯЗЫКА Lisp

Идеи Джона Маккарти (John McCarthy), воплощенные в языке Lisp, сразу вызвали ревнивую критику со стороны как программистов, так и математиков. Математиков смущала противоречивость некоторых построений с точки зрения классической математики, например различие контекстов определения, вызова и вычисления функций[14]. Программисты не могли смириться с отсутствием в языке привычной техники, начиная с «изменения состояний памяти», а также с непредсказуемо медленной обработкой списков в сравнении с быстрой обработкой векторов.

К середине 70-х годов XX в. Дана Скотт (Dana S. Scott) опубликовал конструктивную теорию, смягчившую критицизм математиков, построив первую непротиворечивую модель бестипового λ-исчисления[15]. Скептицизм программистов оказался более устойчивее. Например, общее мнение, что Lisp – это интерпретируемый язык, скорее всего, связано с тем, что реализации языка Lisp

[14] https://en.wikipedia.org/wiki/Funarg problem/ – статья о Funarg-проблеме.

[15] https://ru.wikipedia.org/wiki/ – непрерывность по Скотту.

обычно предоставляют диалоговый – интерактивный стиль работы с программой на базе REPL-цикла и не формируют файл с результатом компиляции, поэтому не заметно, что фрагменты программного кода компилируются по мере необходимости. Такое мнение не исчезло при появлении в 1976 г. диалекта Scheme, использующего совмещение чтения программы с ее полной компиляцией, но, сохраняющего диалог и добавляющего неявную эффективную векторную реализацию списков.

Уже в начале 60-х годов XX в. язык Lisp, включая интерпретатор и компилятор, был описан в виде формализма на самом языке Lisp. Lisp позволяет создавать программы, динамически порождающие код, выполнять любые системные трюки, строить виртуальные машины, специализированные системы, диалекты и пакеты, расширяющие язык. Такой потенциал языка Lisp дает ответ на вопрос: «НАСКОЛЬКО новые задачи компьютерной обработки информации отличаются от традиционных задач обработки чисел?». Основные тезисы ответа представлены следующими утверждениями.

– Любой информации можно дать символьное представление, числа – частный случай символьного представления, элементарные данные другой природы можно представить как атомы.

– Все понятия программирования можно рассматривать как функции или применение функций. Переменные, операторы или команды – не более чем разные категории функций.

– Эксперименты при разработке решений новых задач продуктивнее выполнять в диалоге на базе интерпретаторов. Компиляция полезна для достаточно отлаженных программ.

– Списки произвольной длины из элементов любой природы могут быть гомоиконными конструкциям как высокого уровня постановки задачи, так и низкого уровня системных решений. Вектора, множества, таблицы и другие структуры данных на этапе экспериментов можно моделировать с помощью списков.

Концепции языка Lisp со временем кристаллизовались как парадигма функционального программирования [13][16], хотя реализация языка изначально

[16] Обзор литературы о функциональном программировании – https://alexott.net/ru/fp/books/

поддерживает основные императивные черты ради привлечения опытных программистов и возможности повышать надежность и эффективность программ.

Для быстрого ознакомления с идеями языка Lisp Дж. Маккарти выделил семантический базис – диалект Pure Lisp, включающий в себя пять функций обработки списков (CONS, CAR, CDR, EQ, ATOM), четыре конструктора выражений и функций (QUOTE, COND, LAMBDA, LABEL)[17] и универсальную функцию EVAL, способную вычислить любое выражение, правильно представленное списком, что определяет границы вычислимости без использования семантики изменения состояний памяти, без глобальных переменных. Пары функций QUOTE и EVAL достаточно для поддержки разных схем вычислений, включая ленивые вычисления, оптимизации программ и любую макротехнику как основу метапрограммирования, определения интерпретаторов и компиляторов. Диалект Pure Lisp дал ответ на вопрос «КАКОЙ может быть методика обучения программированию на языке Lisp?». Ответ выглядит следующим образом:

– Выделить краткую базовую семантику, достаточную для определения остальных конструкций языка (выражения, ветвления, рекурсивные функции).

– Дать примеры их символьного представления и применения при решении знакомых задач.

– Показать типовые решения некоторых задач с помощью отображений, ленивых вычислений и метапрограммирования.

– Привести определения интерпретатора и компилятора для изучаемого Pure Lisp на уровне калькулятора, а потом расширить определение Lisp-калькулятора до обобщенного интерпретатора и компилятора.

Такая система понятий и средств обработки данных позволила поддержать все семантические и прагматические принципы чисто функционального программирования, позднее реализованного как основная парадигма программирования на базе ленивых вычислений в языках ML (1973), Hope (1980), Haskell (1990) и др.

Дж. Маккарти ожидал, что оставшиеся проблемы организации вычислений будут решены в более поздней версии, условно названной Lisp 2, в которую

[17] QUOTE – блокировка, COND – ветвление, LAMBDA – безымянная функция, LABEL – именование.

планировал включить обработку многомерных векторов, сопоставление с образцами и организацию параллельных вычислений [11]. Рассказывая о языке Lisp, Дж. Маккарти подчеркивал, что, экспериментируя, программист может изменять в языке Lisp все что угодно, кроме константы Nil[18] [12]. К 1962 г. были готовы версия Lisp 1.5 и описание реализации системы, ставшей преемником самого раннего языка Lisp [2]. Это описание языка стало основой для создания Lisp-систем как в США, так и за их пределами, в нашей стране на БЭСМ-6, ЕС ЭВМ, СМ-4 и других машинах[19]. Сложные данные языка Lisp на уровне синтаксиса выглядят как списки элементов любой природы, хотя неявно в системе программирования на уровне системной прагматики языка поддержаны и другие структуры данных, такие как вектора, множества, хэш-таблицы и изменяемые поля, ставшие явными в более поздних диалектах. Хэш-таблица применяется для идентификации атомов, множество моделируется как список различимых параметров функции, размеченное множество используется при организации списков свойств атомов и пространств имен, изменяемые поля используются при организации рекурсии и оптимизации отложенных или ленивых вычислений, а вектора – для сопряжения со встроенными машинными процедурами.

В начале 60-х годов XX в. Питер Ландин (Peter J. Landin) в работе о λ-исчислении [12] ввел понятие "call-by-name"[20], использованное в описании языка Algol-60. В 1964 г. он предложил машину SECD – виртуальную и/или абстрактную машину, предназначенную для использования в качестве целевого языка (бэкэнд) при компиляции языков функционального программирования [12]. Вскоре появился Lispkit – реализация чисто функционального диалекта Pure Lisp с лексической областью видимости, разработанного в качестве испытательного и учебного стенда при изучении концепций функционального программирования, включая ранние эксперименты с ленивыми вычислениями [13, 14]. Полученные компилятор и виртуальная машина были легко переносимы. Важный

[18] Это утверждение Джон Маккарти произносил в конце декабря 1968 г. в кабинете А. П. Ершова в Новосибирске в цикле лекций, посвященных языку Lisp и верификации программ.

[19] https://www.computer-museum.ru/histsoft/lisp_sorucom_2011.htm

[20] «вызов по имени»

в контексте ленивых вычислений термин «мемоизация» был придуман Дональдом Мичи в 1968 г. [15]. В 1971 г. Кристофер Стрэйчи предложил термин "call-by-need"[21], предшественник термина «ленивые вычисления» [16].

В 1974 г. в Xerox началась разработка аппаратуры и системы машинных команд для аппаратной реализации языка Lisp. Язык Scheme был разработан в 1976 г. в MIT в рамках проекта по созданию Lisp-машин [3]. Появилось доказательство, что так называемая «неэффективность языка Lisp» обусловлена не свойствами языка, а особенностями компьютеров и методов реализации языка программирования. Практически одновременно термин "Lazy evaluation" (ленивые вычисления) был введен в статье "Programming Languages and Their Definition" Кристофера Стрейчи [17], в статье "A Lazy Evaluator" Питера Хендерсона и Джеймса Х. Морриса [16]. В 1978 г. был представлен язык программирования Lazy ML, который стал первым языком, основанным на парадигме ленивых вычислений. В 1978–1979 гг. был разработан язык программирования Норе[22] в Эдинбургском университете Великобритании. Этот язык оказал значительное влияние на Haskell, представленный программистскому сообществу в 1987 г. с целью притормозить создание новых языков функционального программирования, их новизна затрудняет работу экспертов при решении вопросов о приоритете публикуемых результатов.

В 1984 г., через 8 лет после Scheme, появился диалект Common Lisp – мультипарадигмальный язык общего назначения, дополняющий традиционные динамические решения языка Lisp механизмом статического связывания переменных и раздельных пространств имен, специальных средств программирования макросов, функционалов, пакетов и лексических замыканий функций [4]. В 1995 г. Common Lisp был стандартизован ANSI.

В последние годы особое внимание привлекает диалект Racket (ранее – PLT Scheme), созданный в 1994 г. как платформа языково ориентированного программирования [5]. Это симптом перехода практики программирования

[21] «вызов по необходимости»

[22] *Филд А., Харрисон П.* Функциональное программирование. Пер. под редакцией В. А. Горбатова. М. Мир, 1993, 638 с. Содержит описание различных вариантов «мусорщика».

от накопления правильности программ на уровне библиотек к уровню создания диалектов и проблемно ориентированных языков.

В 2007 г. появился Clojure – современный диалект языка Lisp, предлагающий решения по верификации программ и параллельному программированию [6–8].

ДИАЛЕКТ Scheme ДЛЯ ЭФФЕКТИВНОЙ РЕАЛИЗАЦИИ

Язык Scheme был разработан в 1976 г. в MIT в рамках проекта по созданию Lisp-машин [3]. Scheme ответил на вопрос «КАК сделать функциональное программирование столь же эффективным как императивное?». Ответ включал следующее:

– Разграничить универсальность символьных вычислений отказом от представления значений любой природы, а списки свойств символа оставить только для переменных и функций.

– За основу системных структур данных взять вектора, а списки использовать при необходимости как синтаксическое расширение в отдельном модуле.

– Чтение списочного представления программы, фактически являющегося представлением ее абстрактного синтаксического дерева, совместить с принудительной компиляцией, а универсальную функцию EVAL (интерпретатор) вынести из базиса во вспомогательный модуль.

– От рецептов отложенных функциональных аргументов, формируемых в точке вызова функции, перейти к формированию замыканий в точке определения функции, заодно и макротехнику ограничить выполнением на этапе компиляции, что сводит ее потенциал к привычным возможностям препроцессоров во многих языках программирования. При компиляции автоматически выполнять оптимизацию рекурсий, сводимых к циклам.

– Смягчить неизменяемость данных, опираясь на унификацию присваиваний и определений функций, выполненную к тому времени при разработке языка Algol 68, а заодно и разрешить изменять значения системных свойств символа (define = set![23]).

[23] set! работает только на ранее введенных символах, define на любых.

Язык Scheme заимствовал терминологию и синтаксис языка Lisp, несколько изменив смысл ряда понятий и сузив трактовку почти всех принципов функционального программирования. Из характерных особенностей языка Lisp в языке Scheme поддержана примерно треть. Самым заметным отклонением от языка Lisp является введение булевых значений #f и #t вместо использования константы Nil или пустого списка в качестве значения «ложь»[24]. Переход к векторам пошатнул позиции пустого списка и атома Nil – реализация пустых векторов в те годы не практиковалась. Принудительная компиляция программы, теряющая исходное абстрактное синтаксическое дерево, затрудняет возможности динамической оптимизации программ и процессов. Такая компиляция устраняет повторную интерпретацию программ, что является оптимизацией многократно используемых программ.

Можно считать, что так выделилось минимальное ядро грамматики языка Lisp. Теперь существуют реализации Scheme на JVM.

Common Lisp ДЛЯ ПРОИЗВОДСТВЕННОГО ПРИМЕНЕНИЯ

Common Lisp был разработан в начале 80-х годов XX в. с целью объединения полезных механизмов большого числа разрозненных диалектов языка Lisp [4]. Common Lisp часто противопоставляют языку Scheme – это два самых популярных диалекта Lisp. Scheme предшествовал Common Lisp и исходил не только из той же традиции Lisp, но и от тех же разработчиков. Гай Стил, вместе с которым Джеральд Джей Сассман разработал Scheme, возглавлял комитет по стандартизации Common Lisp, в котором было преодолено сужение потенциала языка Lisp. Сохраняя и восстанавливая основные концептуальные решения языка Lisp, диалект Common Lisp их дополняет, расширяя методы формирования пространств допустимых процессов.

[24] Вопреки предостережению Джон Маккарти: «Программист может изменять в языке Lisp все что угодно, кроме константы Nil». Математиков смущало, что Nil одновременно и атом, и список. В теории удобнее, чтобы любое данное принадлежало ровно одному типу.

Common Lisp иногда называют Lisp-2[25], а Scheme – Lisp-1, имея в виду использование раздельных пространств имен для функций и переменных[26]. CLISP проводит различие между временем чтения, временем компиляции, временем загрузки и временем выполнения программы и позволяет пользовательскому коду также учитывать это различие при выборе желаемого типа обработки программы на нужном этапе. В системе CLISP реализован пакет CLOS, дающий полную поддержку популярной в производстве парадигмы ООП.

Этот диалект резко расширил сферу производственного применения языка Lisp, используя на уровне лексикона семантику доступа к многомерным матрицам, хэш-таблицам, программируемым структурам данных, подобным структурам в языке C, изменяемым полям и мультизначениям, удобными для моделирования независимых потоков. Все это внешне представляется как списки, но реализовано как эффективные структуры данных, доступные через функции. Массивы могут содержать любой тип значений в качестве элемента, смешивать разные типы в одном массиве или могут быть специализированы, содержать только определенный тип.

Common Lisp ответил на вопрос: «ЧТО может дать функциональное программирование программной индустрии?». Ответ включает следующие дополнения:

– Универсальность символьной обработки и структур данных неограниченного размера дополнена средствами обработки конечных чисел и структур данных, типичных для большинства языков программирования и приложений.

– Компиляция отдельных функций допускает и компиляцию полной программы без потери исходного списочного представления абстрактного синтаксического дерева.

– Самоопределение в форме рекурсии обогащено разнообразием схем циклов, превосходящих по возможностям типовые схемы.

– Гибкость распределения памяти с возможностью программировать вызов «сборщика мусора» дополнена средствами выяснять время, даты и этапы

[25] Ассоциация с проектом Lisp-2 Дж. Маккарти [11].

[26] Язык Lisp распался на два семейства – Lisp-1 и Lisp-2, различаемые по отношению к статике и динамике, возможностям применения списков свойств и роли атома NIL = ( ) в логике управления вычислениями.

работы программы, чтобы прогнозировать целесообразность применения тех или иных методов.

– Неизменяемость хранимых значений уточняется введением неявного понятия «поле» для работы с изменяемыми системными свойствами символа.

– Для функций, потребляющих много памяти, предоставляются их деструктивные аналоги, приспособленные к более эффективной обработке данных (conc-*nconc, subst-nsubst, reverse-nreverse, union-nunion, mapcar-mapcon и др.*)

– Единственность результата функции расширяется на мультизначения, поддерживающие переход к многозначным функциям и организации параллельных вычислений.

Введены понятия «поле», «пакет» и «мультизначение». Основное отличие от языка Lisp связано с разделением пространств имен в зависимости от их назначения и включения в разные пакеты для решения отдельных классов задач.

Common Lisp поддерживает средства динамического анализа и использовался в разработке автоматизированных средств проверки доказательств теорем (ACL2) и систем компьютерной алгебры (Аксиома, Maxima).

СРЕДА Racket ДЛЯ СОЗДАНИЯ НОВЫХ ЯЗЫКОВ ПРОГРАММИРОВАНИЯ

Очередная версия учебной среды программирования на языке Scheme, разрабатываемая с 1995 г. в Университете Дьюка для обучения созданию, разработке и реализации новых языков программирования, в 2010 г. получила название Racket [5, 20]. Это означает переход интересов от продуктивности программирования к продуктивности разработки компиляторов.

Образовательная направленность повлияла на общую структуру языка Racket как систему диалектов, соответствующих уровням обучения. Это заодно позволило в реализации языка сохранить решения языка Scheme, принятые под прессом компьютерных характеристик середины 70-х годов XX в., и дополнить системную поддержку языка Racket в соответствии со значительно усовершенствованными возможностями современной элементной базы и новыми требованиями ИТ.

Разработчики диалекта Racket выявили ключевые недостатки Scheme, затрудняющие разработку крупных и надежных систем, а именно отсутствие мо-

дульной системы, слабую поддержку обработки исключений, метапрограммирования и расширяемости, ограничения в типизации и структурах данных, отсутствие способов для построения безопасных и масштабируемых систем. Для преодоления таких недостатков Racket был создан как диалект языка Lisp и используется для реализации языков программирования, компиляторов и интерпретаторов, образовательных систем и платформ, языков для обучения и преподавания, включая обучающие и экспериментальные языки. Оставаясь наследником Scheme, Racket делает ставку на практичность, расширяемость и богатство инструментов, уделяя больше внимания удобству использования и обучения, поддержку метапрограммирования и инструментальных средств для разработки языков программирования. Racket включает макросы на этапе и компиляции, и выполнения с удобными функциями для разработки и отладки. Поддержаны диалект Scribble – язык разметки для документации и ряд диалектов, связанных с основными парадигмами программирования, проблемноориентированными языками (*domain-specific language – DSL*) и приложениями ИТ.

Производительность Racket обеспечена JIT-компилятором и механизмом «сборки мусора» с поддержкой поколений объектов. Включена поддержка мелкозернистого параллелизма. Имеется учебная версии Minimal Racket без пакетов, поддержка байт-кода и JIT-компиляции для архитектуры ARM[27], а также быстродействующий Typed Racket и другие диалекты. Разработана собственная виртуальная машина. В экспериментах по разработке разных версий Racket обнаружилось, что компиляция не всегда повышает производительность программ, но может способствовать продуктивности программирования[28]. Система макросов в Racket используется для создания полных языковых диалектов, затрагивая семантику. Racket отвечает на вопрос «КТО будет определять языки программирования в будущем?».

Ответ достигается следующими решениями:

– Универсальность символьных вычислений распространена на приобретение профессиональных навыков, включающих разработку документации, что

[27] Архитектура ARM (*Advanced RISC Machine)* – усовершенствованная RISC-машина для мобильных устройств.

[28] Сотрудники фирм-разработчиков компиляторов утверждают, что необходимость ожидать результат компиляции дает им защищенную нишу времени для продумывания программ.

соответствует предложенной Д. Кнутом парадигме литературного (грамотного) программирования (literate programming)[29].

– Среди диалектов выделены языки, соответствующие уровням способностей, навыков и знаний студентов (Minimal Racket, Racket, Lazy Racket, Typed Racket и др.).

– Независимость параметров поддержана механизмом сопоставления с образцами, достаточными для представления грамматик с переводом.

– Самоопределение и рекурсивные функции подкреплены практикой применения генерации лексеров/парсеров, а также определением языков на уровне абстрактного синтаксического дерева.

– Гибкость распределения памяти сопровождается средствами рефакторинга, тестирования и измерения производительности кода.

– Единственность результатов функции расширена средствами организации асинхронных процессов, подразумевающих мультизначения.

Racket примерно на треть наследует решения языка Scheme и на две трети возвращается к исходным решениям языка Lisp. Самое заметное отличие диалекта Racket от семантики языка Lisp – использование специального булева значения #f в качестве значения «ложь» вместо пустого списка ( ) или атома Nil. Хотя формально язык Racket называют диалектом языка Scheme, по своим особенностям он ближе к Common Lisp.

Clojure – НОВЫЕ ГОРИЗОНТЫ ИТ

Современный диалект языка Lisp Clojure появился в 2007 г., разработан Ричем Хикки (Rich Hickey), независимым разработчиком ПО, ранее разработавшим dotLisp в рамках проекта **.**NET Framework. Clojure наследует особенности языка Lisp, обеспечивающие гибкое и мощное метапрограммирование, поддерживает синтаксическое расширение [6–8] и надежную семантику, дополненную механизмами системной прагматики для программирования приложений на базе

[29] http://www.literateprogramming.com/ – методика профилактики кризиса ухода исполнителя и разбухания описаний, создаваемых техническими писателями. Не исключено, что методика не стала массовой из-за высокого темпа развития ИТ, опережающего созревание речевой практики. Возможно, Д. Кнут хотел привлечь внимание к тому, что искусство программирования основано и на владении естественным языком.

распределенных и параллельных процессов.

В этом языке определения функций могут неформально сопровождаться пред- и постусловиями, что помогает проверке утверждений об аргументах и полученных результатах, а также позволяет при тестировании проверять инварианты функций. Clojure, как язык программирования, не предоставляет встроенных методов верификации программ, но для обеспечения корректности программ использует различные подходы и инструменты, включающие средства типизации и спецификации (clojure.spec, malli), динамический контроль данных и тестирование (clojure.test, test.check), помогающие обнаруживать неожиданности. Поддержана интеграция с внешними инструментами для формальной верификации с возможностью автоматического тестирования свойств, такими как Rosette, Z3, SMT (satisfiability modulo theories) или ACL2, проверки типов данных (clj-kondo, Eastwood), а также статические анализаторы для проверки безопасности кода или соблюдения определенных стандартов.

Динамическая типизация означает, что, кроме статической проверки типов переменных на этапе компиляции, проверяются типы данных, точно известные во время выполнения, что существенно повышает надежность и безопасность вычислений. Заодно это обеспечивает гибкость и быструю разработку, так как не требует явного объявления типов переменных. Clojure уделяет большое внимание тестированию (clojure.test, test.check, midje, kaocha, Unit Testing, Integration Testing, Property-based testing), вызывающему у практиков больше доверия, чем верификация.

Статический анализ выявляет потенциальные ошибки, нарушения стиля кодирования, неиспользуемый код и константные вычисления (эквивалент чисто функционального программирования), поддерживает автодополнения для более раннего обнаружения ошибок. Доступны библиотеки, позволяющие определять контракты (core.contracts, lucid.policy) для функций, и автоматическая генерация документации. Возможна явная реструктуризация структур данных, работающая с любой последовательностью, включая:

- списки, вектора и последовательности Clojure;
- любые коллекции, реализующие java.util.List (например, ArrayLists и LinkedLists);
- Java-массивы;

- строки, реструктурированные как списки символов;
- списки аргументов функций.

Реструктуризация означает переход от ранее созданной структуры данных к структуре с другим методом доступа при сохранении ее наполнения[30], допускается использование подчеркивания «_» для обозначения игнорируемой позиции. Исходная структура сохраняется в соответствии с принципом неизменяемости хранимых значений. Реструктурированная последовательность аргументов функции позволяет вместо имен связанных переменных использовать нумерацию параметров, список которых можно рассматривать как вектор[31]. К первому аргументу функции можно обращаться, просто используя «%».

Параллельное программирование использует транзакционную память, как в базах данных, а также агентов и разные виды указательных переменных. Поддержаны ленивые последовательности, вспомогательные процессы и введено несколько неявных понятий для поддержки параллелизма и программирования своих структур данных с учетом проблем надежности и безопасности.

В качестве компромисса между идеями чисто функционального программирования и необходимостью изменения состояний при организации параллельных процессов введены указательные переменные – атомы из уникальных указателей на списки свойств атома превращаются в указательные переменные и рассматриваются как системные структуры, обеспечивающие разные дисциплины доступа к памяти и стратегии многопоточности (atom, ref, agent).

Clojure отвечает на вопрос «ГДЕ новые горизонты, в освоении которых помогает продуктивность и моделирующая сила языка Lisp?». Ответ опирается на следующее.

– Универсальность символьных вычислений распространена на явный конкретный синтаксис отображений, множеств и векторов. Введены метаданные – кодированный аналог списка свойств, который может быть связан со значением или указательной переменной.

– Можно синхронизовать выполнение потоков: откладывание, ожидание, обещание, передача, готовность, блокировка (delay, future, promise, deliver, is-

[30] Похоже на реорганизацию векторов в языке APL.

[31] Подобно некоторым языкам заданий и макропроцессоров.

done?, deref), контролировать ход вычисления, а также представлять эффективные формы циклов, не использующие стек.

– Неизменяемость данных при необходимости изменений превращена в транзакционную память, как в базах данных, поддержаны агенты и разные виды динамических и указательных переменных.

– Единственность результата функции дополнена возможностью проверки утверждений об аргументах и полученных результатах, при тестировании можно проверять инварианты функций и использовать внешние методы верификации программ.

Диалект Clojure наследует примерно 80% особенностей языка Lisp, уточняет ряд его решений для удобства представления параллельных процессов и дополняет его заметным комплектом средств, соответствующих современной элементной базе и поддерживающих отладку взаимодействующих процессов и удостоверение правильности программ. Самое заметное расширение связано с понятием «атом». Атом из неявного уникального указателя на список свойств атома стал явной указательной переменной, позволяющей поддерживать различные дисциплины обработки памяти, возникающие в разных моделях параллельных вычислений, разнообразие которых непредсказуемо велико. Кроме того, механизм реструктуризации данных распространен на список аргументов, что расширяет форматы определения функций, позволяет отказываться от использования имен связанных переменных[32], что удобно при генерации машинного кода.

ИТОГ СРАВНЕНИЯ ДИАЛЕКТОВ

В ходе эксперимента выяснилось, что результаты сравнения языков программирования следует показывать декомпозированными по отдельным особенностям и конструкциям языка программирования. Обозначения из табл. 1–4

[32] Интересно, что в свое время основатель нашей математической школы Н. Н. Лузин не одобрял термин «связанная переменная», он пояснял, что это вообще не переменная, потому что ее имя можно заменять на другое – смысл формулы не изменится. Своего термина не предложил. *Лузин Н. Н.* Интеграл и тригонометрический ряд. Изд. 2-е. М.: Гостехиздат, 1951.

оказались достаточными для показа наиболее очевидных синтаксических, семантических и прагматических различий между рассмотренными диалектами, причем с выражением отношения наследования. Удалось показать происходившее при создании диалектов изменение особенностей поддержки семантических принципов «универсальность», «самоопределение функций» (рекурсия), «независимость и равноправие параметров функций», «единственность результата функции», а также форматов ветвлений, ленивых вычислений и REPL-цикла. Кроме того, показано произошедшее изменение особенностей поддержки принципов «гибкость границ блоков памяти» и «неизменяемость хранимых в памяти значений» [23].

Отмечая различие в целях создания диалектов языка Lisp, можно заметить, что рассмотренные диалекты обладают стабильным системным ядром, использующим конкретный комплект структур данных и механизмов их обработки. Показано, что вариации структур данных и значений сводятся к пересмотру границ между лексиконом, синтаксисом, семантикой и прагматикой языка, к различию возможностей периода компиляции, выполнения и отладки программы и приоритетов между областями видимости символов, а также к вариантам представления значения «ложь» и расширению понятия «атом» от уникального указателя на список свойств атома до понятия «указательная переменная». Системные решения по обработке структур данных в новых диалектах из неявных уровнях системной прагматики становятся доступными сначала с помощью функций на уровне лексикона абстрактной семантики, затем получают представление на уровне конкретного синтаксиса. Это не нарушает функциональную эквивалентность программ, абстрактное синтаксическое дерева всех диалектов имеет подобное списочное представление. Семантика вычислений в диалектах по-разному взаимодействует с прагматикой изменения состояний памяти, что проявляется, как правило, в именовании функций и выборе границ изменяемых данных.

Принципы и понятия функционального программирования в диалектах языка Lisp уточнялись в зависимости от роли продуктивности программирования и критериев качества программ в разных областях приложения. Для Scheme – это эффективность и сохранение привычки к присваиваниям и векторам, для

Common Lisp – разнообразие структур данных, для Racket – создание специализированных диалектов, освобождающих от нагромождения библиотек, для Clojure – освоение многопроцессорных комплексов и распределенных информационных систем. Более подробно результаты сравнения представлены в препринтах [22, 23].

Разнообразие целей и решений, принятых в рассмотренных диалектах, позволило сформулировать особенности диалектного абстрагирования семантики вычислений от семантики изменения состояний памяти. Каждый из диалектов произвел определенное уточнение особенностей функционального программирования без отказа от его принципов, впервые поддержанных в реализации языка Lisp. Семантические и прагматические принципы функционального программирования (универсальность представления данных и программ, самоприменимость, равноправие параметров и единственность результатов функций, гибкость границ памти и неизменяемость хранимых значений) дополнились конструкциями ветвлений для представления частичных вычислений, ленивых вычислений для управления временем вычислений и REPL-циклом[33] для поддержки удобной отладки программ.

Гомоиконный конкретный синтаксис поддерживает универсальность представления данных и программ с помощью общих структур данных, удобных для самоопределения рекурсивных функций и структур данных, границы которых могут быть заданы как частичные функции с помощью ветвлений и циклов. Начиная с символьных представлений с помощью S-выражений языка Lisp и списков, достаточных для моделирования любых структур данных, диалекты предложили конкретный синтаксис для векторов, множеств и хэш-функций, присутствовавших на уровне системной прагматике:

(S-выражение …) *// список через пробел в языке Lisp.*
(S-выражение «**.**» S-выражение) *// пара в языке Lisp.*
((*S-выражение «.» S-выражение*) …)
// ассоциативный список *в языке Lisp.*
(*индикатор S-выражение* …)
// список свойств атома *в языке Lisp.*

[33] REPL-цикл — Read Eval Print Loop.

```
[S-выражение … ]
        // группировка или вектор в диалекте Scheme.
[S-выражение … ]
        // последовательность в диалектах Racket и в Clojure
#{S-выражение ... }   // множество в диалекте Clojure
{ ( ключ => значение ) … }
    // хэш-таблица или ассоциативный список в диалекте Clojure.
```

На уровне трансформационной абстрактной семантики существуют разные определения функций и представления форм, результаты которых совпадают на одинаковых комплектах аргументов в одинаковых контекстах. Это дает основания для оптимизирующих эквивалентных преобразований программ, включая исключение константных (чисто функциональное программирование) или дублирующих вычислений, перестановочность параметров и аргументов, вынесение подвыражений в аргументы, преобразование рекурсий в циклы, отложенные или ленивые вычисления, а также различные средства проверки правильности программ, предложенные в диалекте Clojure. Развитие вариантов REPL-цикла при отладке программ показывает целесообразность использования частичной или полной компиляции наряду с интерпретацией без потери исходного кода программы. Такие варианты предложены в диалектах Scheme, Common Lisp и Racket.

На уровне системной прагматики поддержка принципов функционального программирования использует исключения, обработка которых необходима для обеспечения продуктивности программирования и производительности программ. В динамике возникает переключение дисциплины функционирования памяти и вызов вариантов мусорщика с оптимизацией. Поддержка параллелизма и асинхронности привела в диалекте Clojure к расширению понятия атом до указательной переменной, а требование производительности программ повлекло поддержку метаданных и транзакций.

ЗАКЛЮЧЕНИЕ

Проведенные исследования форм для представления и обзора результатов сравнения наиболее популярных диалектов языка Lisp показали возможность лаконичного показа наследования их особенностей на уровне синтаксиса,

семантики и прагматики. Основная причина проведения такого эксперимента, а также поиска обозримых форм и выбора кратких обозначений связана с разработкой методик оценки продуктивности языка программирования и программируемых решений в отличие от непосредственного измерения эффективности и производительности программ. Произошло выделение понятия «диалектные абстракции», удобного при декомпозиции определений языка программирования на автономные составляющие, и понятия «логическая интерпретация», допускающего независимое варьирование и развитие сквозных понятий в процессе эволюции ИТ с сохранением общей архитектуры языка.

Сравнение диалектов языка Lisp показывает медленное смягчение программистского скептицизма в отношении к принятым в языке Lisp решениям и парадигме функционального программирования по мере прогресса элементной базы и развития ИТ, что видно по созданию новых языков программирования, рекламирующих включение механизмов функционального программирования как важное преимущество.

Полученные результаты образуют основу для определения номенклатуры семантических систем языков функционального программирования. Следует отметить не только новые диалекты языка Lisp, но и выпуски их реализаций – 25 августа 2025 г. выпущена очередная реализация Armed Bear Common Lisp (ABCL) [25].

При измерении мощности языков программирования как характеристики пространства доступных процессов вычислений самыми мощными являются языки ассемблера. Языки более высокого уровня, даже языки управления заданиями в операционных системах, теряют часть такого пространства ради удобства представления процессов обработки данных и управления ими. Такая потеря отчасти компенсируется моделированием, влекущим снижение эффективности ради продуктивности. Диалекты Scheme, Common Lisp, Racket и Clojure обладают реализацией на JVM, что говорит об их равномощности, они предоставляют одно и то же пространство процессов вычислений.

Сравнение диалектов Scheme, Common Lisp, Racket и Clojure, последний из них появился в 2007 г., показывает, что в них сохранены основные возможности языка Lisp, системные структуры данных, базовые принципы функционального программирования и понятия с определенными вариациями на злобу дня.

Авторы этих диалектов четко называют свои диалекты вариантами языка Lisp. Появление названий “Racket” и “Clojure” не отменяет роль так названных диалектов в успешной адаптации языка Lisp к новым поколениям программистов и новым возможностям аппаратуры. Знакомство с языками функционального программирования, наследующими идеи языка Lisp, такими как Sisal, F# и Haskell [19, 20, 26, 27], дает достаточное основание рассматривать Lisp как базовую математику функционального программирования. Результаты их анализа выходят за пределы настоящей статьи.

Семейство Lisp теперь является одним из старейших и наиболее влиятельных семейств языков программирования, его мощность в различных источниках, начиная с Википедий, оценивается от пятисот до тысяч языков программирования и диалектов. Кроме того, следует учесть языки функционального программирования, наследующие основные идеи языка Lisp, они постоянно разрабатываются, улучшаются и появляются новые, их число также оценивается в сотни или тысячи. Во многих источниках Lisp называют чемпионом по числу диалектов и наследников, хотя встречаются и утверждения, что сравнимое число диалектов имеется у языков C/C++, Bash, Perl, Python, JavaScript, BASIC, Forth, SQL, Fortran, Pascal, Ada, Assembler.

Следующий эксперимент по отладке методики анализа и сравнения языков программирования предполагается выполнить на материале языков функционального программирования, таких как Erlang, Sisal, F# и Haskell, рассматриваемых как наследники языка Lisp. Кроме того, интересно показать результаты сравнения представителей других долгоживущих семейств языков программирования, в первую очередь это Fortran и C, сохранивших значимость до наших дней. Отдельная задача – сравнение наших ЯП с зарубежными аналогами.

Благодарности

Автор искренне благодарен Андрею Валентиновичу Климову за ценные рекомендации по улучшению стиля изложения и поиску форм для представления результатов сравнения ЯП, Николаю Вячеславовичу Шилову, Игорю Сергеевичу Ануурееву и Борису Леонидовичу Файфелю за интерес к языку Lisp и стимулирующие вопросы.

СПИСОК ЛИТЕРАТУРЫ

1. *Городняя Л.В.* О представлении результатов анализа языков и систем программирования. Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018.

2. *McCarthy J. Abrahams P. W., Edwards D. J. et al.* LISP 1.5 Programming Manual. The MIT Press, Cambridge, 1963. 106 p.

3. *Dybvig K.R.* The Scheme Programming Language. https://www.scheme.com/tspl4/

4. *Graham P.* ANSI Common Lisp. Prentice Hall, 1996. 432 p.

5. The Racket Reference. https://docs.racket-lang.org/reference/

6. Clojure Programming. OReilly.com. Retrieved 2013-04-30. https://cdn.oreillystatic.com/oreilly/booksamplers/9781449394707_sampler.pdf

7. *Omm A.* Введение в Clojure. https://alexott.net/ru/clojure/clojure-intro/

8. Differences Clojure with other Lisps. https://clojure.org/reference/lisps/

9. *Backus J.W.* The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM Conference // Proceedings of the International Conference on Information Processing. UNESCO. 1959. P. 125–132.

10. *Backus J*. Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs // 1977 ACM Turing Award Lecture, p. 621–641.

11. *Mitchell R.W.* LISP 2 Specifications Proposal. Stanford Artificial Intelligence Laboratory Memo No. 21, Stanford, Calif., 1964.

12. *Лавров С.С., Силагадзе Г.С.* Входной язык и интерпретатор системы программирования на базе языка ЛИСП для машины БЭСМ-6. М.: ИТМ и ВТ АН СССР, 1969.

13. *Landin P.J.* The Mechanical Evaluation of Expression // Comput. J. 1964. Vol. 6, No. 4. P. 308–320. https://doi.org/10.1093/comjnl/6.4.308

14. *Хендерсон П.* Функциональное программирование. Применение и реализация. М.: Мир, 1983. 349 с.

15. *Henderson Peter; Jones Geraint A.; Jones Simon B.* The LispKit Manual. University of Oxford Computing Lab. 1983.
https://github.com/hanshuebner/secd/tree/master/lispkit/LKIT-2

16. *Michie D.* 'Memo' Functions and Machine Learning" (PDF). Nature. 1968. Vol. 218 (5136), P. 19–22. Bibcode:1968Natur.218...19M.
https://doi.org/10.1038/218019a0. S2CID 4265138

17. *Strachey Ch.* Fundamental Concepts in Programming Languages // Higher-Order and Symbolic Computation. 2000. Vol. 13, No. 1–2. P. 11–49.[34]

18. *Henderson P., Morris JH.* A lazy evaluator. Symposium ACM Sigact-Sigplan sur les principes des langages de programmation // DBLP, Proceedings of the 3rd ACM SIGACT-SIGPLAN symposium on Principles on programming languages (POPL), 1976. P. 95–103.

19. *Душкин Р.В.* Функциональное программирование на языке Haskell. М.: ДМК Пресс, 2008. 544 с., ил.

20. Официальный сайт языка Haskell – "О языке"
http://haskell.org/aboutHaskell.html

21. From PLT Scheme to Racket. Racket-lang.org. Retrieved 2011-08-17. https://docs.racket-lang.org/guide/intro.html Welcome to Racket

22. *Городняя Л.В.* Lisp и его диалекты. Новосибирск, препринт, 2025. https://www.iis.nsk.su/repository/gorod.14408

23. *Городняя Л.В.* Формы для показа результатов сравнения языков программирования на примере диалектов языка LISP.

[34] Предварительные публикации: *Strachey Christopher.* Programming Languages and Their Definition; *Strachey Christopher.* Fundamental Concepts in Programming Languages, 1967

www.iis.nsk.su/files/preprint/gorodnyaya-2025-forms_0.pdf?ysclid=mk9e9ot2mp144838343

24. *Городняя Л.В.* Сравнение диалектов языка Lisp // Материалы конференции «Научный сервис в сети Интернет», 2025. https://keldysh.ru/abrau/2025/temp/17.pdf

25. Armed Bear Common Lisp (ABCL). https://armedbear.common-lisp.dev/

26. *Евстигнеев В.А., Городняя Л.В., Густокашина Ю.В.* Язык функционального программирования SISAL // В сб. «Интеллектуализация и качество программного обеспечения». Новосибирск, 1994. С. 21–42.

27. *Сошников Д.В.* Программирование на F#. М.: ДМК Пресс, 2011. 192 с.

FORMS FOR DISPLAYING THE RESULTS OF COMPARISON OF PROGRAMMING LANGUAGES USING THE EXAMPLE OF DIALECTS OF THE LISP LANGUAGE

L. V. Gorodnyaya[0000-0002-4639-9032]

A. P. Ershov Institute of Informatics Systems, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia

lidvas@gmail.com

Abstract

This article focuses on developing forms for presenting the results of analyzing and comparing the characteristics of programming languages, systems, and paradigms. The proposed form is demonstrated through a comparison of the Lisp language, its most successful dialects (Scheme, Common Lisp, Racket, Clojure), and the functional programming paradigm across different levels of language and system definition. The form allows for a concise presentation of the inheritance of several features of the Lisp language and their evolution in its dialects, at the levels of concrete syntax, abstract semantics, and implementation pragmatics."

***Keywords**:* programming language, Lisp, Scheme, Common Lisp, Racket, Clojure, functional programming, comparison of programming languages, concrete syntax, abstract semantics, implementation pragmatics.

REFERENCES

1. *Gorodnyaya L.V.* O predstavlenii rezul'tatov analiza yazykov i sistem programmirovaniya. Nauchnyy servis v seti Internet: trudy XX Vserossiyskoy nauchnoy konferentsii (17–22 sentyabrya 2018 g., g. Novorossiysk). M.: IPM im. M.V. Keldysha, 2018.

2. *McCarthy J. Abrahams P. W., Edwards D. J. et al.* LISP 1.5 Programming Manual. The MIT Press, Cambridge, 1963. 106 p.

3. *Dybvig K.R.* The Scheme Programming Language.
URL: https://www.scheme.com/tspl4/

4. *Graham P.* ANSI Common Lisp. Prentice Hall, 1996. 432 p.

5. The Racket Reference. URL: https://docs.racket-lang.org/reference/

6. Clojure Programming. OReilly.com. Retrieved 2013-04-30. URL: https://cdn.oreillystatic.com/oreilly/booksamplers/9781449394707_sampler.pdf

7. *Ott A.* Vvedeniye v Clojure.
URL: https://alexott.net/ru/clojure/clojure-intro/

8. Differences Clojure with other Lisps.
URL: https://clojure.org/reference/lisps/

9. *Backus J.W.* The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM Conference // Proceedings of the International Conference on Information Processing. UNESCO. 1959. P. 125–132.

10. *Backus J*. Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs // 1977 ACM Turing Award Lecture, p. 621–641.

11. *Mitchell R.W.* LISP 2 Specifications Proposal. Stanford Artificial Intelligence Laboratory Memo No. 21, Stanford, Calif., 1964.

12. *Lavrov S.S., Silagadze G.S.* Vkhodnoy yazyk i interpretator sistemy programmirovaniya na baze yazyka LISP dlya mashiny BESM-6. M.: ITM i VT AN SSSR, 1969.

13. *Landin P.J.* The Mechanical Evaluation of Expression // Comput. J. 1964. Vol. 6, No. 4. P. 308–320. https://doi.org/10.1093/comjnl/6.4.308

14. *Khenderson P.* Funktsional'noye programmirovaniye. Primeneniye i realizatsiya = Functional Programming. M.: Mir, 1983. 349 p.

15. *Henderson P., Jones G.A.; Jones S.B.* The LispKit Manual. University of Oxford Computing Lab. 1983.
URL: https://github.com/hanshuebner/secd/tree/master/lispkit/LKIT-2

16. *Michie D.* 'Memo' Functions and Machine Learning" (PDF). Nature. 1968. Vol. 218 (5136), P. 19–22. Bibcode:1968Natur.218...19M.
https://doi.org/10.1038/218019a0. S2CID 4265138

17. *Strachey Ch.* Fundamental Concepts in Programming Languages // Higher-Order and Symbolic Computation. 2000. Vol. 13, No. 1–2. P. 11–49.

18. *Henderson P., Morris JH.* A lazy evaluator. Symposium ACM Sigact-Sigplan sur les principes des langages de programmation // DBLP, Proceedings of the 3rd ACM SIGACT-SIGPLAN symposium on Principles on programming languages (POPL), 1976. P. 95–103.

19. *Dushkin R.V.* Funktsional'noye programmirovaniye na yazyke Haskell / Gl. red. D.A. Movchan. M.: DMK Press, 2008. 544 p.

20. Ofitsial'nyy sayt yazyka Haskell. "O yazyke"
http://haskell.org/aboutHaskell.html

21. From PLT Scheme to Racket. Racket-lang.org. Retrieved 2011-08-17.
URL: https://docs.racket-lang.org/guide/intro.html Welcome to Racket

22. *Gorodnyaya L.V.* Lisp i yego dialekty. Novosibirsk, preprint, 2025.
URL: https://www.iis.nsk.su/repository/gorod.14408

23. *Gorodnyaya L.V.* Formy dlya pokaza rezul'tatov sravneniya yazykov programmirovaniya na primere dialektov yazyka LISP.
URL: www.iis.nsk.su/files/preprint/gorodnyaya-2025-forms_0.pdf?ysclid=mk9e9ot2mp144838343

24. *Gorodnyaya L.V.* Sravneniye dialektov yazyka Lisp // Materialy konferentsii "Nauchnyy servis v seti Internet", 2025.
URL: https://keldysh.ru/abrau/2025/temp/17.pdf

25. Armed Bear Common Lisp (ABCL). URL: https://armedbear.common-lisp.dev/

26. *Yevstigneyev V.A., Gorodnyaya L.V., Gustokashina Yu.V.* Yazyk funktsional'nogo programmirovaniya SISAL // v sb. «Intellektualizatsiya i kachestvo programmnogo obespecheniya». Novosibirsk, 1994. S. 21–42.

27. *Soshnikov D.V.* Programmirovane na F#. M.: DMK Press, 2011. 192 p.

СВЕДЕНИЯ ОБ АВТОРЕ

ГОРОДНЯЯ Лидия Васильевна – к. ф.-м. н., старший научный сотрудник Института систем информатики им. А.П. Ершова СО РАН, доцент Новосибирского государственного университета, специалист в области системного программирования и образовательной информатики.

Lidia Vasiljevna GORODNYAYA – Senior Researcher at the A. P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Associate Professor at the Novosibirsk State University, specialist in system programming and educational informatics.

email: gorod@iis.nsk.su

ORCID: 0000-0002-4639-9032

Материал поступил в редакцию 6 января 2026 года

УДК 519.178+004.9

НАУЧНОЕ СОАВТОРСТВО ПО ДАННЫМ РИНЦ И SCOPUS ЗА 2000–2020 ГОДЫ: ТЕНДЕНЦИИ РОСТА

С. А. Дурнев[1] [0009-0005-0338-5430], Е. А. Знаменская[2] [0000-0003-3630-712X], А. А. Печников[3] [0000-0002-0683-0019], Д. Е. Чебуков[4] [0000-0001-9738-8707]

[1]*ООО «Научная электронная библиотека», г. Москва, Россия*
[2, 4]*Математический институт им. В. А. Стеклова РАН, г. Москва, Россия*
[3]*Институт прикладных математических исследований – обособленное подразделение ФИЦ «Карельский научный центр РАН», г. Петрозаводск, Россия*
[2] *Институт программных систем им. А. К. Айламазяна РАН, с. Веськово, Ярославская обл., Россия*

[1]durnev@elibrary.ru, [2]ekaterin@mi-ras.ru, [3]pechnikov@krc.karelia.ru, [4]tche@mi-ras.ru

Аннотация

Научное соавторство является непосредственным отражением научного сотрудничества. Зарубежные исследования, выполненные на основе данных Web of Science и Scopus, показывают, что на протяжении последних десятилетий наблюдается рост числа соавторов научных публикаций в международных журналах в различных дисциплинах.

В работе проведено сравнение тенденций роста числа соавторов по данным РИНЦ и Scopus для пяти тематических областей (химия, история, математика, медицина и физика) за период с 2000 по 2020 г. Получены схожесть тенденций научного соавторства в случаях публикаций по истории и математике и заметное различие по остальным научным направлениям.

Ключевые слова*: публикация, соавторство, количество соавторов статьи, РИНЦ, Scopus.*

ВВЕДЕНИЕ

Научное соавторство статьи – это формальное признание двух или более исследователей в качестве равноправных создателей научной публикации, осно-

ванное на их существенном интеллектуальном вкладе в ее содержание и разделяемой ответственности за достоверность представленных результатов. Именно в таком понимании первым фактом соавторства можно считать статью математиков Клейна и Ли 1870 г. [1].

Научное соавторство является непосредственным отражением научного сотрудничества и является частым объектом для исследований из-за очевидности определения «знакомства» ученых. Согласно [2], соавторство определяется так: «…два ученых считаются связанными, если они совместно написали статью. Это кажется разумным определением научного знакомства: большинство людей, которые написали статью вместе, будут хорошо знать друг друга».

В 2024 г. сотрудниками Научной электронной библиотеки, Математического института им. В.А. Стеклова и Института прикладных математических исследований Карельского научного центра Российской академии наук было проведено совместное исследование динамики соавторства в России по пяти тематическим областям (химии, истории, математики, медицине и физике), источником данных для которого был Российский индекс научного цитирования (РИНЦ, https://elibrary.ru/). Это исследование базировалось на ограниченном количестве журналов, что позволило оценить, насколько хорошо работают процедуры сбора данных и взаимодействия участников, а также скорректировать цели и задачи дальнейших исследований. Полученные результаты исследования опубликованы в работе [3]. Кратко основной содержательный вывод можно сформулировать так: за период с 2000 по 2020 г. в российских научных публикациях наблюдаются тенденции постоянного роста среднего числа соавторов и доли статей, написанных в соавторстве с достаточно четко выраженными различиями между научными направлениями.

В зарубежных исследованиях, начиная с работы [4], рост количества соавторов вызывает следующую озабоченность: что именно значит упоминание имени среди соавторов; какую роль могут играть соавторы в самом исследовании, анализе результатов и подготовке текстов [5–7]. Проведенное нами исследование не выявило подобной озабоченности в России. Научные направления, традиционно требующие индивидуальной работы, показывают незначительный рост соавтор-

ства. Более значительный рост, наблюдаемый в естественно-научных направлениях и медицине, вполне объясним усложнением характера исследований и расширением научных границ, что и ведет к увеличению состава научных коллективов.

Естественно, что нам хотелось бы сравнить тенденции изменения соавторства в российских журналах с соответствующими результатами для зарубежных журналов. Материал для такого сравнения дает исследование, проведенное английскими коллегами по данным Scopus, основные результаты которого изложены в [8], а наборы данных, промежуточные результаты обработки в табличном виде и графики представлены в [9].

В настоящей статье изложены результаты проведенного сравнения тенденций соавторства в российских журналах по пяти перечисленным выше тематическим областям с аналогичными или близкими научными направлениями, отраженными в Scopus, по данным [8, 9].

ИСХОДНЫЕ ДАННЫЕ ИССЛЕДОВАНИЯ

Источником данных для исследования российского фрагмента сравнения является библиографическая база данных «Российский индекс научного цитирования» (РИНЦ), «…аккумулирующая более 12 миллионов публикаций российских авторов, а также информацию о цитировании этих публикаций из более 6000 российских журналов» [10].

На начало декабря 2025 г. в РИНЦ представлена информация о более 16 тыс. российских научных журналах, из которых почти 13 тыс. издаются в настоящее время, а 6509 имеют полные тексты на этом портале. Журналы классифицированы по 70 тематикам, определяемым по рубрикам верхнего уровня Государственного рубрикатора научно-технической информации (ГРНТИ) [11]. Количество журналов, в зависимости от тематики, существенно различается (от 1590 по медицине и здравоохранению до 18 по стандартизации).

Были выбраны пять тематик по существенно разным направлениям науки, представленных в РИНЦ большим количеством российских журналов, выходящих в настоящее время (в скобках указано количество журналов): Химия (181), История. Исторические науки (680), Математика (250), Медицина и здравоохранение (1266), Физика (239).

Для каждой тематики было отобрано 20 журналов, отвечающих следующему ряду условий: год основания не позднее 2000, наличие полностью проиндексированного архива в РИНЦ с 2005 по 2020 г., присутствие в рейтинге Science Index РИНЦ [12], включение в Russian Science Citation Index (RSCI) [13]. Полные списки из 20 журналов по каждой тематике приведены в [14].

Сводные итоги по выбранным 20 журналам каждой тематики из базы данных РИНЦ за период с 2000 по 2020 г. таковы:

- Химия: 70.6 тыс. статей / 37.2 тыс. авторов;
- История. Исторические науки: 34.2 / 9.9;
- Математика: 23.8 / 9.3;
- Медицина и здравоохранение: 52.7 / 40.6;
- Физика: 72.8 / 43.1.

М. Телвалл и Н. Мафлахи оценили изменения в частоте соавторства в журнальных статьях за 1900–2020 гг. по рубрикам научных областей Scopus [8]. Ими отмечено, что начиная с 1900 г. число соавторов возросло как в крупных научных направлениях, так и в более узких областях, с существенными различиями этого показателя между ними.

В [8] сказано, что «…исходными данными для этого исследования являются записи метаданных всех документов Scopus типа journal article, опубликованных в период с 1900 по 2020 г., загруженных с помощью Scopus API до сентября 2021 г.». Охвачено 88 млн журнальных статей по 27 рубрикам первого уровня классификатора и 332 рубрикам второго уровня Scopus [15].

Для нас было важно, что в качестве приложения к статье [8] на интернет-ресурсе [9] приведены данные о среднем количестве авторов на одну статью для 27 рубрик первого уровня и большого количества рубрик второго уровня Scopus по годам за весь период с 1900 по 2020 г. Очевидно, что сравнение тенденций соавторства по РИНЦ и Scopus возможно лишь за период с 2000 по 2020 г.

В исследовании Тевалл и Мафлахи в качестве среднего было взято среднее геометрическое, поскольку оно «…является более подходящим показателем центральной тенденции для сильно искаженных данных… что важно из-за наличия нескольких крупных команд» [8, с. 334].

В нашем исследовании по данным РИНЦ мы находили среднее арифметическое количество авторов как величину, более естественную для содержательной интерпретации. Однако из-за отсутствия у нас исходных данных Scopus наши результаты из [3] были пересчитаны как средние геометрические значения, и сравнение проводилось по ним.

Более сложным оказался вопрос о сопоставлении тематик РИНЦ и рубрик Scopus. Как уже было сказано выше, в РИНЦ тематика определяется по рубрикам верхнего уровня системы классификации ГРНТИ, а в Scopus используется классификатор All Science Journal Classification (ASJC) Научного издательского дома Elsevier.

Сопоставление различных классификационных систем научной и технической информации является отдельной задачей [16], далеко выходящей за рамки настоящей работы.

В качестве примера приведем медицину. В ГРНТИ рубрика первого уровня называется «Медицина и здравоохранение» и содержит девять рубрик второго уровня, а в Scopus рубрика “Medicine” содержит 48 рубрик второго уровня, причем в ГРНТИ «Фармакология» входит в «Медицину и здравоохранение», а «Pharmacology, Toxicology and Pharmaceutics» в Scopus является рубрикой первого уровня.

Для сопоставления классификаторов в нашей задаче был использован электронный ресурс [17]. Результаты сопоставления, используемые далее, представлены в табл. 1.

Табл. 1. Сопоставление тематик РИНЦ и рубрик Scopus

ГРНТИ (тематика РИНЦ)	**Scopus**
31 Химия	1600 Chemistry
03 История. Исторические науки	*1200 Arts and Humanities* 1202 History 1204 Archaeology
27 Математика	2600 Mathematics 2603 Analysis
76 Медицина и здравоохранение	2700 Medicine
29 Физика	3100 Physics and Astronomy

И хотя для химии, математики, медицины и физики сопоставления рубрик первого уровня представляются достаточно естественными, мы также использовали для математики рубрику второго уровня “2603 Analysis”.

Не столь простая ситуация складывается по истории: в Scopus рубрика второго уровня “1202 History” входит в рубрику “1200 Arts and Humanities”, которая, очевидно, неадекватна рубрике “03 История. Исторические науки”. По этой причине для сравнения были взяты две рубрики Scopus: “1202 History” и “1204 Archaeology”, а не “1200 Arts and Humanities” (в табл. 1 выделена курсивом).

СРАВНЕНИЕ ТЕНДЕНЦИЙ СОАВТОРСТВА ПО ДАННЫМ РИНЦ И SCOPUS

На рис. 1 приведены графики изменения по годам среднего геометрического количества соавторов по химии, медицине и физике. Обозначения графиков очевидны, к примеру, “chem eLib” соответствует графику по химии по данным РИНЦ, а “Chem Scop” – графику по химии по данным Scopus. Графики по данным РИНЦ показаны сплошными линиями, а графики по данным Scopus – пунктирными.

На рис. 2 представлены аналогичные графики по истории и математике. Отличие от рис. 1 заключается в том, что в каждом случае сравниваются по три графика: два по данным Scopus и один по РИНЦ.

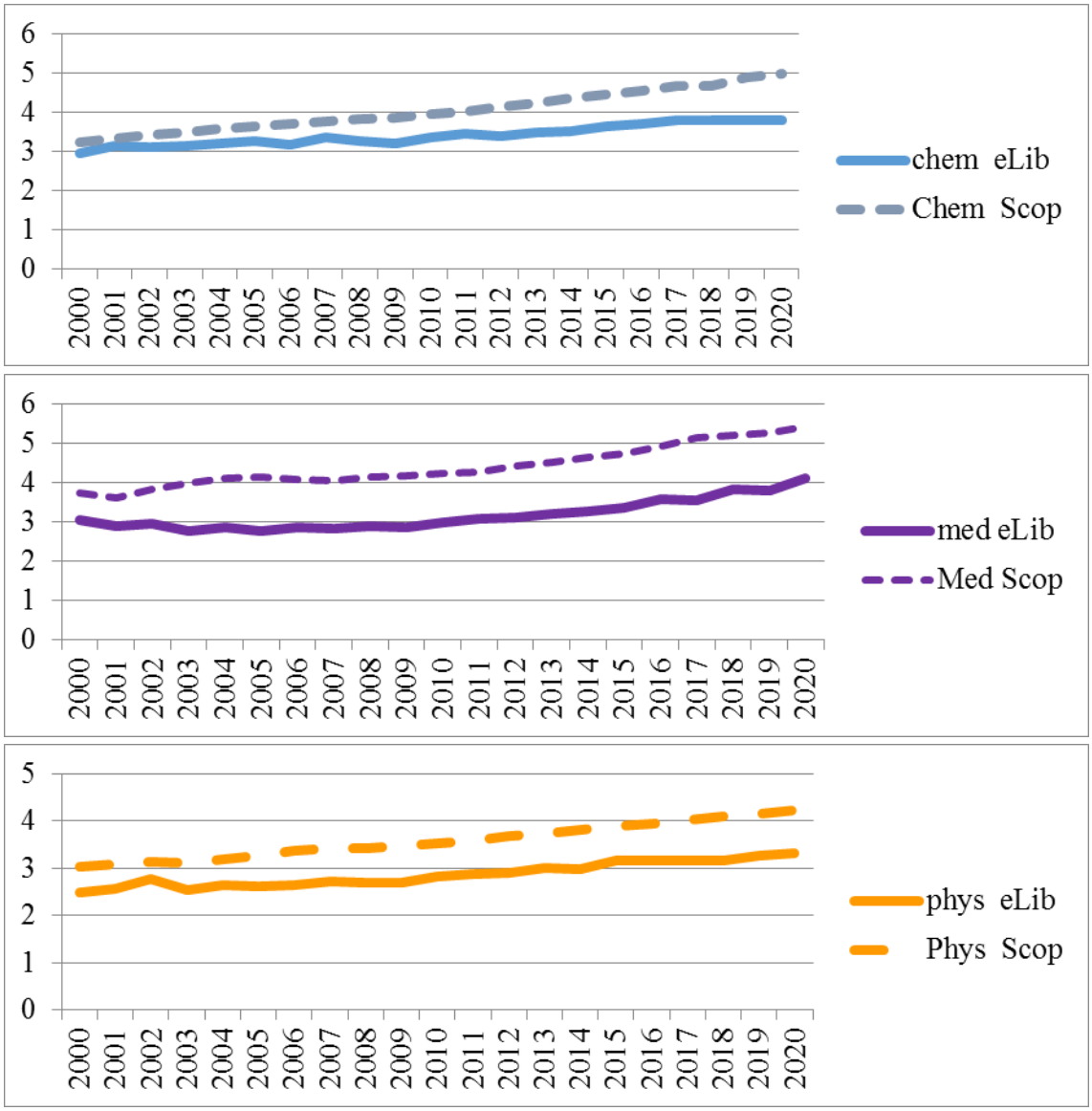


Рис. 1. Графики изменения среднего геометрического количества авторов по РИНЦ и Scopus (Химия, Медицина и здравоохранение, Физика)

Сравнение тенденций в изменении соавторства в российских и зарубежных журналах показывает, что научные направления, требующие больших коллективов исследователей, отличаются довольно существенно. Например, в статьях по химии среднее геометрическое количество авторов по Scopus примерно в 1.15 раз больше, чем такое же значение по РИНЦ. Учитывая, что среднее арифметическое всегда больше или равно среднего геометрического, примерная разница в среднем арифметическом количестве соавторов в группе составляет 1.35 в пользу Scopus. Те же тенденции выявлены в медицине и физике.

В то же время тенденции соавторства в исторических и математических статьях практически идентичны для тематик РИНЦ и родственных рубрик второго уровня Scopus, но существенно отличаются от рубрик первого уровня.

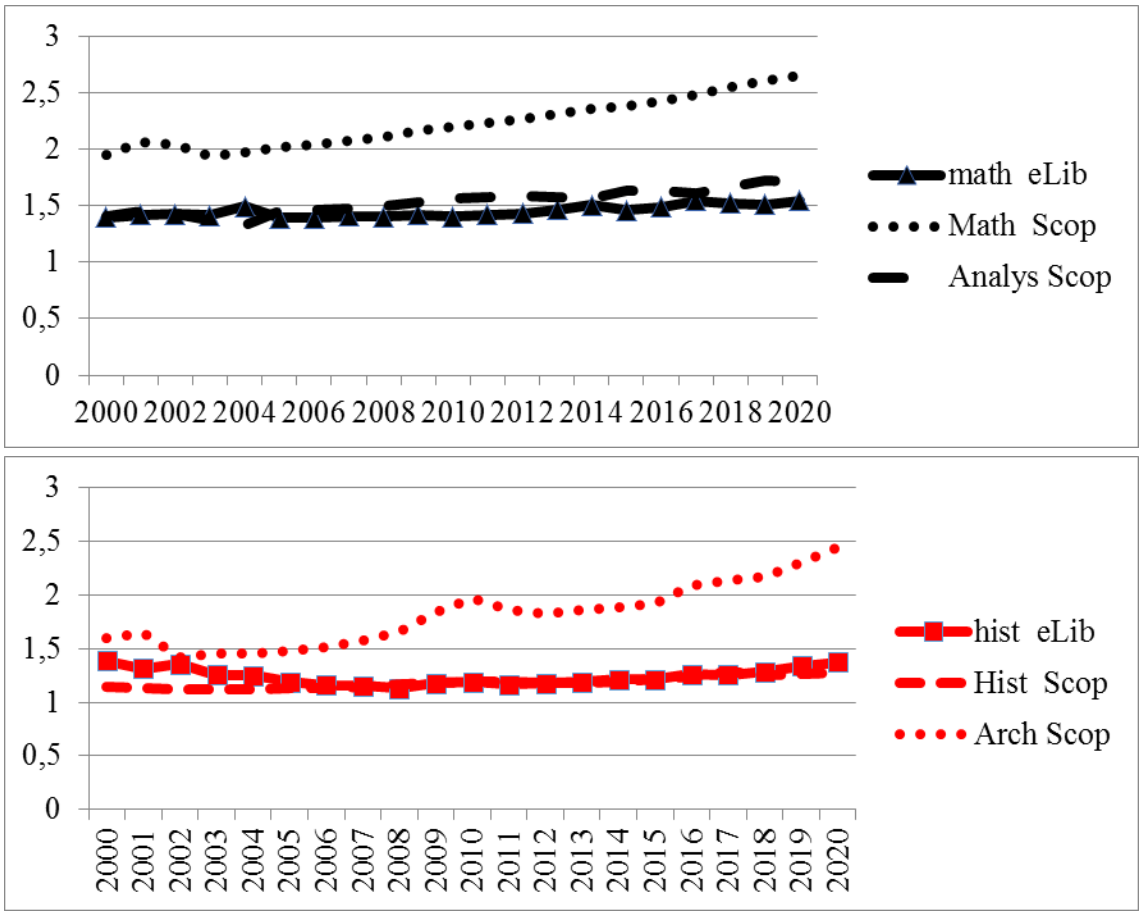


Рис. 2. Графики изменения среднего геометрического количества авторов по РИНЦ и Scopus (История. Исторические науки, Математика)

Для проверки было проведено вычисление среднего геометрического количества соавторов в статьях из двух журналов РИНЦ, которые, судя по названиям и содержанию, являются чисто археологическими, и построен график, практически аналогичный графику “Arch Scop” на рис. 2. Возможно, наличие этих журналов в списке по тематике «История. Исторические науки» задает незначительное превышение графика “hist eLib” над “Hist Scop”.

Значительное превышение графика “Math Scop” над “math eLib” вполне объяснимо тем, что в Scopus рубрика первого уровня Mathematics содержит 14 рубрик второго уровня, например, таких как Applied Mathematics и Modelling

and Simulation, где исследования выполняются группами ученых. Совпадение графиков “math eLib” и “Analys Scop” говорит о том, что журналы по математической тематике РИНЦ содержательно близки к журналам рубрики Analysis из Scopus (то есть математическому анализу).

ЗАКЛЮЧЕНИЕ

Одним из возможных объяснений различий в тенденциях роста соавторства между РИНЦ и Scopus может быть следующее. Абсолютное большинство журналов Scopus издается на английском языке, и редакции оценивают публикации, в том числе, и по качеству текста. Когда над статьей работают большие коллективы соавторов, для которых язык не является родным, неизбежно привлечение специалиста по английскому языку, имеющего достаточные знания в указанной научной области, которого, вполне возможно, включают в соавторы. Для этого в таксономии соавторства [18] есть специальная позиция “Writing – original draft”, увеличивающая количество соавторов на 1. Для историков и математиков, более склонных к индивидуальной научной работе, вопрос стоит менее остро, а для российских ученых, публикующихся в российских журналах, он отсутствует совсем.

Анализ последствий внедрения оценок научной деятельности в России показывает распространение нечестных практик соавторства, в частности включение в состав авторов людей, не имеющих отношения к исследованию [19–21]. Заметим, что указанные работы во многом опираются на данные Web of Science и Scopus. Проведенное исследование, показывающее «эволюционный» рост соавторства, исключает подобные практики в 80 исследованных журналах РИНЦ. Вполне возможно (по условиям отбора), журналы, отобранные для данного исследования, относятся по терминологии [22] к «высококачественному сегменту публикаций», а в «сегменте публикаций более низкого качества» тенденции роста соавторства будут совсем другими.

Результаты, опубликованные в настоящей работе, не содержат персональных данных, в том числе уникальных кодов, по которым можно идентифицировать авторов.

Благодарности

Работа выполнена при частичном финансировании по проекту FMEN-2024-0005 «Случайные графы, структура и информационный поиск, кооперация и конкуренция в сетях и приложения в сложных системах».

СПИСОК ЛИТЕРАТУРЫ

1. *Klein F., Lie S.* Sur une certaine famille de courbes et de surfaces // Comptes rendus de l'Académie des sciences. 1870. P. 1222–1226, 1275–1279.

2. *Newman M.E.J.* The structure of scientific collaboration networks // Proceedings of the National Academy of Sciences of the USA. 2001. Vol. 98. No. 2. P. 404–409. https://doi.org/10.1073/pnas.98.2.404

3. *Дурнев С.А., Знаменская Е.А., Печников А.А., Чебуков Д.Е.* Сравнительный анализ научного соавторства в России // Библиосфера. 2025. № 3. С. 110–119. https://doi.org/10.20913/1815-3186-2025-3-110-119

4. *Cronin B.* Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? // Journal of the American Society for Information Science and Technology. 2001. Vol. 52. No 7. P. 558–569. https://doi.org/10.1002/asi.1097

5. *Clement T.P.* Authorship matrix: a rational approach to quantify individual contributions and responsibilities in multi-author scientific articles // Science and Engineering Ethics. 2014. Vol. 20. No. 2. P. 345–361. http://dx.doi.org/10.1007/s11948-013-9454-3

6. *Resnik D.B., Tyle A.M., Black J.R., Kissling G.* Authorship policies of scientific journals // Journal of Medical Ethics. 2016. Vol. 42. No. 3. P. 199–202. https://doi.org/10.1136/medethics-2015-103171

7. *Osborne J.W., Holland A.* What is authorship, and what should it be? A survey of prominent guidelines for determining authorship in scientific publications // Practical Assessment, Research, and Evaluation. 2019. Vol. 14. No. 1. Article Number 15. 19 pp. https://doi.org/10.7275/25pe-ba85

8. *Thelwall M., Maflahi N.* Research coauthorship 1900–2020: Continuous, universal, and ongoing expansion // Quantitative Science Studies. 2022. Vol. 3. No. 2. P. 331–344. https://doi.org/10.1162/qss_a_00188

9. Research Co-authorship 1900–2020: Continuous, universal, and ongoing expansion. URL: https://figshare.com/articles/dataset/Research_Co-authorship_1900-2020_Continuous_universal_and_ongoing_expansion/17064419.

10. Российский индекс научного цитирования (РИНЦ). URL: https://elibrary.ru/project_risc.asp.

11. Коды ГРНТИ. URL: https://грнти.рф/kody-grnti.

12. О новом рейтинге журналов Science Index. URL: https://elibrary.ru/projects/science_index/ranking_info.asp.

13. Список журналов, входящих в базу данных RSCI. URL: https://www.elibrary.ru/projects/rsci/rsci.pdf.

14. Appendix 1. URL: https://homepage.mi-ras.ru/~tche/download/appendix1.xlsx.

15. Классификация ASJC в Scopus. URL: https://научныепереводы.рф/classification_asjc_scopus.

16. *Антопольский А.Б., Белоозеров В.Н., Маркарова Т.С., Дмитриева Е.Ю.* Установление соответствий рубрик ГРНТИ рубрикам других систем классификации научной и технической информации // Научно-техническая информация. Сер. 1. Организация и методика информационной работы. 2015. № 3. С. 3–19.

17. Сеть классификационных систем. URL: http://rffi20.viniti.ru.

18. NISO. CRediT, Contributor Roles Taxonomy. URL: https://groups.niso.org/higherlogic/ws/public/download/26466/ANSI-NISO-Z39.104-2022.pdf.

19. *Губа К.С., Словогородский Н.А.* Publish or Perish в российских социальных науках: паттерны соавторства в «хищных» и «чистых» журналах // Вопросы образования. 2022. № 4. С. 80–106. https://doi.org/10.17323/1814-9545-2022-4-80-106

20. *Gureev V.N., Lakizo I.G., Mazov N.A*. Unethical authorship in scientific publications (a review of the problem) // Scientific and Technical Information Processing. 2019. Vol. 46. № 4. P. 219–232. https://doi.org/10.3103/S0147688219040026

21. *Ефимова Г.З.* Соавторство или соло-авторство: соблюдение традиций или свободный выбор? // Социология науки и технологий. 2022. Т. 13. №1. С. 130–148. https://doi.org/10.24412/2079-0910-2022-1-130-148

22. *Матвеева Н.Н.* Библиометрический анализ взаимодействия ученых в

российских вузах: кооперация vs индивидуальная продуктивность // Университетское управление: практика и анализ. 2020. Т. 24. №. 2. С. 26–43.
https://doi.org/10.15826/umpa.2020.02.012

SCIENTIFIC CO-AUTHORSHIP ACCORDING TO RSCI AND SCOPUS DATA FOR 2000-2020: GROWTH TRENDS

S. A. Durnev[1] [0009-0005-0338-5430], E. A. Znamenskaya[2] [0000-0003-3630-712X], A. A. Pechnikov[3] [0000-0002-0683-0019], D. E. Chebukov[4] [0000-0001-9738-8707]

[1]*Scientific Electronic Library (eLIBRARY.RU), Moscow, Russia*
[2, 4]*Steklov Mathematical Institute of RAS, Moscow, Russia*
[3]*Institute of Applied Mathematical Research of the Karelian Research Centre of RAS, Petrozavodsk, Russia*
[2]*Ailamazyan Program Systems Institute of RAS, s. Ves'kovo, Yaroslavl Oblast, Russia*
[1]durnev@elibrary.ru, [2]ekaterin@mi-ras.ru, [3]pechnikov@krc.karelia.ru, [4]tche@mi-ras.ru

Abstract

Scientific co-authorship is a direct reflection of scientific collaboration. Foreign studies based on Web of Science and Scopus data show that over the past decades there has been an increase in the number of co-authors of scientific publications in international journals in various disciplines. The paper compares the growth trends in the number of co-authors according to the RSCI and Scopus data. The study was conducted in five thematic areas (chemistry, history, mathematics, medicine, and physics) from 2000 to 2020. The article shows the identity of the trends in the growth of the number of co-authors in the cases of publications on history and mathematics, and a noticeable difference in other scientific fields.

Keywords*: publication, co-authorship, number of co-authors of the article, RSCI, Scopus.*

REFERENCES

1. *Klein F., Lie S.* Sur une certaine famille de courbes et de surfaces // Comptes rendus de l'Académie des sciences. 1870. P. 1222–1226, 1275–1279.

2. *Newman M.E.J.* The structure of scientific collaboration networks // Proceedings of the National Academy of Sciences of the USA. 2001. Vol. 98. No. 2. P. 404–409. https://doi.org/10.1073/pnas.98.2.404

3. *Durnev S.A., Znamenskaya E.A., Pechnikov A.A., Chebukov D.E.* Comparative analysis of scientific co-authorship in Russia // Bibliosphere. 2025. No. 3. P. 110–119. https://doi.org/10.20913/1815-3186-2025-3-110-119

4. *Cronin B.* Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? // Journal of the American Society for Information Science and Technology. 2001. Vol. 52. No. 7. P. 558–569. https://doi.org/10.1002/asi.1097

5. *Clement T.P.* Authorship matrix: a rational approach to quantify individual contributions and responsibilities in multi-author scientific articles // Science and Engineering Ethics. 2014. Vol. 20. No. 2. P. 345–361. http://dx.doi.org/10.1007/s11948-013-9454-3

6. *Resnik D.B., Tyle A.M., Black J.R., Kissling G.* Authorship policies of scientific journals // Journal of Medical Ethics. 2016. Vol. 42. No. 3. P. 199–202. https://doi.org/10.1136/medethics-2015-103171

7. *Osborne J.W., Holland A.* What is authorship, and what should it be? A survey of prominent guidelines for determining authorship in scientific publications // Practical Assessment, Research, and Evaluation. 2019. Vol. 14. No. 1. Article Number 15. 19 pp. https://doi.org/10.7275/25pe-ba85

8. *Thelwall M., Maflahi N.* Research coauthorship 1900–2020: Continuous, universal, and ongoing expansion // Quantitative Science Studies. 2022. Vol. 3. No. 2. P. 331–344. https://doi.org/10.1162/qss_a_00188

9. Research Co-authorship 1900-2020: Continuous, universal, and ongoing expansion. URL: https://figshare.com/articles/dataset/Research_Co-authorship_1900-2020_Continuous_universal_and_ongoing_expansion/17064419.

10. Rossiiskiy indeks nauchnogo citirovaniya. URL: https://elibrary.ru/project_risc.asp.

11. Kody GRNTI. URL: https://грнти.рф/kody-grnti.

12. O novom reitinge zhurnalov Science Index.

URL: https://elibrary.ru/projects/science_index/ranking_info.asp.

13. Spisok zhurnalov, vhodyaschih v bazu dannyh RSCI.

URL: https://www.elibrary.ru/projects/rsci/rsci.pdf.

14. Appendix 1.

URL: https://homepage.mi-ras.ru/~tche/download/appendix1.xlsx.

15. Klassifikaciya ASJC v Scopus.

URL: https://научныепереводы.рф/classification_asjc_scopus.

16. *Antopol'skii A.B., Belozerov V.N., Markarova T.S., Dmitrieva E.Yu.* Ustanivlenie sootvetstvii rubric GRNTI rubrikam drugih system klassificacii // Nauchno-tehnicheskaya informaciya. Ser. 1. Organizaciya i metodika informacionnoi raboty. 2015. No. 3. S. 3–19.

17. Set' klassificacionnyh system. URL: http://rffi20.viniti.ru.

18. NISO. CRediT, Contributor Roles Taxonomy.

URL: https://groups.niso.org/higherlogic/ws/public/download/26466/ANSI-NISO-Z39.104-2022.pdf.

19. *Guba K.S., Slovogorodskii N.A.* Publish or Perish v rossiiskih social'nyh naukah: pattern soavtorstva v "hischnyh" i "chistyh" zhurnalah // Voprosy obrazovaniya. 2022. No. 4. S. 80–106. https://doi.org/10.17323/1814-9545-2022-4-80-106

20. *Gureev V.N., Lakizo I.G., Mazov N.A.* Unethical authorship in scientific publications (a review of the problem) // Scientific and Technical Information Processing. 2019. Vol. 46. No. 4. P. 219–232. https://doi.org/10.3103/S0147688219040026

21. *Efimova G.Z.* Soavtorstvo ili solo-avtorstvo: soblyudenie tradicii ili svobodnyi vybor? // Sotsiologiya nauki i tekhnologiy. 2022. T. 13. No. 1. S. 130–148. https://doi.org/10.24412/2079-0910-2022-1-130-148

22. *Matveeva N.N.* Bibliometricheskii analiz vzaimodeistviya uchenyh v rossiiskih vyzah: kooperaciya vs individual'naya produktivnost' // Universitetskoe upravlenie: Praktika i analiz. 2020. T. 24. No. 2. S. 26–43. https://doi.org/10.15826/umpa.2020.02.012

СВЕДЕНИЯ ОБ АВТОРАХ

ДУРНЕВ Сергей Андреевич – старший программист Отдела разработок Научной электронной библиотеки. Сфера научных интересов – электронные библиотеки, библиометрия, наукометрия.

Sergey Andreevich DURNEV – Senior Programmer at the Development Department, Scientific Electronic Library (eLIBRARY.RU). Research interests include digital libraries, bibliometrics, scientometrics.

email: durnev@elibrary.ru

ORCID: 0009-0005-0338-5430

ЗНАМЕНСКАЯ Екатерина Александровна – ведущий программист Отдела компьютерных сетей и информационных технологий Математического института им. В.А. Стеклова Российской академии наук. Младший научный сотрудник ИЦИИ Института программных систем им. А. К. Айламазяна Российской академии наук. Сфера научных интересов – электронные библиотеки, технология разметки библиографии, библиометрия.

Ekaterina Aleksandrovna ZNAMENSKAYA – Leading programmer at the Department of Computer Networks and Information Technology, Steklov Mathematical Institute of Russian Academy of the Sciences. Reseacher at the Ailamazyan Program Systems Institute of the Russian Academy of Sciences. Research interests include digital libraries, bibliographic tagging technology, bibliometrics.

email: ekaterin@mi-ras.ru

ORCID: 0000-0003-3630-712X

ПЕЧНИКОВ Андрей Анатольевич – ведущий научный сотрудник Института прикладных математических исследований обособленного подразделения ФИЦ «Карельский научный центр Российской академии наук», доцент, д. т. н. Сфера научных интересов – математическое моделирование, графы и сети, дискретная оптимизация, вебометрика, наукометрия.

Andrey Anatolievich PECHNIKOV – Leading Researcher, Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, Associate Professor, Doctor (DSc) of Technics. Research interests include mathematical modelling, graphs and nets, discrete optimization, webometrics, scientometrics.

email: pechnikov@krc.karelia.ru

ORCID: 0000-0002-0683-0019

ЧЕБУКОВ Дмитрий Евгеньевич – зав. Информационно-издательским сектором Математического института им. В. А. Стеклова Российской академии наук, к. х. н. Сфера научных интересов – библиометрия, наукометрия, электронные библиотеки.

Dmitry Evgen'evich CHEBUKOV – Head of the Information and Publishing Sector, Steklov Mathematical Institute of the Russian Academy of Sciences, Candidate Chem. Sci. Research interests include bibliometrics, scientometrics, digital libraries.

email: tche@mi-ras.ru

ORCID: 0000-0001-9738-8707

Материал поступил в редакцию 16 января 2026 года

УДК 004.65(075.8)+004.89

ЗАПРОСЫ К НЕРЕЛЯЦИОННЫМ ДАННЫМ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ НА ОСНОВЕ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ

А. О. Еркимбаев[1] [0000-0002-5239-2208], В. Ю. Зицерман[2] [0000-0003-3327-3139], Г. А. Кобзев[3] [0000-0001-9987-1823]

[1-3]*Объединенный институт высоких температур РАН, г. Москва, Россия*

[1]adilbek@jiht.ru, [2]vz1941@mail.ru, [3]gkbz@mail.ru

Аннотация

В работе рассмотрены новые возможности организации запросов на естественном языке к научным локальным базам данных нереляционного типа. Проведенный анализ исследований, выполненных за последние годы, показал активное внедрение запросов на естественном языке к базам данных различного типа. Отмечено активное применение методов машинного обучения (нейронных алгоритмов). Показано широкое использование в последние два года большой языковой модели для подготовки запросов в различных языковых средах и областях знаний. Проведено исследование новых возможностей графовой базы данных AllegroGraph по использованию больших языковых моделей для организации поиска на естественном языке. Функционал базы данных изучен на примере системы метаданных по теплофизическим свойствам веществ в форме предметной онтологии «Термаль». Тестирование поисковых запросов в двуязычной (английская и русская) среде базы данных выявило в целом преодолимые проблемы и дает хорошие надежды на дальнейшее применение новых прикладных сервисов с использованием больших языковых моделей.

Ключевые слова*: запрос на естественном языке, большая языковая модель, эмбеддинг, нереляционные базы данных, графовая база данных, онтология предметной области.*

ВВЕДЕНИЕ

Работа посвящена расширению функциональности баз данных (БД), работающих при поддержке средств искусственного интеллекта (ИИ). Эта проблема изучается последние годы с преимущественной ориентацией на реляционные (SQL) БД с их хорошо структурированной системой хранения и поиска. Наш опыт по систематизации естественно-научных данных показал существование потребности в полуструктурированных данных с изменчивой структурой, что требует перехода на нереляционные (NoSQL) БД [1, 2].

Активное внедрение средств ИИ открывает новые возможности при работе с подобного рода данными, в частности в организации поиска. Одной из них является возможность организации запросов на естественном языке (ЕЯ), что избавляет пользователя от необходимости точного знания классификации и лексики предметной области, а также сложных правил составления поисковых запросов для NoSQL БД.

Применение запросов на ЕЯ к хранилищам данных имеет давнюю историю и восходит к информационной системе, часто упоминаемой как LUNAR [3], созданной в 1972 г. Полное наименование системы “The Lunar Sciences Natural Language Information System”. Она была предназначена для обычных пользователей и отвечала на вопросы о химическом анализе лунных пород, полученных с «Аполлона-11». Система состояла из трех основных компонентов: формальной грамматики общего назначения, синтаксического анализатора для большого подмножества естественного английского языка и компоненты семантической интерпретации, управляемой правилами.

ОБЗОР СОВРЕМЕННЫХ ПОДХОДОВ

На примере ряда работ установлены общие принципы составления запросов на ЕЯ. Так, авторы [4] подчеркивают, что при решении этой задачи всегда требуются: а) преобразование предложений на ЕЯ в информацию, пригодную для машинной обработки; б) транслирование запросов к БД, составленных на основании информации, извлеченной из текста. Преобразование текстов на ЕЯ предложено выполнять методами компьютерной лингвистики в полуавтоматическом или автоматическом режимах с применением графематического, морфологического и синтаксического анализов [5]. Итог реализации данного шага –

синтаксическое дерево с определенным набором объектов и свойств, которое впоследствии может быть преобразовано в запросы на формальном языке. Классифицируя существующие БД по их типу как дореляционные, реляционные и постреляционные (NoSQL), авторы [4] предложили наиболее реализуемую и оптимальную организацию запросов на ЕЯ к реляционным и графовым БД (которые относятся к NoSQL). С учетом выделенных вариантов БД ими указаны два пути обработки полученного синтаксического дерева: а) с помощью семантических сетей, отражающих связи между объектами; б) с помощью средств логического программирования, например Prolog. В дальнейшем результаты [4] были реализованы в системе запросов на ограниченном ЕЯ к реляционным БД [6].

Особенность другой работы [7] состояла в том, что использовалась семантическая модель БД как на этапе формирования естественно-языкового интерфейса, так и в ходе его эксплуатации. Первоначальный процесс обработки естественно-языкового запроса пользователя состоял из последовательного выполнения анализа. Следующий шаг обработки запроса на ЕЯ заключался в построении его морфологического, синтаксического и семантического представлений. При этом семантическое представление естественно-языкового запроса пользователя строилось на основе семантической модели БД, имеющей обязательное текстовое представление, доступное для ручной правки экспертами или машинной обработки. На основе этого представления в дальнейшем формировался SQL-запрос к БД.

Отметим, что работы, названные выше, не использовали методы машинного обучения (например, нейронные алгоритмы) для формирования запросов и делали акцент на семантические модели и логическое программирование.

Запрос на ЕЯ и выделенные после его разбора объекты синтаксического и семантического анализа (слова, части речи и предложения) представляют собой категориальный (нечисловой) тип данных. Как правило, для дальнейшей их обработки компьютером требуется их векторизация – преобразование текста в числовой формат, который могут понимать и обрабатывать программы, например алгоритмы машинного обучения. Активное применение нейронных сетей привело к созданию методами машинного обучения процедур векторизации текста [8]. Эта процедура преобразования текста получила название текстового *эмбеддинга* (text embeddings), или «встраивания текста». При этом различают

векторизацию слова (word embedding) или предложений (sentence embeddings). Существуют эмбеддинги иных типов данных (например, изображений, графовых структур и т. д.). К наиболее распространенным сейчас видам текстовых эмбеддингов в машинном обучении относят Word2Vec [8], Glove [9] и BERT [10].

Развитие методов глубокого обучения в настоящее время привело к тому, что встраивание текста или эмбеддинг стало основополагающей технологией в области обработки ЕЯ, способствующей прогрессу в решении множества последующих задач, связанных с языком. При этом по-прежнему одной из важных задач является определение семантического сходства текстов, что требует векторизации (встраивания) текста для вычисления сходства. В таких технологиях важную роль играют размеры и качество текстовых массивов или корпусов при организации обучения. На их основе создано множество фреймворков с готовыми решениями для создания запросов данных. Так, выполненный в обзоре [11] анализ 35 фреймворков, разработанных в период с 2008 по 2018 г., учитывал поддержку языка, эвристические правила, функциональную совместимость, объем данных и оценку производительности. Оказалось, что 70% запросов на ЕЯ было выполнено для SQL, а на долю NoSQL приходилось только 15% (SPARQL), 10% (CYPHER) и 5% (GREMLIN).

Результаты [11] трудно обобщить, поскольку тестировались они на различных БД и в разных предметных областях. Но при этом следует подчеркнуть, что подавляющее большинство фреймворков, рассмотренных в [11], выполняет запросы на ЕЯ к данным на английском языке, исключение составил лишь один продукт, работающий с арабским языком.

В работе [12] представлен обзор инструментов (фреймворков, платформ) по преобразованию запросов на ЕЯ к формату SQL с применением нейронных сетей, в котором приведены результаты тестирования на единой реляционной БД SPIDER [13]. Было проведено тестирование девяти моделей формирования запросов на ЕЯ. Наиболее эффективной моделью оказалась RAT-SQL, объединенная с методом векторизации BERT [10], которая обеспечивала точность 65.6% при выполнении междоменных SQL-запросов. Остальные модели обеспечивали точность формирования запросов от 30% до 50%. Важно отметить, что все методы организации запросов на ЕЯ, рассмотренные в обзорах [11, 12], были также созданы для работы с английским языком.

ПРИМЕНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ЗАПРОСОВ ДАННЫХ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Рассмотренные выше источники не отражают тенденции последних 3–5 лет, связанные с применением больших языковых моделей (Large Language Model, LLM [14]) для выполнения запросов к БД на ЕЯ. По типу это модели глубокого обучения, которые обучены на огромных объемах данных и содержат более миллиарда параметров. Благодаря экстремальному числу параметров они реализуют возможности распознавать, переводить, прогнозировать или генерировать текст, как и другой контент (изображения, звук и т. д.). В 2022 г. появился продукт ChatGPT [15], разработанный компанией OpenAI на базе ИИ и основанный на LLM-модели. Эта система способна работать в диалоговом режиме, отвечать на вопросы и генерировать тексты, что особенно важно, на разных языках, включая русский, относящиеся к различным областям знаний. Важной особенностью ChatGPT является также возможность генерации по запросу программ на различных языках программирования. В течение всего нескольких лет многие разработчики программных продуктов, оперирующих различными типами БД, создали собственные сервисы для запросов на ЕЯ на основе LLM. Отметим следующие из них.

- Сервис Microsoft Copilot [16] для выполнения запросов на ЕЯ к БД SQL Azure на контролируемом внешнем ресурсе Microsoft с платными и бесплатными возможностями (доступен на территории РФ только через VPN). Запросы можно выполнять на более чем 10 языках, в том числе на русском. Сервис основан на применении платформы вычислительного кластера OpenAI.
- Функциональные возможности по созданию запросов на ЕЯ к NoSQL документной базе данных MongoDb [17]. Запросы на ЕЯ на версии MongoDb Compass доступны, начиная с версии 1.40.x. В качестве текущего поставщика запросов используется Azure OpenAI.
- GraphDB предлагает палитру моделей ИИ с набором аналитических возможностей и инструментов [18]. GraphDB предоставляет инструменты LLM, использующие спецификацию OpenAI API (Application Programming Interface).

- Функционал LLM-ориентированных сервисов по созданию запросов на ЕЯ в графовой БД AllegroGraph [19]. Сервисы основаны на применении платформы вычислительного кластера OpenAI API.
- Сервис Stardog Voicebox в графовой БД Stardog [20] – интеллектуальный помощник по знаниям, работающий на базе LLM и автономных агентов для предоставления основных сервисов. Использует свою кластерную среду LLM и работает только в облачном пространстве Stardog Cloud.

Таким образом, четыре из пяти перечисленных примеров программных продуктов при использовании запросов на ЕЯ основаны на обязательном применении возможностей OpenAI [15].

Для Объединенного института высоких температур (ОИВТ) РАН с его обширной системой БД по свойствам веществ и материалов представляют особый интерес последние достижения в организации запросов к нереляционным БД с характерной для них сложной структурой, отражающей специфику предметной области. Ниже приведены результаты начальных экспериментов по реализации запросов на ЕЯ к онтологическим моделям, размещенным на платформе графовой БД AllegroGraph, указанной ранее в списке новых предлагаемых сервисов.

В ОИВТ функционирует несколько БД по теплофизическим свойствам веществ в текстовом формате ISO 2709 [21] дореляционного типа, в частности БД «Термаль». Необходимость переноса подобных данных в Интернет с реализацией на новой программно-аппаратной технологии привела нас к разработке системы метаданных в виде онтологической модели «Термаль» с применением технологий семантического веба. В качестве платформы для онтологической модели используется нереляционная графовая БД AllegroGraph, а носителем основной части данных является нереляционная БД MongoDb [17].

С целью оценки возможного применения были проведены тестовые испытания новых сервисов, предлагаемых AllegroGraph, с LLM-технологиями формирования запросов к онтологии на ЕЯ. В качестве объекта испытаний был выбран один из текущих вариантов разрабатываемой онтологии «Термаль».

После загрузки онтологической модели на серверный вариант БД AllegroGraph в облаке и для применения сервиса LLM по работе с ЕЯ был приобретен ключ доступа к прикладным функциям API кластера OpenAI. Функционал LLM для создания запросов на ЕЯ позволил провести векторизацию онтологии

«Термаль» при ее сохранении в векторном хранилище БД AllegroGraph. В настоящее время БД AllegroGraph предлагает на своей платформе два новых сервиса для создания запросов на ЕЯ с применением в нереляционной среде запросов SPARQL.

1. Сервис так называемых «магических» предикатов и функций, которые можно использовать в запросах, связанных с LLM. Магический предикат может использоваться в позиции предиката в запросах SPARQL, а функции могут преобразовывать значение переменной, что значительно расширяет возможности запросов SPARQL. Предложен ряд «магических предикатов», таких, например, как:

- llm: response: как функция, так и предикат, в зависимости от того, хотим ли мы, чтобы LLM возвращала один элемент или список элементов;
- llm:askMyDocuments: предикат высокого уровня, позволяющий запрашивать у LLM информацию о локальном хранилище векторов;
- llm:node: функция для генерации уникального URI для текстового литерала. llm:NearestNeighbor: предикат, который работает в хранилище векторов AllegroGraph. Он принимает в качестве входных данных строку или запрос и находит наилучшие совпадения в хранилище векторов;
- llm:askForTable: предикат высокого уровня, позволяющий запрашивать у LLM информацию и возвращать результаты в табличной форме.

Здесь конструкция llm: – это предопределенный префикс в системе AllegroGraph (http://franz.com/ns/allegrograph/8.0.0/llm/).

2. Сервис функции преобразования запроса на ЕЯ в запрос SPARQL – Natural Languge to SPARQL (новая функция, находящаяся еще в стадии разработки). Для использования сервиса необходимо создать специализированную векторную БД (VDB), в которой хранятся пары запросов на ЕЯ и соответствующие им SPARQL-запросы. База VDB напрямую связана с триплетным хранилищем AllegroGraph и действует как хранилище сопоставлений между тем, как пользователи могут задать вопрос на ЕЯ, и тем, как этот запрос должен быть выражен в SPARQL. По мере того как сохраняется все больше сопоставлений в VDB, сервис становится все более способным преобразовывать запросы пользователей в более точные запросы SPARQL.

В качестве примера общей работоспособности предлагаемых системой БД AllegroGraph возможностей выполнения запросов на русском («Список республик в России») и английском (“List the USA states”) ЕЯ на рис. 1 представлены результаты применения использования «магического» предиката llm:response из первого сервиса к глобальным ресурсам интернета через API функции OpenAI.

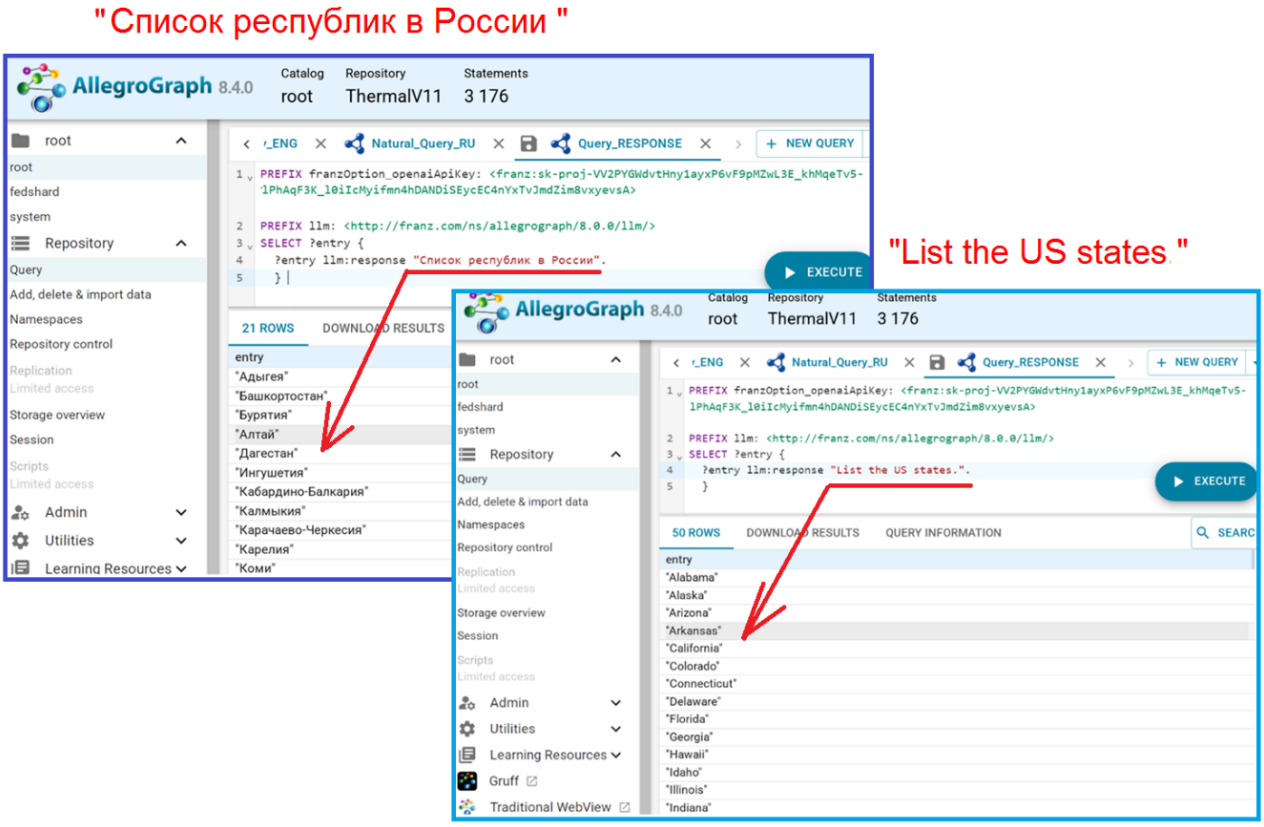


Рис. 1. Сканы интерфейсов БД AllegroGraph [19] при выполнении запросов на русском и английском ЕЯ к глобальным ресурсам интернета с использованием «магического» предиката llm:response.

Далее на рис. 2 представлены интерфейсы, демонстрирующие технологию реализации на платформе БД AllegroGraph двух разных сервисов для создания запросов на ЕЯ к локальным ресурсам через API функции OpenAI.

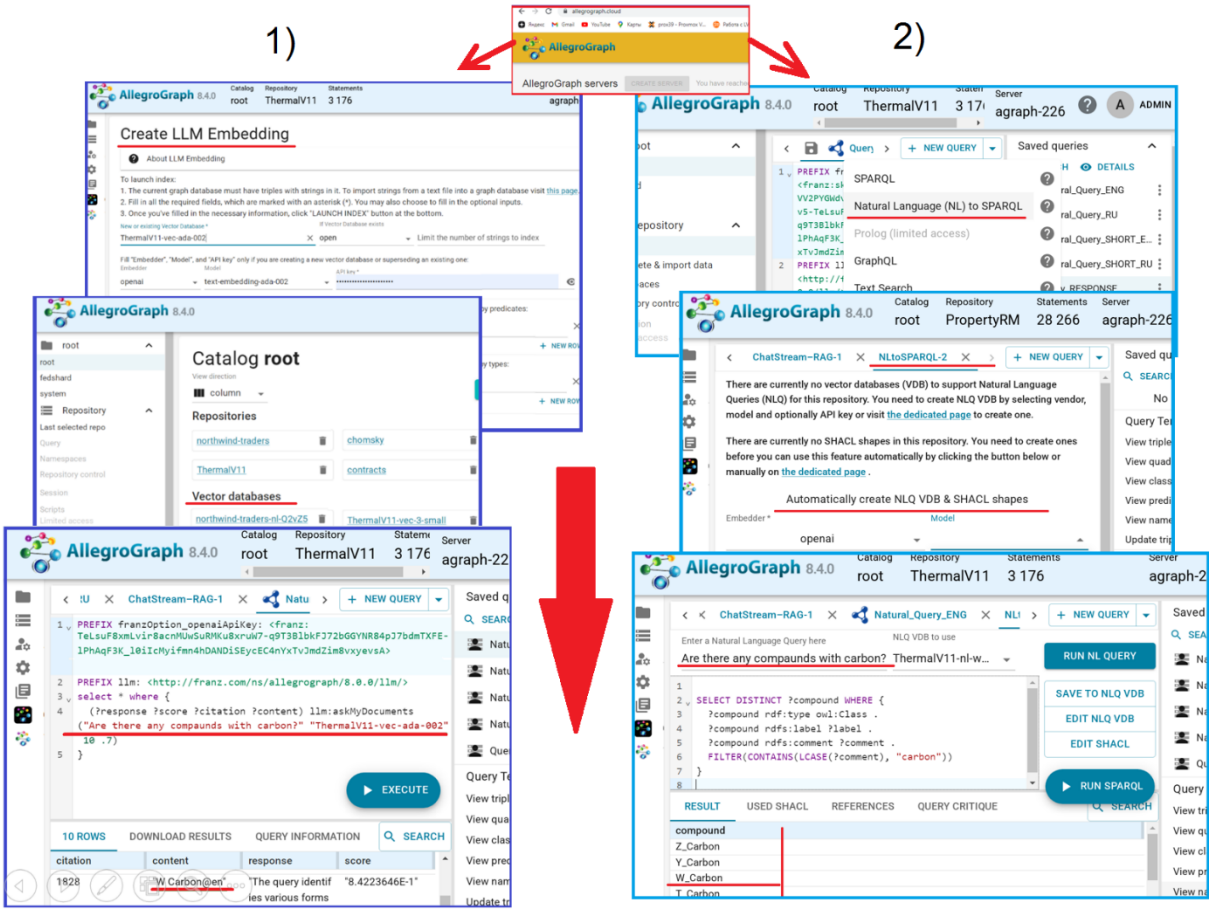


Рис. 2. Последовательности сканов интерфейсов БД AllegroGraph [19], реализующие два сервиса для создания запросов на ЕЯ к локальным ресурсам: 1) сервис «магических» предикатов; 2) сервис функции преобразования запроса на ЕЯ в SPARQL запрос.

Были выполнены тестовые запросы с использованием двух сервисов организации запросов на ЕЯ к локальной системе метаданных в форме онтологии «Термаль», загруженной в нереляционную БД AllegroGraph. В качестве проверочных данных были выбраны произвольные выражения на русском и английском языках для формирования запроса: «Есть ли соединения с углеродом?»; “Are there any compaunds with carbon?”.

Для проверки первого сервиса организации запросов на ЕЯ был использован «магический предикат» llm:askMyDocuments. В условиях запроса в атрибутах функции предписано выдать ближайшие 10 результатов с оценкой приближения не ниже 0.7 (1.0 – полное совпадение, 0.0 – отсутствие совпадения). На рис. 3 представлены интерфейсы запросов на ЕЯ с результатами запроса к векторному образу «ThermalV11-vec-ada-002» онтологии.

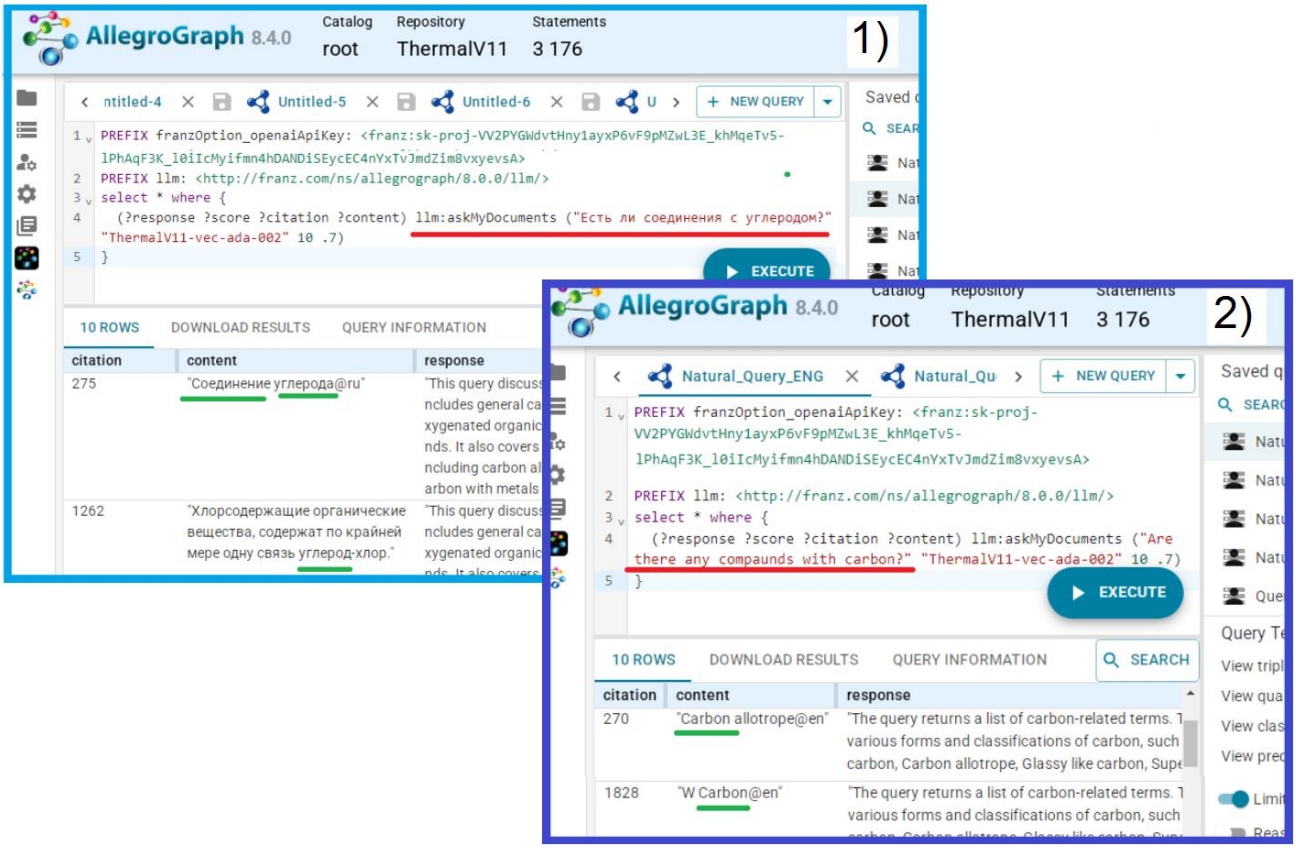


Рис. 3. Сканы интерфейса БД AllegroGraph с запросами на русском (1) и английском (2) языках к онтологии «Термаль»

На рис. 4 представлены четыре примера найденных классов онтологии «Термаль» с содержанием различных аннотационных свойств. На рис. 5 показаны (рядом, для сравнения) полные списки результатов поиска на русском и английском ЕЯ с указанием содержания найденного контента, оценки приближения результата применяемой функциональной модели LLM и соответствующего контенту класса онтологии. Для пояснения на рис. 3–5 зеленным цветом выделены совпавшие поисковые слова из запроса на ЕЯ.

Отметим следующие важные моменты в полученных результатах.

1. Поиск, выполненный в векторном образе онтологии, выдал 10 ближайших результатов, содержащих поисковые слова из запроса на ЕЯ с максимальной оценкой приближения 0.8941 для запроса на русском языке и 0.8531 на английском для классов “Carbon_ME” и “Elemental_Carbon_ME” соответственно по их аннотационным свойствам rdfs:label "Соединение углерода"@ru и

rdfs:label "Elemental carbon"@en (см. рис. 3 и 5). Это свидетельствует о формальной работоспособности функционала LLM графовой БД для выполнения запросов на русском и английском ЕЯ.

2. Из 10 результатов только в одном случае запросы на разных языках к онтологии показали одинаковый результат – класс “Carbon_allotrope” благодаря наличию аннотационного свойства на обоих языках: rdfs:label "Carbon allotrope"@en, rdfs:label "Аллотроп углерода"@ru (см. п. 2 на рис. 4 и строки с желтым фоном на рис. 5). Отмеченный результат показал, что не все классы тестовой онтологии обладают полным набором аннотационных свойств на русском и английском языках, что и объясняет наличие всего одного совпадения.

3. При формально правильном совпадении поисковых слов «соединения» и “compounds” со значениями аннотационных свойств класса “Oxygen_ME” (rdfs:label "Соединение кислорода"@ru см. п. 4 рис. 4) и класса “Addition_compound” (rdfs:label "Addition compound"@en см. рис. 5) результат оказался неверным по смыслу запроса. Это указывает на важность выбора выражений на ЕЯ при формировании запроса и более тщательной подготовки значений свойств онтологии. В то же время это является указателем необходимости проверки всех возможностей векторизации онтологии в данном функционале LLM графовой БД: отдельных слов, словосочетаний и предложений.

4. Результаты запроса на русском языке в большинстве своем получены на основе значений аннотационного свойства rdfs:comment @ru (комментарий) на русском ЕЯ и аннотационного свойства rdfs:label @en (ярлык) на английском ЕЯ, см. рис. 5. Это обстоятельство еще раз подчеркивает важность подготовки свойств и содержания онтологии на разных языках.

Полученные результаты показывают, что одинаковый по смыслу вопрос дает различный набор ответов на разных языках (см. рис. 5), что с формальной точки программирования не является нарушением. Однако это указывает на неполноту и неустойчивость подобной модели поиска данных на ЕЯ, особенно при работе с семантическим объектом – онтологией. Но отмеченные выше моменты при анализе результатов дают объяснение возникшей проблемы и пути к ее устранению. Полноценная подготовка значений свойств онтологий, выбор кор-

ректных выражений запросов на ЕЯ и использование всех возможностей векторизации должны обеспечить приемлемое совпадение ответов на различных языках и помочь преодолеть этот недостаток модели поиска.

В целях проверки были проведены частичная коррекция онтологии добавлением недостающих аннотационных свойств на английском и русском языках для классов, описывающих соединения углерода, и уточнение выражений при формулировке запроса. Использование пробных уточняющих выражений для запросов на русском и английском ЕЯ в этом случае приводит к значительному росту совпадающих ответов независимо от языка запроса. Таким образом, предварительные результаты проверки дают обнадеживающие перспективы преодоления обнаруженной проблемы рассмотренной модели поиска на ЕЯ в двуязычной предметной онтологии по теплофизическим свойствам.

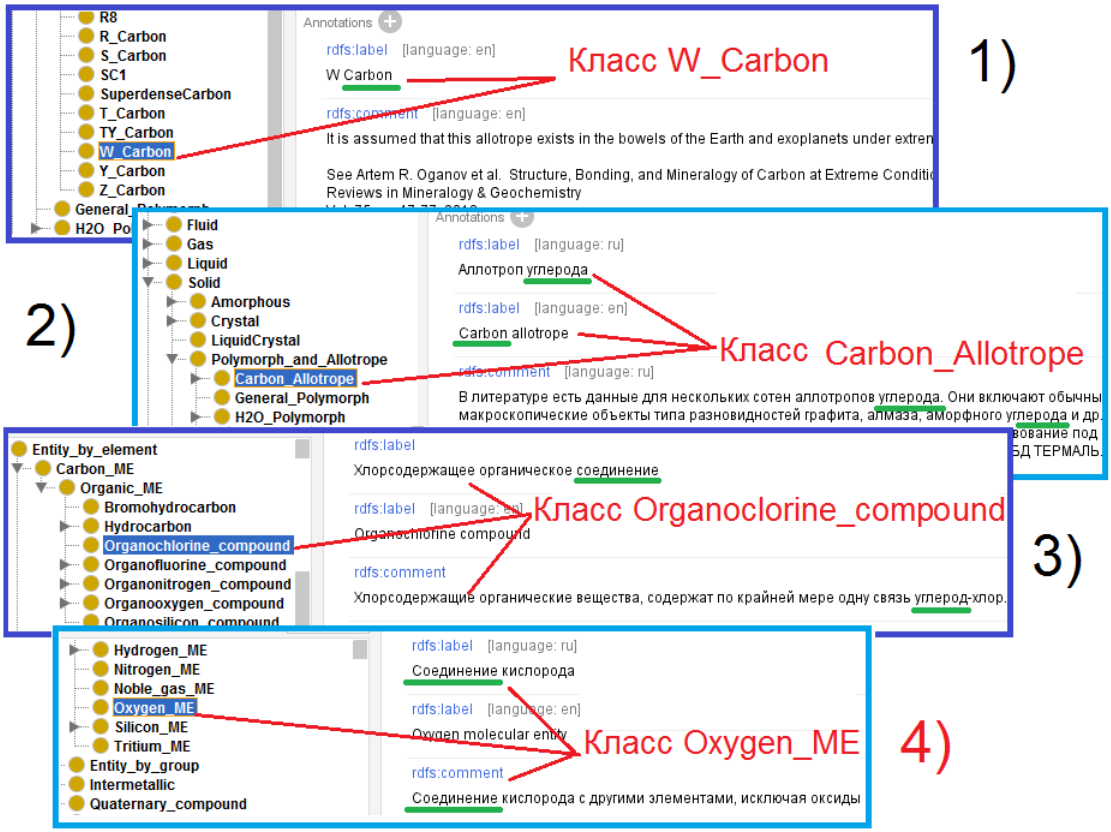


Рис. 4. Фрагменты онтологии «Термаль» с указанием найденных классов по запросу на английском и русском ЕЯ. Представлено содержание аннотационных свойств классов rdfs:label, rdfs:comment

Результаты запроса на русском языке "Есть ли соединения с углеродом"

Содержание контента, ЗЕЛЕННЫМ выделены найденные объекты	score	Имя класса в онтологии "Термаль"
Соединение углерода	0.8941	Carbon_ME
Хлорсодержащие органические вещества; содержат по крайней мере одну связь ***углерод***-хлор.	0.8671	Organochlorine_compound
Кислородсодержащие органические вещества; содержат по крайней мере одну ***углерод***-кислородную связь.	0.867	Organooxygen_compound
Фторсодержащие органические вещества; содержат по крайней мере одну связь ***углерод***-фтор.	0.859	Organofluorine_compound
Углеводород; содержащий замкнутую в кольцо цепь атомов ***углерода***.	0.8578	Cyclic_hydrocarbon
Соединение кислорода	0.8565	Oxygen_ME
Азотсодержащие органические вещества; содержат по крайней мере одну ***углерод***-азотную связь.	0.8544	Organonitrogen_compoun
Аллотроп ***углерода***	0.8536	Carbon_Allotrope
Карбиды; соединения ***углерода*** с металлами; а также с бором и кремнием.	0.8525	Carbide
3-элементное соединение переходного металла с ***углеродом*** и серой	0.8502	Carbosulphide

Результаты запроса на английском языке "Are there any compounds with carbon"

Содержание контента, ЗЕЛЕННЫМ выделены найденные объекты	score	Имя класса в онтологии "Термаль"
Elemental ***carbon***	0.8531	Elemental_carbon_ME
W ***Carbon***	0.8506	W_Carbon
Carbon allotrope	0.8495	Carbon_Allotrope
Amorphous ***Carbon***	0.8386	AmorphousCarbon
T ***Carbon***	0.8403	T_Carbon
R ***Carbon***	0.8394	R_Carbon
M-***carbon***	0.839	M_Carbon
Z ***Carbon***	0.8374	Z_Carbon
Addition ***compound***	0.8346	Addition_compound
Superdense ***carbon***	0.8336	SuperdenseCarbon

Рис. 5. Сравнение результатов запроса на русском и английском ЕЯ к онтологии «Термаль». На рисунке: желтый цвет фона – совпадение результатов поиска на русском и английском языках по классу в онтологии; красная линия подчеркивания – найденные классы не соответствуют смысловому содержанию запроса.

Результаты проверки второго сервиса организации запросов на ЕЯ представлены на сканах интерфейсов БД AllegroGraph, см. рис. 6–8. Они показали принципиальную работоспособность предлагаемого сервиса и возможности его самоулучшения в процессе применения. Проблема одинаковых по смыслу запросов на русском и английском языках проявилась в различном содержании составленных сервисом запросов SPARQL, и соответственно, в получаемых результатах. Так, на рис. 6 представлены два варианта ответа на запрос на русском ЕЯ, отражающие различные решения, предлагаемые сервисом. В одном случае сервис принимает решение использовать в составленном запросе SPARQL слово «углерод», а в другом – его перевод на английский язык “carbon”, что, разумеется, привело к разным результатам. Данная проблема была нами описана выше, при обсуждении результатов применения сервиса «магических» предикатов.

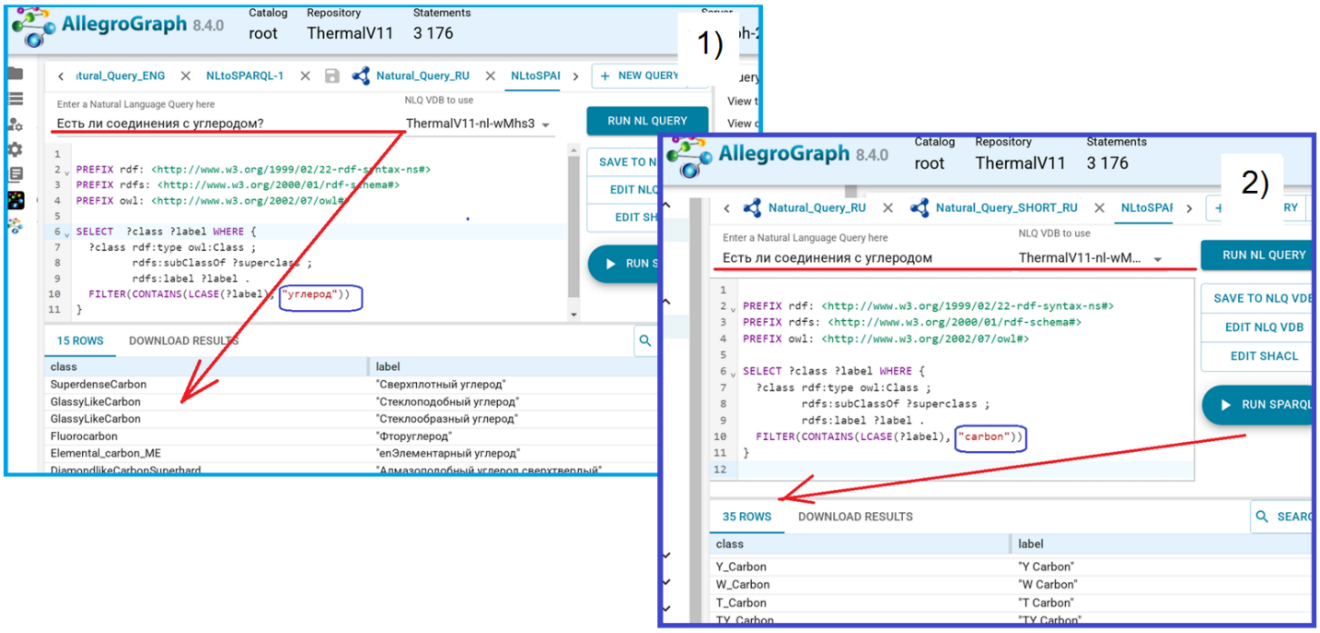


Рис. 6. Варианты запроса на ЕЯ на русском языке, выполненные разными запросами SPARQL. В запросе использовано слово: 1) «углерод»; 2) “carbon”

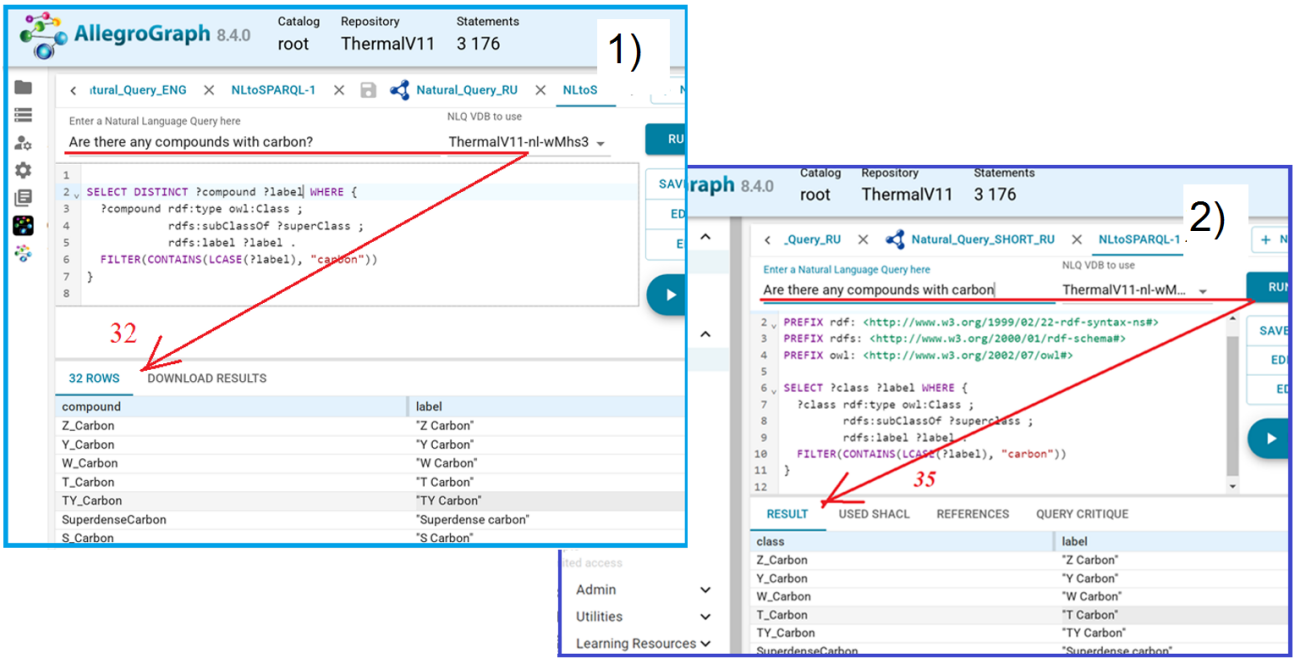


Рис. 7. Примеры уточнения запроса в процессе выполнения запроса на английском языке: 1) правильный результат 32 записи; 2) первоначальный запрос с неверным результатом в 35 записей.

На рис. 7 показан пример, когда в процессе многократного применения запросов произошли уточнение составленного запроса, его так называемое «самоулучшение» и выдача верного результата. Это выразилось в том, что во вновь сформулированном запросе SPARQL была добавлена опция уникальности (не повторяемости) “DISTINCT”, что привело к верному результату.

Наиболее интересный результат был отмечен при получении ответа на запрос ЕЯ количественного характера на двух языках, см. рис. 8. Запрос на английском ЕЯ выдал верный ответ – 32, в отличие от запроса на русском ЕЯ – 35. Данное обстоятельство в целом подтверждает особенность модели LLM, заключающуюся в более точных ответах на запросы на английском ЕЯ в глобальном интернете.

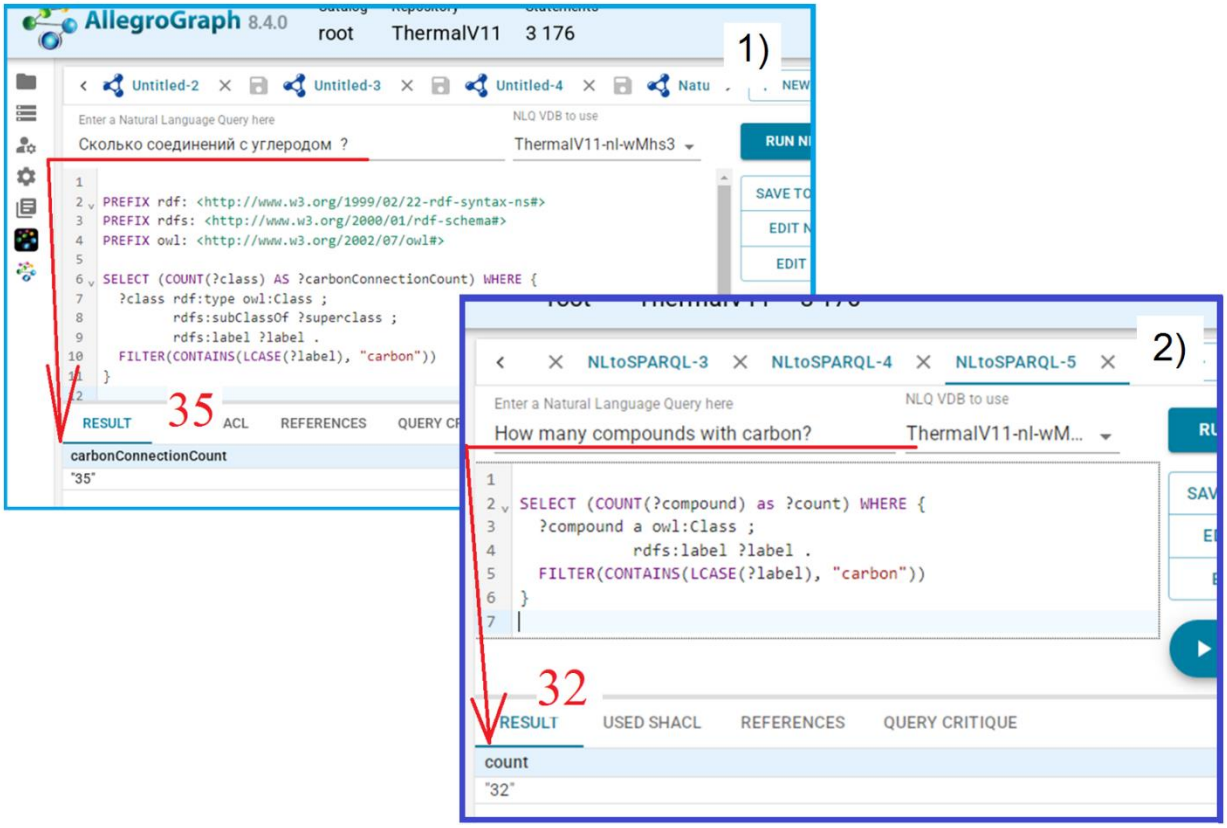


Рис. 8. Пример выполнения запроса на ЕЯ о количестве записей: 1) запрос на русском языке «Сколько соединений с углеродом?» – 35; 2) запрос на английском языке “How many compounds with carbon?” – 32.

Результаты, приведенные на рис. 7 и 8, объяснимы, если мы посмотрим на дерево онтологии «Термаль» (1 на рис. 9) и список из 35 классов (2 на рис. 9), полученные в запросе без применения опции “DISTINCT” из рис. 7 (2). Здесь на рис. 9 выделены три класса онтологии, имеющие дубли в силу того, что они имеют нескольких «родителей».

Так, например, класс «K6 Carbon» является одновременно подклассом класса «Аллотроп углерода» и подклассом класса «Металлический углерод».

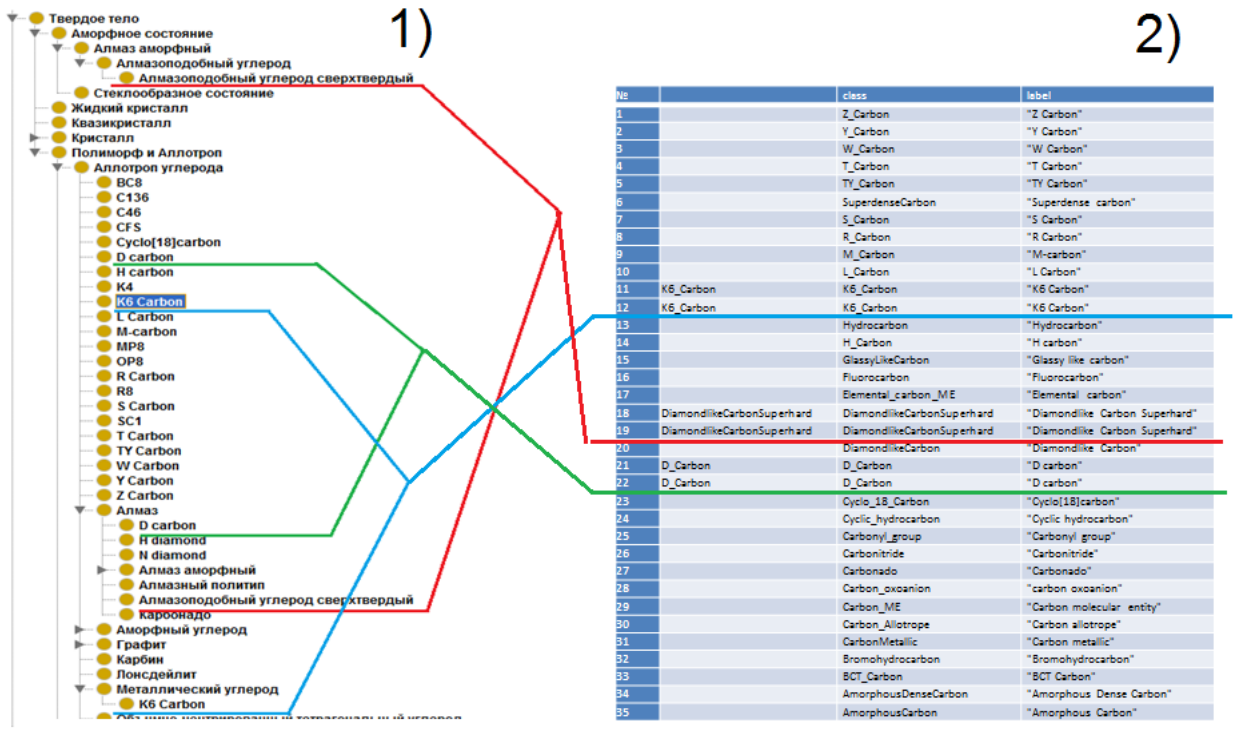


Рис. 9. Демонстрация фрагмента онтологии и списка результатов, полученных после выполнения запроса на ЕЯ, с примерами дублирования классов: 1) фрагмент онтологии; 2) список из 35 классов. На рисунке одинаковыми цветами отмечены продублированные классы в онтологии и списке результатов.

Таким образом, в целом тестирование показало работоспособность сервиса функции преобразования запроса на ЕЯ в запрос SPARQL.

ЗАКЛЮЧЕНИЕ

Проведенный анализ современного состояния средств и технологий, способных реализовать на ЕЯ запросы к БД различного типа, подтвердил развитие

возможностей в решении этих задач средствами LLM и их активное внедрение в прикладные сервисы во множество применяемых БД. Следует особо отметить одно из преимуществ LLM, которое состоит в способности работать на разных языках, включая русский. При этом языковое многообразие допустимо как в самом запросе, так и в семантическом ресурсе, которому направлен запрос. Другое преимущество, обеспеченное новыми сервисами БД, – это возможность «переключить внимание» LLM с глобальной среды на узкий терминологический ресурс с подбором более адекватной лексики.

В работе предложен подход к организации запросов на ЕЯ, использующий сервисы, реализованные в графовой БД AllegroGraph. Семантический ресурс, на котором изучались возможности и проблемы предложенного подхода, представлял собой онтологическую модель «Термаль», выполняющую роль управляющей надстройки и метаданных БД по теплофизическим свойствам вещества. Особенность ресурса – это богатый объем терминологии по классам веществ, их физическим свойствам и взаимосвязям при активном использовании двух языков, русского и английского.

После загрузки онтологической модели на облачный вариант БД AllegroGraph была проведена серия экспериментов по проверке релевантности ответов на переданные запросы. В целом проверка выполнения запросов к онтологии на русском и английском языках выявила работоспособность сервисов графовой БД AllegroGraph. Наряду с этим были зафиксированы явные различия в полноте ответов при запросах на различных языках. Анализ этих различий позволил выявить неполноту в представлении аннотационных свойств онтологии, а проведенная коррекция позволила улучшить совпадение ответов при разных языках запросов. В дальнейшем, помимо поиска в хранилище онтологий, планируется проведение организации аналогичного поиска в хранилище фактографических данных (текстов, таблиц, рисунков), размещенных на документно-ориентированной БД NoSQL типа Mongo DB.

Благодарности

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (Государственное задание № 075-00269-25-00).

СПИСОК ЛИТЕРАТУРЫ

1. *Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А.* Типология материаловедческих данных // Научно-техническая информация. Сер. 2. 2023. № 6. С. 25–39.

2. *Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Косинов А.В.* О представлении и оценке научных данных числового и нечислового типа при проведении исследований по свойствам материалов // Научно-техническая информация. Сер. 2. 2023. № 2. С. 8–16.

3. *Woods W.A.* Semantics and quantification in natural language question answering // Advances in computers. N.Y. etc.: Acad. Press, 1978. Vol. 17. P. 1–87. https://web.stanford.edu/class/linguist289/woods.pdf

4. *Бородин Д.С., Строганов Ю.В.* К задаче составления запросов к базам данных на естественном языке // Новые информационные технологии в автоматизированных системах: материалы 19-го научно-практического семинара. М.: ИПМ им. М.В. Келдыша, апрель 2016. С. 119–125.

5. *Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. М.: МИЭМ, 2011. 272 с.

6. *Бородин Д.С., Строганов Ю.В., Волкова Л.Л., Рудаков И.В., Просуков Е.А.* Транслятор запросов на ограниченном естественном языке в запросы к реляционным базам данных // Системный администратор. 2019. Выпуск №01-02. С. 194–195.

7. *Посевкин Р.В.* Применение семантической модели базы данных при реализации естественно-языкового пользовательского интерфейса // Научно-технический вестник информационных технологий, механики и оптики. 2018. Том 18. №2. С. 262–267.

8. *Mikolov T., et al.* Distributed representations of words and phrases and their compositionality // Proc. 26th Int. Conf. on Neural Information Processing Systems. 2013. P. 3111–3119.

9. *Pennington J., et al*. Glove: Global vectors for word representation // Proc. Conf. Empirical Methods in Natural Language Processing. 2014. P. 1532–1543.

10. *Kenton J.D.M.-W. C., Toutanova L.K.* Bert: Pre-training of deep bidirectional transformers for language understanding // Proc. Conf. of North American Chapter of Association for Computational Linguistics. 2019. P. 4171–4186.

11. *Hafsa Shareef Dar, M. Ikramullah Lali, Khalid Mahmood Malik, Syed Ahmad Chan Bukhari.* Frameworks for Querying Databases Using Natural Language: A Literature Review. arXiv preprint. 2019. URL: https://arxiv.org/abs/1909.01822

12. *Baig Muhammad Shahzaib, et al.* Natural Language to SQL Queries: A Review Original Article // International Journal of Innovations in Science & Technology. 2022. Vol. 4. Issue 1. P. 147–162.

13. *Tao Yu, et al.* Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. arXiv preprint. 2018.
URL: https://arxiv.org/abs/1809.08887

14. *Manning C.D*. Human language understanding & reasoning // Daedalus 2022. Vol.151. Issue 2. P. 127–138.

15. *Meyer Jesse G., et al.* ChatGPT and large language models in academia: opportunities and challenges. // BioData Mining. 2023. Vol. 16. Art. numb. 20.

16. Microsoft Copilot в Azure с базой данных SQL Azure. URL: https://learn.microsoft.com/ru-ru/azure/azure-sql/copilot/copilot-azure-sql-overview?view=azuresql

17. MongoDB Query Generator using OpenAI.
URL: https://www.mongodb.com/docs/compass/current/query-with-natural-language/#std-label-compass-query-natural-language

18. Lower your Large Language Model costs with Graphwise GraphDB. URL: https://www.ontotext.com/blog/lower-your-llm-costs-with-graphwise-graphdb/

19. AllegroGraph 8.4.0 LLM Embed Specification.
URL: https://franz.com/ agraph/support/documentation/llmembed.html

20. Stardog Voicebox FAQ: How LLM, Generative AI, and Knowledge Graphs are the Future of Data Management.
URL: https://www.stardog.com/blog/stardog-voicebox-faq-how-llm-generative-ai-and-knowledge-graphs-are-the-future-of-data-management/

21. *Трахтенгерц М.С.* Технология подготовки информации для баз данных в обменном формате ISO 2709 // Научно-техническая информация. Сер. 2. 2006. № 7. С. 28–31.

QUERIES TO NON-RELATIONAL DATA USING NATURAL LANGUAGE BASED ON A LARGE LANGUAGE MODEL

A. O. Erkimbaev[1] [0000-0002-5239-2208], V. Yu. Zitserman[2] [0000-0003-3327-3139], G. A. Kobzev[3] [0000-0001-9987-1823]

[1-3]*Joint Institute for High Temperatures, RAS, 125412, Moscow, Russia*

[1]adilbek@jiht.ru, [2]vz1941@mail.ru, [3]gkbz@mail.ru

Abstract

The main purpose of this work is to explore new opportunities for organizing natural language queries in scientific local databases that are not relational. A brief review of recent research shows that there has been an active introduction of natural language queries into databases of various types, and the use of machine learning methods, such as neural algorithms, is noted. The widespread use of large language models in the last two years for query generation in various language settings and fields of expertise has been demonstrated. A study has been conducted to explore the potential of the AllegroGraph graph database in using large language models for natural language search. The functionality of the database has been examined using the example of a metadata system for thermophysical properties in the form of the "Thermal" domain ontology. Testing search queries in a bilingual (English and Russian) database environment has revealed some general problems that can be overcome, and it gives us good hope for the future application of new services using large language models.

Keywords*: natural language query, large language model, embedding, non-relational databases, graph database, domain ontology.*

REFERENCES

1. *Erkimbaev A.O., Zitserman V.Iu., Kobzev G.A.* Tipologiia materialovedcheskikh dannykh // Nauchno-tekhnicheskaia informatsiia. Ser. 2. 2023. № 6. S. 25–39.

2. *Erkimbaev A.O., Zitserman V.Iu., Kobzev G.A., Kosinov A.V.* O predstavlenii i otsenke nauchnykh dannykh chislovogo i nechislovogo tipa pri provedenii issledovanii

po svoistvam materialov // Nauchno-tekhnicheskaia informatsiia. Ser. 2. 2023. № 2. S. 8–16.

3. *Woods W.A.* Semantics and quantification in natural language question answering. // Advances in computers. N.Y. etc.: Acad. Press, 1978. Vol. 1 7. P. 1–87. https://web.stanford.edu/class/linguist289/woods.pdf

4. *Borodin D.S., Stroganov Iu.V.* K zadache sostavleniia zaprosov k bazam dannykh na estestvennom iazyke // Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh: materialy 19 nauchno-prakticheskogo seminara. M.: IPM im. M.V. Keldysha, aprel 2016. P. 119–125.

5. *Bolshakova E.I., Klyshinskii E. S., Lande D.V., Noskov A.A., Peskova O.V., Iagunova E.V.* Avtomaticheskaia obrabotka tekstov na estestvennom iazyke i kompiuternaia lingvistika: uchebnoe posobie. M.: MIEM, 2011. 272 s.

6. *Borodin D.S., Stroganov Iu.V., Volkova L.L., Rudakov I.V., Prosukov E.A.* Transliator zaprosov na ogranichennom estestvennom iazyke v zaprosy k reliatsionnym bazam dannykh // Sistemnyi administrator. 2019. Vypusk №01-02. S. 194–195.

7. *Posevkin R.V.* Primenenie semanticheskoi modeli bazy dannykh pri realizatsii estestvenno-iazykovogo polzovatelskogo interfeisa // Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki. 2018. Tom 18. № 2. S. 262–267.

8. *Mikolov T., et al.* Distributed representations of words and phrases and their compositionality // Proc. 26th Int. Conf. on Neural Information Processing Systems. 2013. P. 3111–3119.

9. *Pennington J., et al.* Glove: Global vectors for word representation // Proc. Conf. Empirical Methods in Natural Language Processing. 2014. P. 1532–1543.

10. *Kenton J.D.M.-W. C., Toutanova L.K.* Bert: Pre-training of deep bidirectional transformers for language understanding // Proc. Conf. of North American Chapter of Association for Computational Linguistics. 2019. P. 4171–4186.

11. *Hafsa Shareef Dar, M. Ikramullah Lali, Khalid Mahmood Malik, Syed Ahmad Chan Bukhari*. Frameworks for Querying Databases Using Natural Language: A Literature Review. 2019. P. 1–18. arXiv preprint.
URL: https://arxiv.org/abs/1909.01822

12. *Baig Muhammad Shahzaib, et al.* Natural Language to SQL Queries: A Review Original Article // International Journal of Innovations in Science &

Technology. 2022. Vol. 4. Issue 1. P. 147–162.

13. *Tao Yu, et al.* Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. arXiv preprint. 2018. URL: https://arxiv.org/abs/1809.08887

14. *Manning C.D.* Human language understanding & reasoning // Daedalus 2022. Vol. 151. Issue 2. P. 127–138.

15. *Meyer Jesse G., et al.* ChatGPT and large language models in academia: opportunities and challenges // BioData Mining 2023. Vol. 16. Art. numb. 20.

16. Microsoft Copilot в Azure с базой данных SQL Azure. URL: https://learn.microsoft.com/ru-ru/azure/azure-sql/copilot/copilot-azure-sql-overview?view=azuresql

17. MongoDB Query Generator using OpenAI. URL: https://www.mongodb.com/docs/compass/current/query-with-natural-language/#std-label-compass-query-natural-language

18. Lower your Large Language Model costs with Graphwise GraphDB. URL: https://www.ontotext.com/blog/lower-your-llm-costs-with-graphwise-graphdb/

19. AllegroGraph 8.4.0 LLM Embed Specification. URL: https://franz.com/agraph/support/documentation/llmembed.html

20. Stardog Voicebox FAQ: How LLM, Generative AI, and Knowledge Graphs are the Future of Data Management. URL: https://www.stardog.com/blog/stardog-voicebox-faq-how-llm-generative-ai-and-knowledge-graphs-are-the-future-of-data-management/

21. *Trakhtengerts M.S.* Tekhnologiia podgotovki informatsii dlia baz dannykh v obmennom formate ISO 2709 // Nauchno-tekhnicheskaia informatsiia. Ser. 2. 2006. № 7. S. 28–31.

СВЕДЕНИЯ ОБ АВТОРАХ

ЕРКИМБАЕВ Адильбек Омирбекович – старший научный сотрудник лаборатории теплофизических баз данных Объединенного института высоких температур (ОИВТ РАН), кандидат техн. наук. Область научных интересов: теплофизика, теплофизические свойства веществ, технологии баз данных.

Adilbek Omirbekovich ERKIMBAEV – Senior Researcher at the Thermophysical Databases Laboratory, Joint Institute for High Temperatures (JIHT RAS), PhD. Research interests: thermophysics, thermophysical properties of substances, database technologies.

email: adilbek@jiht.ru

ORCID: 0000-0002-5239-2208

ЗИЦЕРМАН Владимир Юрьевич – ведущий научный сотрудник лаборатории теплофизических баз данных Объединенного института высоких температур (ОИВТ РАН), кандидат физ.-мат. наук. Область научных интересов: теплофизика, химическая физика, технологии баз данных.

Vladimir Yurievich ZITSERMAN – Leading Researcher at the Thermophysical Databases Laboratory, Joint Institute for High Temperatures (JIHT RAS), Ph.D. in Physico-mathematical Sciences. Research interests: thermophysics, chemical physics, database technologies.

email: vz1941@mail.ru

ORCID: 0000-0003-3327-3139

КОБЗЕВ Георгий Анатольевич – главный научный сотрудник лаборатории теплофизических баз данных Объединенного института высоких температур (ОИВТ РАН), доктор физ.-мат. наук. Область научных интересов: теплофизика, физика неидеальной плазмы, систематизация научных данных

George Anatolyevich KOBZEV – Principal Research Scientist at the Thermophysical Databases Laboratory, Joint Institute for High Temperatures (JIHT RAS), DSc (Phys), Research interests: thermophysics, the physics of non-ideal plasmas, scientifical data categorization.

email: gkbz@mail.ru

ORCID: 0000-0001-9987-1823

Материал поступил в редакцию 12 декабря 2025 года

УДК 004.4

ВЕБ-СИСТЕМЫ ПО ТЕОРЕТИКО-ГРАФОВЫМ МОДЕЛЯМ И МЕТОДАМ В ПРОГРАММИРОВАНИИ

В. Н. Касьянов[1] [0000-0002-4899-9429], Е. В. Касьянова[2] [0000-0002-3412-0997]

[1, 2]*Институт систем информатики им. А. П. Ершова СО РАН,*
г. Новосибирск, Россия

[1]kvn@iis.nsk.su, [2]kev@iis.nsk.su

Аннотация

Теория графов из академической дисциплины все более превращается в средство, владение которым становится решающим для успешного применения компьютеров во многих прикладных областях. Несмотря на наличие обширной специальной литературы по решению задач на графах, широкое применение в практике программирования полученных математических результатов затруднено в силу отсутствия систематического их описания, ориентированного на программистов. Поэтому значительный класс практических задач, по существу сводящихся к простому выбору подходящего способа решения и построению конкретных формулировок абстрактных алгоритмов, для многих программистов все еще остается полем для интеллектуальной деятельности по «переоткрытию» известных методов. Статья посвящена разрабатываемому в Институте систем информатики им. А. П. Ершова СО РАН цифровому вики-словарю WikiGRAPP по теории графов и ее применениям в информатике и программировании и цифровой вики-энциклопедии WEGA теоретико-графовых алгоритмов решения задач информатики и программирования.

Ключевые слова*: теоретико-графовые модели, теоретико-графовые методы, программирование, цифровой вики-словарь, цифровая вики-энциклопедия.*

ВВЕДЕНИЕ

Современное программирование невозможно представить себе без применения теоретико-графовых моделей и методов. Хорошо известно, что многие задачи повышения эффективности и надежности конструирования программ с использованием языков высокого уровня и трансляторов были сформулированы и решены как задачи на графах. К ним в первую очередь относятся задачи, связанные с представлением алгоритмов, программ и систем в виде теоретико-графовых моделей. Роль теоретико-графовых методов в программировании существенно возросла в последние годы в связи с появлением параллельных компьютеров и сетей, а также новых предметных областей, таких как веб-графы, социальные сети, семантический веб, базы знаний, библиографические сети, сети белок-белковых взаимодействий и многие другие.

Началом широкого внедрения методов теории графов в практику научных и технических исследований следует считать 50-е г. XX в. В те годы были опубликованы отчеты американской корпорации RAND по математическим исследованиям в военной области, которые проводились в США во время и после окончания Второй мировой войны. Книги по теории графов К. Бержа [1] и О. Оре [2], вышедшие в начале 1960-х г., уже содержали материалы, относящиеся к приложениям теории графов в исследовании операций, дискретной оптимизации, электротехнике и пр. Затем последовали книги, посвященные алгоритмическим проблемам теории графов и вопросам применения теории графов в отдельных областях знаний, в том числе программировании. Среди них нужно назвать «Введение в теоретическое программирование. Беседы о методе» А. П. Ершова [3], «Применение теории графов в программировании» В. А. Евстигнеева [4], «Оптимизирующие преобразования программ» В. Н. Касьянова [5], «Комбинаторика для программистов» В. Липского [6]. Заметим, что любая из книг, содержащих теоретико-графовые алгоритмы, оставалась либо книгой по теории графов, либо книгой по основной предметной области, использующей теоретико-графовые методы. Это же относится и к фундаментальному труду «Искусство программирования» Д. Кнута [7–9], который посвящен рассмотрению и анализу важнейших алгоритмов, используемых в информатике. В отличие от них книга «Графы в программировании: обработка, визуализация и применение» В. Н. Касьянова и В. А. Евстигнеева объединяет эти два

направления с точки зрения программирования [10]. Она также впервые в отечественной литературе содержит монографическое изложение материала по вопросам визуализации информации на основе графовых моделей.

В последние годы в сети появилось большое количество открытых веб-систем, аккумулирующих знания по различным предметным областям, связанным с теоретико-графовыми моделями и методами. Среди них всем известная WikipediA [11], а также такие более специализированные системы, посвященные отдельным прикладным областям, как MathWorld [12] и AlgoWiki [13].

MathWorld – это наиболее обширный математический ресурс паутины, предоставляемый в качестве бесплатной услуги компанией Wolfram Research, создавшей хорошо известную систему Mathematica. С 1995 г. сайт MathWorld [12] активно развивается и поддерживается. В настоящее время MathWorld – это открытая энциклопедия по математике, которая считается не только самым ярким и самым читаемым интернет-ресурсом по математике, но и одним из самых надежных. Ее статьи широко упоминаются в журналах и книгах. MathWorld продолжает расти и развиваться при поддержке тысяч активных пользователей и в настоящее время содержит более 13900 статей по всем основным разделам математики.

AlgoWiki [13] – это открытая энциклопедия по свойствам алгоритмов и особенностям их реализации на различных программно-аппаратных платформах от мобильных платформ до экзафлопсных суперкомпьютерных систем с возможностью коллективной работы всего мирового вычислительного сообщества. Она разрабатывается как вики-система с 2014 года на базе Научно-исследовательского вычислительного центра Московского государственного университета и в настоящее время содержит более 300 статей, посвященных в основном алгоритмам линейной алгебры.

Широкая применимость графов связана с тем, что они являются естественным средством объяснения сложных ситуаций на интуитивном уровне. Эти преимущества представления сложных структур и процессов графами становятся более ощутимыми при наличии хороших средств их визуализации. Поэтому неслучайно в последнее время в мире растет интерес к методам и системам рисования и визуальной обработки графов и графовых моделей [14, 15]. Многие про-

граммные системы, особенно те, которые используют информационные модели, включают элементы визуальной обработки графовых объектов. Среди них – системы и окружения программирования, инструменты CASE-технологии, системы автоматизации проектирования и многие другие.

Поскольку информация, обрабатываемая на компьютерах, постоянно увеличивается и усложняется, возникает все больше ситуаций, в которых классические графовые модели перестают быть адекватными. Требуются и возникают более мощные графовые формализмы для представления информационных моделей, обладающих иерархической структурой, такие как кластерные графы [16] и составные графы [17], описывающие иерархию для неориентированных и ориентированных графов соответственно. Использование иерархического представления является основой многочисленных методов анализа и синтеза сложных информационных моделей в различных областях применения компьютеров и позволяет поддерживать интерактивную визуализацию сложных информационных моделей большого размера [18, 19].

Теория графов из академической дисциплины все больше превращается в средство, владение которым становится решающим для успешного применения компьютеров во многих прикладных областях.

Поэтому неслучайно, что в Лаборатории конструирования и оптимизации программ Института систем информатики (ИСИ) им. А. П. Ершова СО РАН с момента ее создания в 1990 г. ведутся исследования методов и средств повышения эффективности и надежности конструирования программ и систем на основе теоретико-графовых моделей и методов (см., например, [20–22]).

Одним из важных направлений этих работ является разработка методов и средств поддержки применения теоретико-графовых методов в информатике и программировании.

В рамках этого направления была опубликована серия книг и учебных пособий по теоретико-графовым методам в информатике и программировании [4, 5, 10, 15, 23–26]. Разрабатываются методы и средства для визуализации сложно структурированной информации большого объема на основе предложенного формализма иерархических графов и графовых моделей [27–35]. Ведется работа по созданию цифрового вики-словаря WikiGRAPP [36] по теории графов и его применениям в информатике и программировании и цифровой вики-

энциклопедии WEGA [37] теоретико-графовых алгоритмов решения задач информатики и программирования.

Создаваемым цифровым вики-системам WikiGRAPP и WEGA поддержки применения теоретико-графовых методов в информатике и программировании, а также поддерживающим их системам визуализации атрибутированных иерархических графов и посвящена настоящая статья.

ЦИФРОВОЙ ВИКИ-СЛОВАРЬ WIKIGRAPP ПО ТЕОРИИ ГРАФОВ И ЕЕ ПРИМЕНЕНИЯМ В ИНФОРМАТИКЕ И ПРОГРАММИРОВАНИИ

В связи с активным развитием теоретико-графовых методов решения задач на компьютере, а также с их постоянным расширением на новые предметные области проблема терминологии является одной из основных проблем в применении теоретико-графовых методов в программировании и информатике.

Расширяемый цифровой вики-словарь по графам в информатике и программировании WikiGRAPP (см. рис. 1), создаваемый в ИСИ СО РАН, призван если не решить проблему терминологии, то значительно ее облегчить. Для его создания было использовано написанное на препроцессоре гипертекста (PHP) свободно распространяемое программное обеспечение MediaWiki [38], предназначенное для поддержки гипертекстовой среды «вики» (wiki) – такого вебсайта, структуру и содержимое которого пользователи могут сообща изменять с помощью инструментов, предоставляемых самим сайтом.

В основу вики-словаря WikiGRAPP были положены две книги [23, 24].

Первая книга – это словарь [23], опубликованный в 1999 г. в издательстве «Наука», предварительная версия которого была издана в 1995–1996 гг. тремя выпусками в Новосибирском государственном университете. Это был первый словарь по теории графов в информатике и программировании, и он вызвал большой интерес читателей.

Вторая книга [24] – это исправленная и расширенная английская версия словаря 1999 г., опубликованная в Сибирском научном издательстве в 2009 г, которая включила в себя дополнительно более 1000 новых терминов из статей, рефераты которых публиковались в РЖ «Математика» в разделе «Теория графов».

В словаре 1999 г. были собраны теоретико-графовые термины из таких известных монографий по теории графов, как книги Ф. Харари, К. Бержа, О. Оре, А. А. Зыкова и др., а также доступных для отечественного читателя книг по информатике и программированию, с указанием источника и вариантов. Кроме описаний собственно теоретико-графовых терминов в словарь были также включены необходимые для их понимания термины из программирования, комбинаторного анализа, прикладной алгебры и исследования операций, что расширяло круг пользователей словаря.

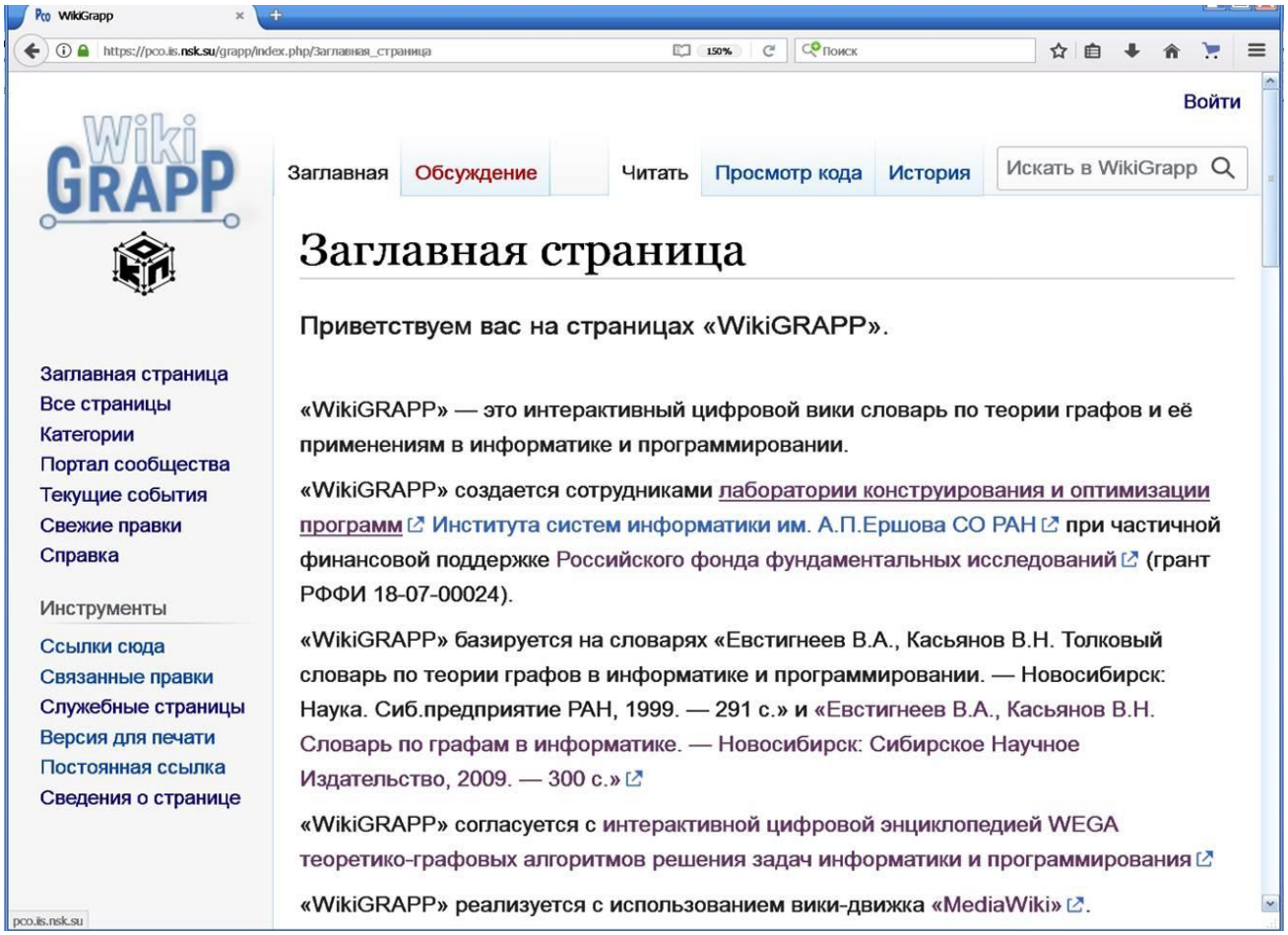


Рис. 1. Заглавная страница вики-словаря WikiGRAPP.

Статьи словаря 1999 г. [23] были снабжены иллюстрациями, перекрестными ссылками и ссылками на доступную литературу. Русские термины в словаре сопровождались их английскими эквивалентами, что позволяло использовать книгу как русско-английский словарь, а прилагаемый к статьям словаря краткий англо-русский словарь был призван помочь при чтении англоязычной литерату-

ры. Последнее, на взгляд авторов, может препятствовать размножению вариантов русских эквивалентов английских терминов, используемых в литературе.

Приведенные выше свойства словаря 1999 г. не только сохранились в создаваемом цифровом словаре WikiGRAPP (см. рис. 2), но и усилились за счет использования в цифровом словаре гипертекстовых ссылок и категорий, присущих вики-системам, а также за счет включения в него всех статей на английском языке из словаря 2009 г. [24].

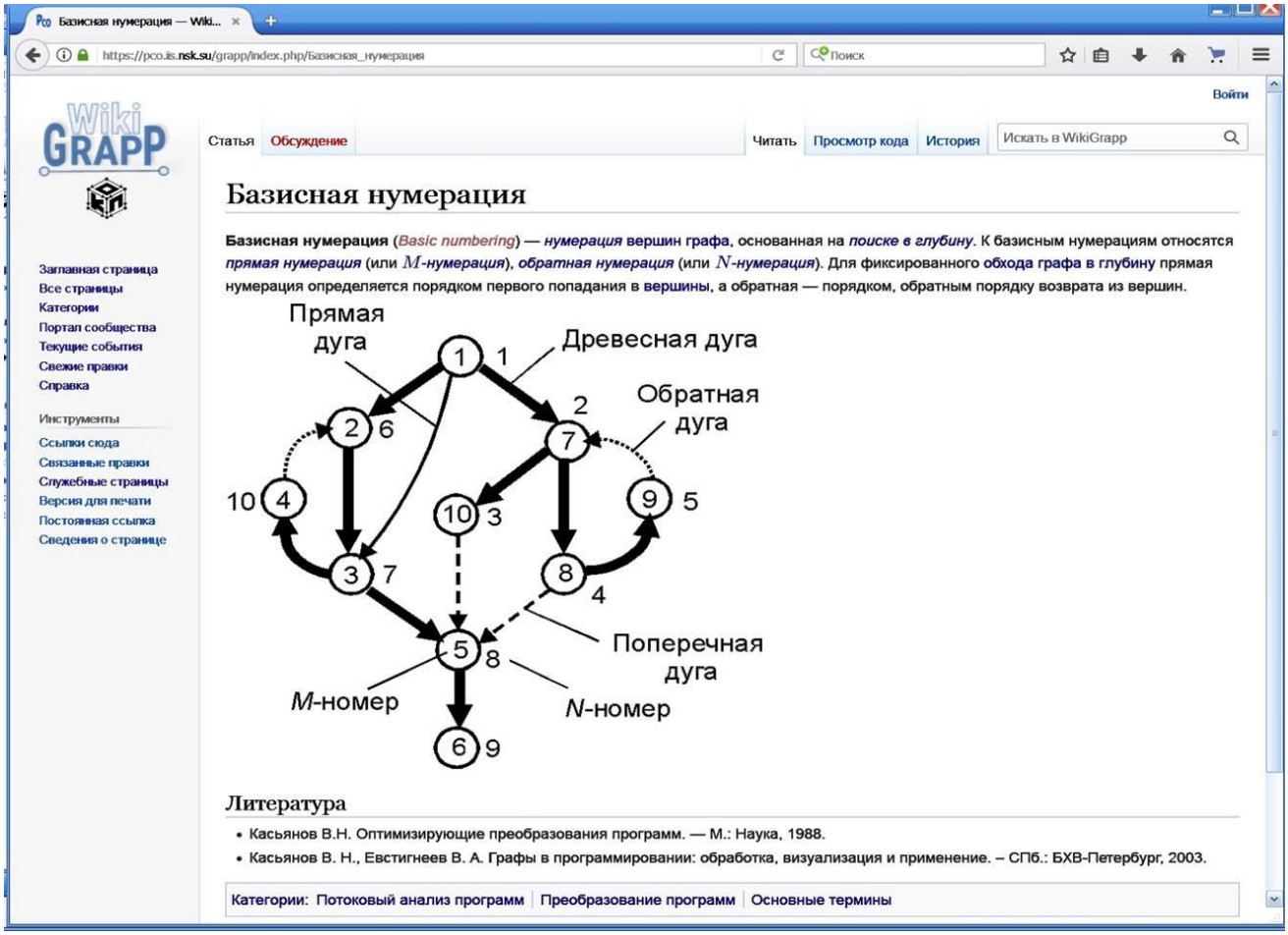


Рис. 2. Страница вики-словаря WikiGRAPP о базисных нумерациях

При отборе терминов для словаря 1999 г. авторы поступали следующим образом. В качестве основного было выбрано множество понятий, представленных в известной монографии [39], как издания по теории графов, наиболее полного и доступного отечественному читателю. Затем оно было пополнено терминами из других отечественных и переводных книг по теории графов, а также из

монографий по информатике и программированию, существенно использующих методы теории графов, таких, например, как [3–5].

Чтобы как-то уменьшить разрыв между включенной в словарь терминологией вышедших в свет монографий и терминологией, еще не использующейся в монографиях, словарь был расширен за счет тех терминов, которые встречаются в докладах на ежегодной конференции “Graph Theory Concepts in Computer Science” и книгах серии “Graph Theory Notes of New York”.

В дальнейшем отмеченное отставание было еще более сокращено в английской версии словаря [24] благодаря включению в него более 1000 новых терминов из докладов конференций и журнальных статей, опубликованных в журналах, ведущих по данной тематике (“Discrete Mathematics”, “Journal of Graph Theory” и др.), рефераты которых публиковались в РЖ «Математика» в разделе «Теория графов». При этом при включении в словарь того или иного термина из статьи или доклада в соответствующей статье словаря делалась лишь общая ссылка на название журнала или конференции с данной публикацией.

Начальная версия цифрового вики-словаря WikiGRAPP, покрывающая печатные издания [23, 24], прошла государственную регистрацию в 2013 г. [40], и до настоящего времени разработчиками словаря ведется постоянная работа по его пополнению и совершенствованию.

ЦИФРОВАЯ ВИКИ-ЭНЦИКЛОПЕДИЯ WEGA ТЕОРЕТИКО-ГРАФОВЫХ АЛГОРИТМОВ РЕШЕНИЯ ЗАДАЧ ИНФОРМАТИКИ И ПРОГРАММИРОВАНИЯ

С использованием MediaWiki [38] в ИСИ СО РАН создается еще и другая вики-система WEGA (см. рис. 3), также ориентированная на поддержку применения графов в программировании, – расширяемая интерактивная цифровая энциклопедия теоретико-графовых алгоритмов решения задач информатики и программирования [37].

В отличие от монографий Д. Кнута [7–9], содержащих низкоуровневое (в терминах машины MIX) описание алгоритмов, цифровая энциклопедия WEGA, равно как и книга [8], на которой она базируется, ориентируется на абстрактную модель современных компьютеров (равнодоступная адресная машина – РАМ) и высокоуровневое описание алгоритмов в терминах специального языка высокого уровня (ВУ-язык).

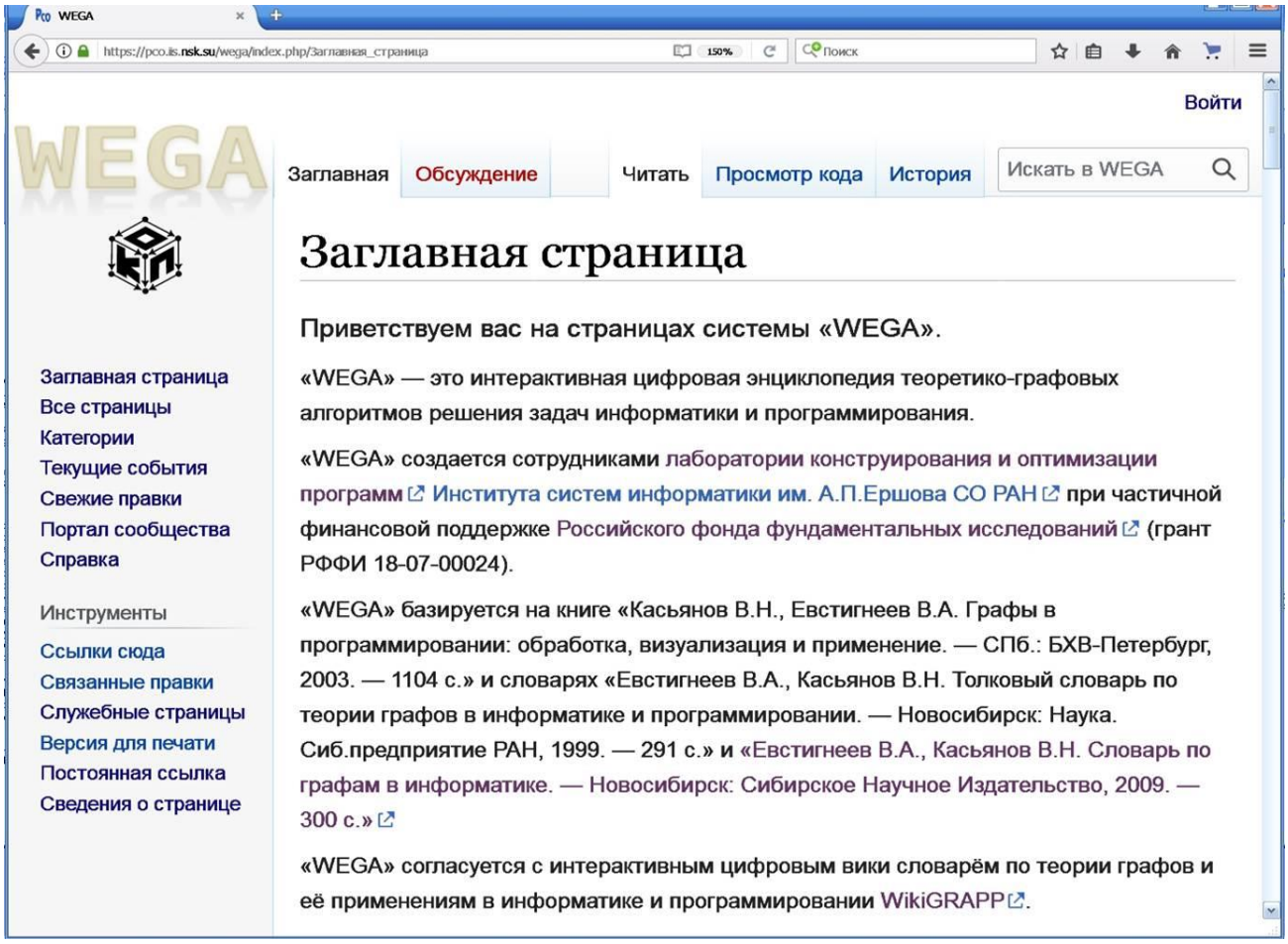


Рис. 3. Заглавная страница вики-энциклопедии WEGA

По существу, ВУ-язык является псевдоязыком (лексиконом) программирования и содержит в качестве базовых традиционные конструкции языков программирования, для каждой из которых фиксируется класс ее допустимых реализаций на РАМ. Предполагается также, что ВУ-язык наряду с базовыми конструкциями позволяет использовать любые необходимые конструкции, если очевидны или заранее зафиксированы оценки их сложности, а также те реализации этих конструкций на РАМ, которые допускают такие оценки. В частности, наряду с типами простых и составных данных, обычными для современных языков, ВУ-язык допускает использование более сложных структур данных, как, например, деревья и графы (см. рис. 4).

Такой подход позволяет формулировать алгоритмы в естественной форме, допускающей прямой анализ их корректности и сложности, а также простой пе-

ренос сформулированных алгоритмов на реальные языки программирования и компьютеры с сохранением полученных оценок сложности.

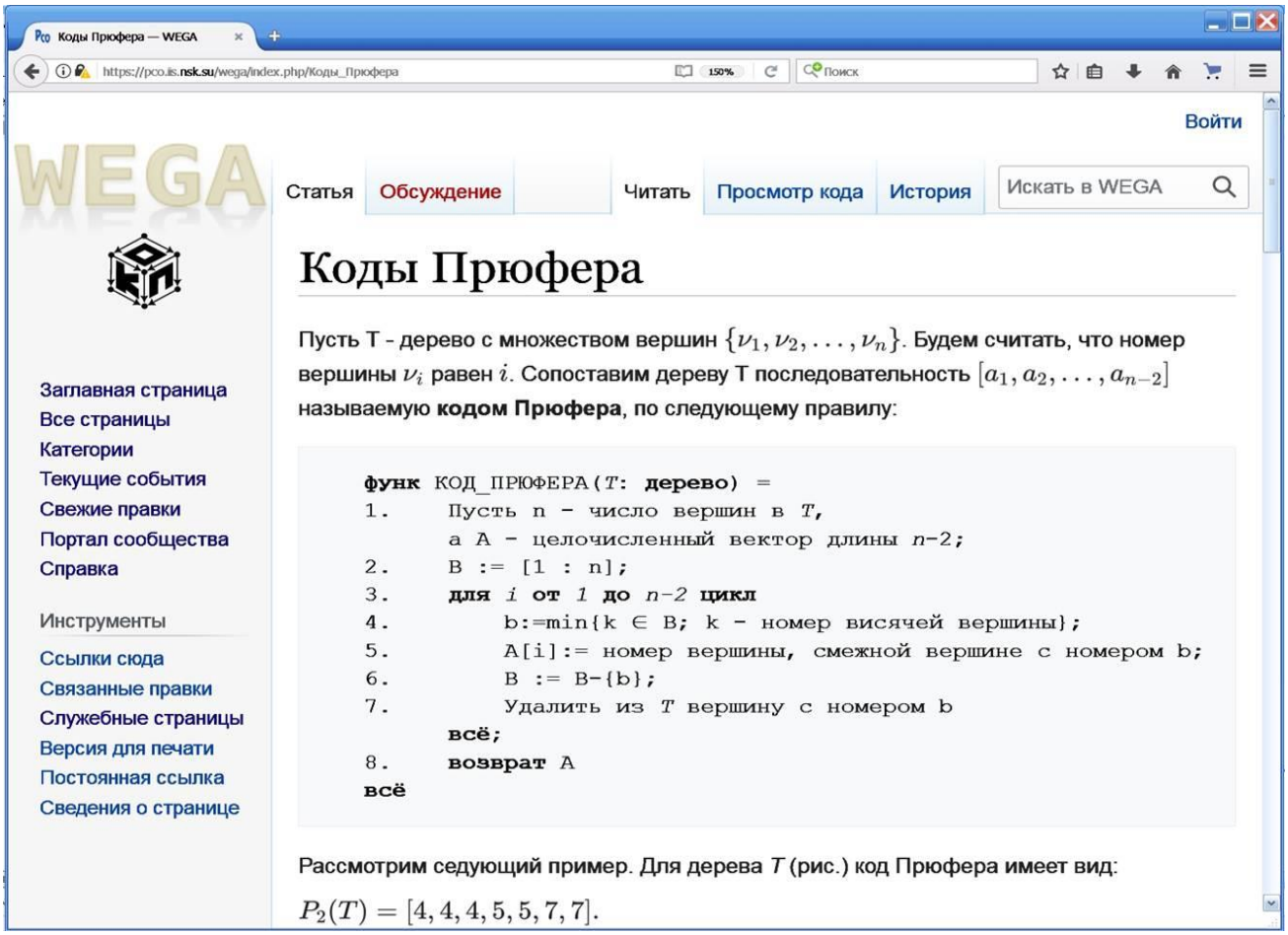


Рис. 4. Страница энциклопедии WEGA о кодах Прюфера

Начальная версия вики-энциклопедии WEGA прошла государственную регистрацию в 2013 г. [41]. Но разработчики энциклопедии до сих пор ведут постоянную работу по ее пополнению и совершенствованию. И хотя энциклопедия WEGA согласуется со словарем WikiGRAPP, обе системы являются полностью независимыми и содержат гиперссылки друг на друга только в заглавных страницах. В отличие от энциклопедии, которая ориентируется главным образом на системных и прикладных программистов, словарь предназначен для всех тех, кто использует теоретико-графовые методы при решении своих задач. Поэтому у словаря значительно более широкий круг пользователей и существенно большая посещаемость, чем у энциклопедии. С этим связано, в частности, то, что, как правило, при запросах по графовой тематике многие браузеры выдают в первую

очередь ссылку на словарь, а не на энциклопедию, а также то, что статьи словаря как по размеру, так и по структуре существенно уступают статьям энциклопедии.

СИСТЕМЫ ВИЗУАЛИЗАЦИИ НА ОСНОВЕ ИЕРАРХИЧЕСКИХ ГРАФОВ И ГРАФОВЫХ МОДЕЛЕЙ

Системы визуализации графов, создаваемые в ИСИ СО РАН, ориентированы на визуализацию атрибутированных иерархических графов – иерархических графовых моделей, в которых семантика выражена с помощью системы атрибутов, сопоставленных элементам иерархических графов [27].

Иерархический граф [27] состоит из основного графа произвольного вида и корневого дерева вложенности некоторого такого выделенного подмножества его фрагментов (частей основного графа), что сам основной граф входит в подмножество и соответствует корню дерева, а фрагменты, соответствующие листьям дерева, образуют некоторое разбиение основного графа на попарно непересекающиеся части. Важный частный случай иерархических графов образуют так называемые простые иерархические графы, в которых все выделенные фрагменты являются подграфами основного графа, порожденными соответствующими подмножествами своих вершин. В частности, подкласс простых иерархических графов, использующих в качестве основных только неориентированные графы, включает все кластерные графы.

Система HIGRES [32, 33], созданная в ИСИ СО РАН, – это первый отечественный универсальный визуализатор и редактор графовых моделей.

Система HIGRES ориентирована на многооконную работу с изображениями на плоскости простых атрибутированных иерархических графов небольшого размера (рис. 5). Каждому выделенному фрагменту в изображении атрибутированного иерархического графа соответствует некоторый прямоугольник плоскости, внутри которого располагаются изображения всех составляющих его элементов и их атрибутов. Кроме того, для каждого выделенного фрагмента можно открыть отдельное окно, в котором видны только атрибутированные элементы данного фрагмента. При этом каждый фрагмент при его изображении в любом окне можно объявить закрытым, тогда изображаются только его контуры, либо открытым, тогда помимо контура фрагмента изображаются и составляющие его

элементы и их атрибуты. Для изображения контуров фрагментов в системе используется прием создания эффекта тени. Закрытые фрагменты выглядят слегка выступающими вверх: как будто они закрыты крышками, открытые же слегка утоплены вниз.

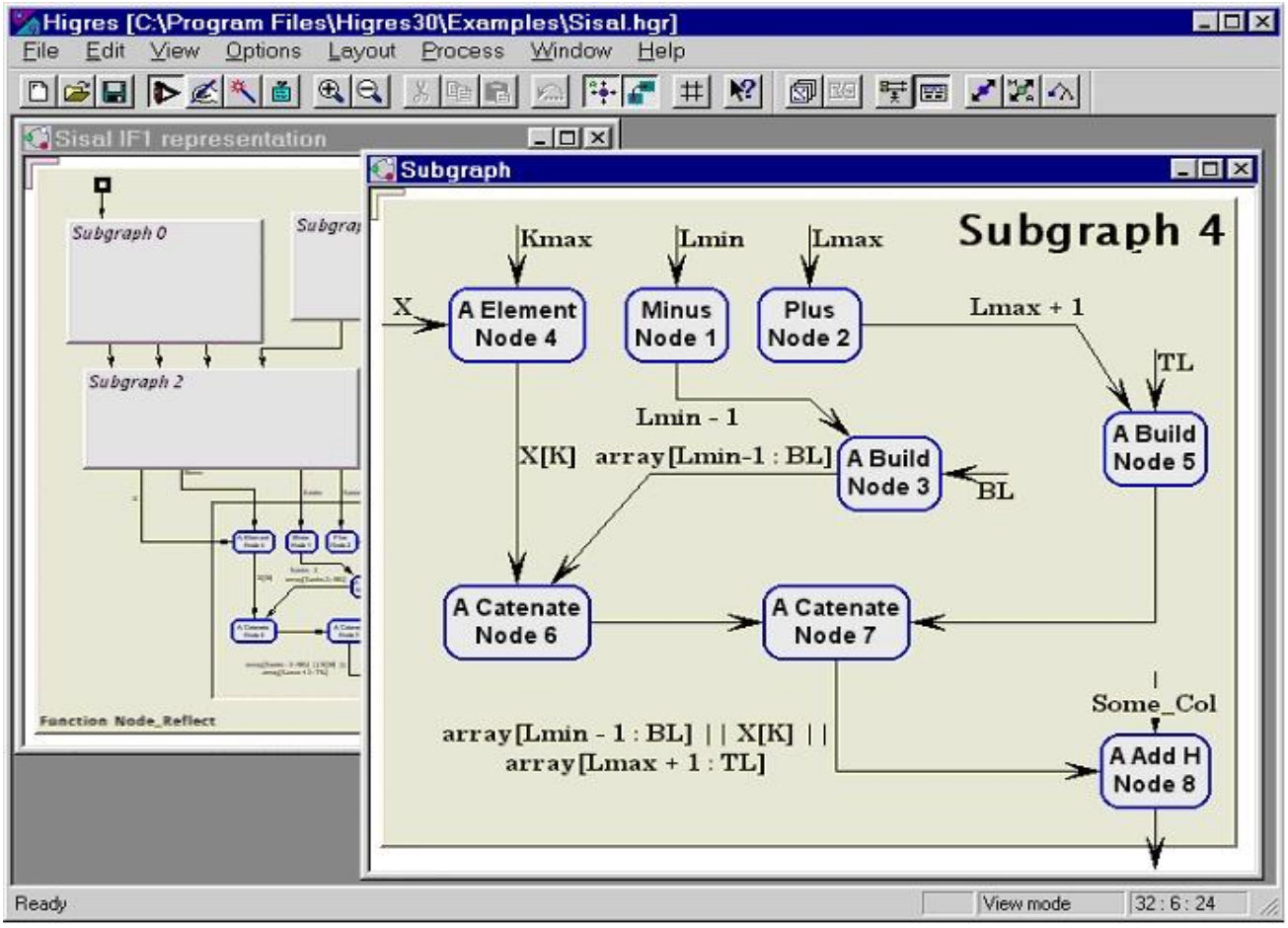


Рис. 5. Система HIGRES: двухоконное изображение простого атрибутированного иерархического графа

Важным отличием системы HIGRES от других универсальных систем визуализации является ее способность сохранять во внутреннем представлении и визуализировать не только сам граф, но и его семантику, представленную в виде системы типов атрибутированных вершин, дуг и фрагментов графа, а также библиотеки алгоритмов обработки, так называемых внешних модулей, причем пользователь системы может легко управлять методами визуализации графовой модели, а также корректировать и доопределять ее семантику. Такой подход обеспечивает, с одной стороны, универсальность системы HIGRES, с другой – возможность ее специализации. Он также позволяет использовать систему как

платформу для работы и анимации алгоритмов работы с иерархическими графами и графовыми моделями. Запустив внешний модуль, пользователь может регулировать параметры обработки графовой модели, прерывать алгоритм на любом шаге, просматривать в любую сторону последовательность изображений промежуточных результатов шагов работы алгоритма как в форме анимации, так и в покадровом режиме (рис. 6).

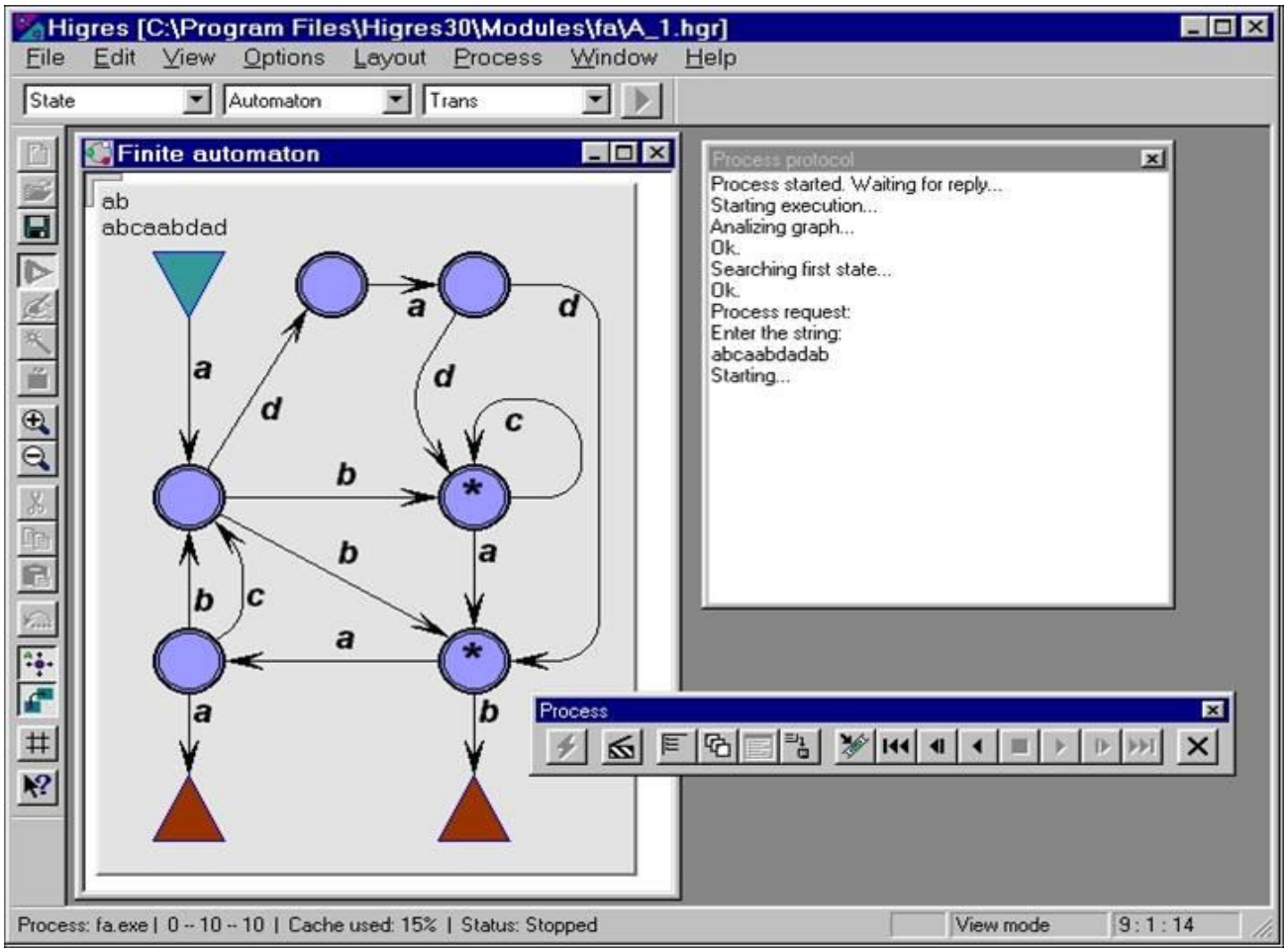


Рис. 6. Система HIGRES: анимация работы конечного автомата

Разработана система визуализации Visual Graph [28], которая работает под управлением ОС Windows, Linux и MacOS, поддерживает обработку произвольных атрибутированных иерархических графов (в том числе составных и кластерных) достаточно большого размера. Она ориентирована на визуализацию структур данных, возникающих в компиляторах, и позволяет одновременно работать с ними как в графовой, так и в текстовой формах. Система Visual Graph предоставляет богатые возможности для навигации по графовой модели и рабо-

ты с атрибутами ее элементов, а также для настройки системы на нужды конкретного пользователя. Система использует для спецификации входного (визуализируемого) графа стандартный язык описания графов GraphML [42] и обеспечивает плавность выполнения основных операций над графами, содержащими до 100000 элементов (вершин и дуг).

Во многих приложениях объекты, моделируемые вершинами графа, являются сложными и могут содержать по несколько разных логических частей (так называемых портов [42]), через которые эти объекты находятся во взаимосвязи, моделируемой дугами и/или ребрами, и которые при изображении графа могут или должны представляться разными точками (или разными непересекающимися частями) изображений вершин, в которых соответствующие вершины соединяются с инцидентными им ребрами или дугами.

Разработан формализм атрибутированных иерархических графов с портами и предложены способы изображения таких графов на плоскости, ориентированные на визуализацию графовых представлений транслируемых программ [31]. Важно, что предложенный подход позволяет строить наглядные иерархические изображения объектов, в которых элементы, моделируемые вершинами, связаны между собой отношениями как объекты целиком, так и через разные свои логические части. Например, он позволяет строить классическое изображение операторной схемы над распределенной памятью (P-схемы) [10, 43], в котором вершины (операторы) изображаются прямоугольниками, а операнды вершины (оператора) – кругами или выступами у этого прямоугольника (сверху прямоугольника располагаются входы, а снизу – выходы), и в котором помимо дуг информационного графа, соединяющего операнды вершин, есть дуги управляющего графа, соединяющие вершины.

На основе предложенной атрибутированной иерархической графовой модели с портами разработаны новые методы и эффективные алгоритмы анализа и визуализации сложно организованной информации большого объема. С их использованием были существенно расширены возможности системы Visual Graph по визуальному представлению и анализу структурной информации, возникающей при работе компиляторов.

Выполнен цикл работ по созданию на основе системы Visual Graph начальной версии веб-сервиса визуализации атрибутированных иерархических графов

с портами (Visual Graph Web Service) для последующей его интеграции со словарем WikiGrapp и энциклопедией WEGA с целью расширения последних возможностями создания и поддержки интерактивных иллюстраций.

На наш взгляд, возможность визуализировать графовую модель, а также наблюдать за процессами ее функционирования и работы алгоритма на этой модели через динамику ее изменений позволит программисту быстрее и глубже понять на конкретных примерах смысл модели и алгоритма, представленных в словаре WikiGrapp или энциклопедии WEGA. В этом плане динамическую визуализацию функционирования графовой модели и работы графового алгоритма нельзя заменить никакими текстовыми описаниями и пояснениями. При этом, на наш взгляд, возможность пользователя управлять не только процессом просмотра анимации (например, останавливать ее, изменять скорость показа, просматривать процесс анимации как вперед, так и назад), но и процессом самой динамической визуализации (например, изменять входную графовую модель, ее промежуточный вид, а также множество демонстрируемых событий и их визуальные представления) в большинстве случаев может существенно упростить пользователю словаря WikiGrapp и энциклопедии WEGA задачу понимания описанных в них графовых моделей и алгоритмов, а также оценку их применимости для решения конкретных задач, стоящих перед пользователем.

ЗАКЛЮЧЕНИЕ

Цифровой словарь WikiGRAPP [36] и цифровая энциклопедия WEGA [37], рассмотренные выше, находятся в открытом доступе в Интернете начиная с 1999 г. Все это время они вместе с книгами и учебными пособиями [4, 5, 10, 15, 23–26], опубликованными разработчиками систем, успешно применяются в учебном процессе Новосибирского государственного университета, в частности в рамках годового спецкурса «Графы в программировании». Но редактировать статьи словаря и энциклопедии до сих пор могут только сотрудники Лаборатории, названной выше, причем далеко не все: среди тех, кто это может делать, есть и такие, которым разрешено изменять содержимое только одной из этих систем. Поэтому в статье говорилось о создании авторских пилотных версий вики-словаря и вики-энциклопедии, перевод которых в вики-системы «общего

пользования» хотя возможен и планируется, но потребует решения непростых вопросов организации модерации.

Основная идея, на которую опирается создание словаря WikiGRAPP и энциклопедии WEGA, состоит в том, что основные свойства самих теоретико-графовых моделей и алгоритмов никак не зависят от вычислительных систем и языков программирования, которые используются в настоящее время и которые планируются использовать в будущем. С этой точки зрения абстрактное описание свойств моделей и алгоритмов имеет очень высокую самостоятельную ценность.

Начальные версии словаря WikiGRAPP и энциклопедии WEGA прошли государственную регистрацию в 2013 г. [40, 41], и до настоящего времени разработчики систем продолжают их совершенствовать и пополнять, добавляя примерно по шесть новых терминов в месяц. Поэтому как весь словарь или всю энциклопедию целиком, так и каждую отдельную статью словаря или энциклопедии можно отнести к так называемым живым публикациям [44]. В 2024 г. системы (словарь и энциклопедия) суммарно подвергались в среднем трем изменениям в день, и к настоящему времени каждая из систем стала содержать порядка 5 тыс. статей.

Создан программный комплекс Wiki2TeX [45], поддерживающий работу со словарем и энциклопедией. Он позволяет строить набор TeX-документов, образующих оффлайн-версию базы данных вики, построенной с помощью MediaWiki, а также выполнять и обратную операцию – преобразовывать TeX-документы в статьи MediaWiki и добавлять их к заданной вики.

В начале 2025 г. вышел «Англо-русский словарь по графам для программиста» [26], который отражает текущее состояние словаря WikiGRAPP и энциклопедии WEGA и дополнительно охватывает более 1000 новых терминов по сравнению с соответствующей частью словаря [25], подготовленного на базе этих систем в 2011 г. Готовится к изданию также «Русско-английский словарь по графам для программиста».

На основе развития методов и систем визуализации сложно структурированной информации с использованием атрибутированных иерархических графовых моделей с портами [27–35] начаты работы по расширению словаря

WikiGRAPP и энциклопедии WEGA возможностями создания и поддержки интерактивных иллюстраций.

Благодарности

Авторы благодарны всем коллегам, которые принимали участие в работах, рассмотренных в статье.

СПИСОК ЛИТЕРАТУРЫ

1. *Берж К.* Теория графов и ее применения. М.: Изд-во иностр. лит., 1962. 319 с.

2. *Оре О.* Графы и их применение. М.: Мир, 1965. 174 с.

3. *Ершов А.П.* Введение в теоретическое программирование (беседы о методе). Новосибирск: Наука, 1977.

4. *Евстигнеев В.А.* Применение теории графов в программировании. М.: Наука, 1985. 352 с.

5. *Касьянов В.Н.* Оптимизирующие преобразования программ. М: Наука, 1988. 336 с.

6. *Липский В.* Комбинаторика для программистов. М.: Мир, 1988.

7. *Кнут Д.* Искусство программирования для ЭВМ. М.: Мир, 1976. Т. 1. 735 с.

8. *Кнут Д.* Искусство программирования для ЭВМ. М.: Мир, 1977. Т. 2. 724 с.

9. *Кнут Д.* Искусство программирования для ЭВМ. М.: Мир, 1978. Т. 3. 844 с.

10. *Касьянов В.Н., Евстигнеев В.А.* Графы в программировании: обработка, визуализация и применение. СПб.: БХВ-Петербург, 2003. 1104 с.

11. WikipediA. URL: https://www.wikipedia.org

12. MathWorld. URL: http://mathworld.wolfram.com/

13. AlgoWiki. URL: http://algowiki-project.org

14. *Di Battista G., Eades P., Tamassia R., Tollis I.G.* Graph drawing: algorithms for visualization of graphs. Prentice Hall, 1999. 397 p.

15. *Касьянов В.Н., Касьянова Е.В.* Визуализация графов и графовых моделей. Новосибирск: Сибирское Научное Издательство, 2010. 123 с.

16. *Feng Q., Cohen R.F., Eades P.* Planarity for clustered graphs // Lecture Notes in Computer Science. 1995. Vol. 979. P. 213–226.

17. *Sugiyama K., Misue K.* Visualization of structural information: automatic drawing of compound digraphs // IEEE Trans. on Systems, Man and Cybernetics. 1991. Vol. 21. No. 4. P. 876–892.

18. *Herman I, Melançon G., Marshall M.S.* Graph visualization and navigation in information visualization: a survey // IEEE Trans. on Visualization and Computer Graphics. 2000. Vol. 6. P. 24-43.

19. *Касьянов В.Н., Касьянова Е.В.* Визуализация информации на основе графовых моделей // Научная визуализация. 2014. Том. 6. № 1. С. 31–50.

20. *Касьянов В.Н.* Применение графов в программировании // Программирование. 2001. № 3. С. 51–70.

21. *Касьянов В.Н., Касьянова Е.В.* Теоретико-графовые методы и системы программирования // Проблемы информатики. 2016. № 1. С. 26–38.

22. *Касьянов В.Н., Касьянова Е.В.* Методы и технологии конструирования эффективных и надежных программ и программных систем на основе графовых моделей и семантических преобразований // Системная информатика. 2021. № 19. С. 1–14.

23. *Евстигнеев В.А., Касьянов В.Н.* Толковый словарь по теории графов в информатике и программировании. Новосибирск: Наука, 1999. 291 с.

24. *Евстигнеев В.А., Касьянов В.Н.* Словарь по графам в информатике. Новосибирск: Сибирское Научное Издательство, 2009. 300 с.

25. *Евстигнеев В.А., Касьянов В.Н.* Русско-английский и англо-русский словарь по графам в информатике. Новосибирск: Сибирское Научное Издательство, 2011. 216 с.

26. *Касьянов В.Н., Касьянова Е.В.* Англо-русский словарь по графам для программиста. Новосибирск: СО РАН, 2025. 160 с.

27. *Касьянов В.Н.* Иерархические графы и графовые модели: вопросы визуальной обработки // Проблемы систем информатики и программирования. Новосибирск: ИСИ СО РАН, 1999. С. 7–32.

28. *Касьянов В.Н., Золотухин Т.А.* Программная система для визуализации сложных больших данных на основе графовых моделей (Visual Graph). Свиде-

тельство о государственной регистрации программы для ЭВМ № 2017612824 от 03.03.2017.

29. *Касьянов В.Н., Касьянова Е.В.* Визуализация информации на основе графовых моделей. Новосибирск: Новосиб. гос. ун-т, 2014. 149 с.

30. *Касьянов В.Н., Золотухин Т.А., Гордеев Д.С.* Методы и алгоритмы визуализации графовых представлений функциональных программ // Программирование. 2019. № 4. С. 19–27.

31. *Касьянов В.Н.* Методы и средства визуализации информации на основе атрибутированных иерархических графов с портами // Сибирский аэрокосмический журнал. 2023. Том 24. № 1. С. 8–17.

32. *Lisitsyn I.A., Kasyanov V.N.* HIGRES – visualization system for clustered graphs and graph algorithms // Lecture Notes in Computer Science. 1999. Vol. 1731. P. 82–89.

33. *Kasyanov V.N., Lisitsyn I.A.* Hierarchical graph models and visual processing // Proc. of Intern. Conf. on Software: Theory and Practice (ICS-2000). 16th World Computer Congress IFIP. Beijing: PHEI, 2000. P. 179–182.

34. *Kasyanov V.N. Kasyanova E.V.* Information visualization based on graph models // Enterprise Information Systems. 2013. Vol. 7, No. 2. P. 187–197.

35. *Kasyanov V.N., Zolotuhin T.A.* A system for visualization of big attributed hierarchical graphs // International Journal of Computer Networks & Communications (IJCNC). 2018. Vol.10. No. 2. P. 55–67.

36. WikiGRAPP. URL: https://pco.iis.nsk.su/grapp/

37. WEGA. URL: https://pco.iis.nsk.su/wega/

38. MediaWiki. URL: http://www.mediawiki.org/wiki/MediaWiki/ru/

39. *Емеличев В.А., Мельников О.И., Сарванов В.И., Тышкевич Р.И.* Лекции по теории графов. М.: Наука, 1990. 384 с.

40. *Касьянов В.Н., Евстигнеев В.А., Касьянова Е.В.* Электронный словарь WikiGRAPP по теории графов и ее применениям в информатике и программировании. Свидетельство о государственной регистрации базы данных № 2013620433 от 25.03.2013.

41. *Касьянов В.Н., Евстигнеев В.А., Касьянова Е.В.* Электронная энциклопедия WEGA теоретико-графовых алгоритмов решения задач информатики и

программирования. Свидетельство о государственной регистрации базы данных № 2013620463 от 01.04.2013.

42. *Brandes U., Marshall M.S., and North S.C.* Graph data format workshop report // Lecture Notes in Computer Science. 2001. Vol. 1984. P. 410–418.

43. *Ershov A.P.* Theory of program schemata // Information Processing 71: Proc. IFIP Congr. 71, Lubliana, 1971. Amsterdam: North Holland, 1972. P. 28–45.

44. Г*орбунов-Посадов М.М.* Жизнь как форма существования научной публикации // Научный сервис в сети Интернет: труды XXVI Всероссийской научной конференции (23–25 сентября 2024 г., онлайн). М.: ИПМ им. М.В. Келдыша, 2024. С. 50–56.

45. *Касьянов В.Н., Касьянова Е.В., Малышев А.А.* Программный комплекс Wiki2Tex. Свидетельство о государственной регистрации программы для ЭВМ № 2016616426 от 01.04.2016.

WEB-SYSTEMS ON GRAPH-THEORETIC MODELS AND METHODS IN PROGRAMMING

V. N. Kasyanov[1] [0000-0002-4899-9429], E. V. Kasyanova[2] [0000-0002-3412-0997]

[1, 2]A. P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia

[1]kvn@iis.nsk.su, [2]kev@iis.nsk.su

Abstract

Graph theory is increasingly turning from an academic discipline into a tool, mastery of which is becoming decisive for the successful use of computers in many applied areas. Despite the existence of extensive specialized literature on solving problems on graphs, the widespread use of the obtained mathematical results in programming practice is difficult due to the lack of a systematic description of them oriented towards programmers. Therefore, a significant class of practical problems, essentially reduced to a simple choice of a suitable solution method and to the construction of specific formulations of abstract algorithms, for many programmers still remains a field for intellectual activity in the “rediscovery” of known methods. The paper is devoted to the digital wiki dictionary WikiGRAPP on graph theory and its applications in computer science and programming and the digital wiki encyclopedia

WEGA of graph-theoretical algorithms for solving computer science and programming problems, being developed at the A.P. Ershov Institute of Informatics Systems SB RAS.

Keywords*: graph-theoretic models, graph-theoretic methods, programming, digital wiki dictionary, digital wiki encyclopedia*.

REFERENCES

1. *Berge K.* Teoriya grafov i eyo primeneniya. M.: Izd-vo inostr. lit.,1962. 319 p.

2. *Ore O.* Grafy i ih primenenie. M.: Mir, 1965. 174 p.

3. *Ershov A.P.* Vvedenie v teoreticheskoe programmirovanie (besedy o metode). Novosibirsk: Nauka, 1977. 288 p.

4. *Evstigneev V.A.* Primenenie teorii grafov v programmirovanii. M.: Nauka 1985. 352 p.

5. *Kasyanov V.N.* Optimiziruyushchie preobrazovaniya programm. M: Nauka, 1988. 336 p.

6. *Lipskiy V.* Kombinatorika dlya programmistov. Moscow: Mir, 1988. 213 p.

7. *Knuth D.* Iskusstvo programmirovaniya dlya EHVM. Moscow: Mir, 1976. Vol. 1. 735 p.

8. *Knuth D.* Iskusstvo programmirovaniya dlya EHVM. Moscow: Mir, 1977. Vol. 2. 724 p.

9. *Knuth D.* Iskusstvo programmirovaniya dlya EHVM. Moscow: Mir, 1978. Vol. 3. 844 p.

10. *Kasyanov V.N., Evstigneev V.A.* Grafy v programmirovanii: obrabotka, vizualizaciya i primenenie. St. Petersburg: BHV-Petersburg, 2003. 1104 p.

11. WikipediA. URL: https://www.wikipedia.org

12. MathWorld. URL: http://mathworld.wolfram.com/

13. AlgoWiki. URL: http://algowiki-project.or

14. *Di Battista G., Eades P., Tamassia R., Tollis I.G.* Graph drawing: algorithms for visualization of graphs. Prentice Hall, 1999. 397 p.

15. *Kasyanov V.N., Kasyanova E.V.* Vizualizaciya grafov i grafovykh modelej. Novosibirsk: Sibirskoe Nauchnoe Izdatel'stvo, 2010. 123 p.

16. *Feng Q., Cohen R.F., Eades P.* Planarity for clustered graphs // Lecture Notes in Computer Science. 1995. Vol. 979. P. 213–226.

17. *Sugiyama K., Misue K*. Visualization of structural information: automatic drawing of compound digraphs // IEEE Trans. on Systems, Man and Cybernetics. 1991. Vol. 21, No. 4. P. 876–892.

18. *Herman I, Melançon G., Marshall M.S.* Graph visualization and navigation in information visualization: a survey // IEEE Trans. on Visualization and Computer Graphics. 2000. Vol. 6. P. 24–43.

19. *Kasyanov V.N., Kasyanova E.V.* Information visualization based on graph models // Scientific visualization. 2014. Vol. 6. No. 1. P. 31–50.

20. *Kasyanov V.N.* Graph applications in programming // Programming and Computer Software. 2001. Vol. 27, No. 3. P. 146–164.

21. *Kasyanov V.N., Kasyanova E.V.* Graph-theoretical methods and programming systems // Problems of Informatics. 2016. No. 1. P. 26–38.

22. *Kasyanov V.N., Kasyanova E.V.* Methods and technologies for constructing efficient and reliable programs and software systems based on graph models and semantic transformations // System informatics. 2021. No. 19. P. 1–14.

23. *Evstigneev V.A., Kasyanov V.N.* Tolkovyj slovar' po teorii grafov v informatike i programmirovanii. Novosibirsk: Nauka, 1999. 291 p.

24. *Evstigneev V.A., Kasyanov V.N.* Slovar' po grafam v informatike. Novosibirsk: Sibirskoe Nauchnoe Izdatel'stvo, 2009. 300 p.

25. *Evstigneev V.A., Kasyanov V.N.* Russko-anglijskij i anglo-russkij slovar' po grafam v informatike. Novosibirsk: Sibirskoe Nauchnoe Izdatel'stvo, 2011. 216 p.

26. *Kasyanov V.N., Kasyanova E.V.* Anglo-russkij slovar' po grafam dlya programmista. Novosibirsk: SO RAN, 2025. 160 p.

27. *Kasyanov V.N.* Ierarkhicheskie grafy i grafovye modeli: voprosy vizual'noj obrabotki // Problemy sistem informatiki i programmirovaniya. Novosibirsk: ISI SO RAN, 1999. P. 7–32.

28. *Kasyanov V.N., Zolotukhin T.A.* Programmnaya sistema dlya vizualizacii slozhnykh bol'shikh dannykh na osnove grafovykh modelej (Visual Graph). Svidetel'stvo o gosudarstvennoj registracii programmy dlya EHVM No. 2017612824 ot 03.03.2017.

29. *Kasyanov V.N., Kasyanova E.V.* Vizualizaciya informacii na osnove grafovykh modelej. Novosibirsk: Novosib. gos. un-t, 2014. 149 p.

30. *Kasyanov V.N., Zolotuhin T.A., Gordeev D.S.* Visualization methods and algorithms for graph representation of functional programs // Programming and Computer Software. 2019. Vol. 45, No. 4. P. 156–162.

31. *Kasyanov V.N.* Methods and tools for information visualization on the basis of attributed hierarchical graphs with ports // Siberian Aerospace Journal. 2023. Vol. 24, No. 1. P. 8–17.

32. *Lisitsyn I.A., Kasyanov V.N.* HIGRES — visualization system for clustered graphs and graph algorithms // Lecture Notes in Computer Science. 1999. Vol. 1731. P. 82–89.

33. *Kasyanov V.N., Lisitsyn I.A.* Hierarchical graph models and visual processing // Proc. of Intern. Conf. on Software: Theory and Practice (ICS-2000). 16th World Computer Congress IFIP. Beijing: PHEI, 2000. P. 179–182.

34. *Kasyanov V.N. Kasyanova E.V*. Information visualization based on graph models // Enterprise Information Systems. 2013. Vol. 7, No. 2. P. 187–197.

35. *Kasyanov V.N., Zolotuhin T.A*. A system for visualization of big attributed hierarchical graphs // International Journal of Computer Networks & Communications (IJCNC). 2018. Vol.10, No. 2. P. 55–67.

36. WikiGRAPP. URL: https://pco.iis.nsk.su/grapp/

37. WEGA. URL: https://pco.iis.nsk.su/wega/

38. MediaWiki. URL: http://www.mediawiki.org/wiki/MediaWiki/ru/

39. *Emelichev V.A., Melnikov O.I., Sarvanov V.I., Tyshkevich R.I.* Lekcii po teorii grafov. M.: Nauka, 1990. 384 p.

40. *Kasyanov V.N., Evstigneev V.A., Kasyanova E.V*. Electronic dictionary WikiGRAPP of graph theory and its applications in computer science and programming. Certificate of state registration of the database No. 2013620433 ot 25.03.2013.

41. *Kasyanov V.N., Evstigneev V.A., Kasyanova E.V.* Ehlektronnaya ehnciklopediya WEGA teoretiko-grafovykh algoritmov resheniya zadach informatiki i programmirovaniya. Svidetel'stvo o gosudarstvennoj registracii bazy dannykh № 2013620463 ot 01.04.2013.

42. *Brandes U., Marshall M.S., and North S.C.* Graph data format workshop report // Lecture Notes in Computer Science. 2001. Vol. 1984. P. 410–418.

43. *Ershov A.P.* Theory of program schemata // Information Processing 71. Proc. IFIP Congr. 71. Lubliana. 1971. Amsterdam: North Holland, 1972. P. 28–45.

44. *Gorbunov-Posadov M.M.* Aliveness as a form of existence of scientific publication // Scientific service on the Internet: proceedings of the XXVI All-Russian scientific conference (September 23–25, 2024, online). M.: IPM im. M.V. Keldysh, 2024. P. 50–56.

45. *Kasyanov V.N., Kasyanova E.V., Malyshev A.A.* Programmnyj kompleks Wiki2Tex. Svidetel'stvo o gosudarstvennoj registracii programmy dlya EHVM № 2016616426 ot 01.04.2016

СВЕДЕНИЯ ОБ АВТОРАХ

КАСЬЯНОВ Виктор Николаевич – доктор физико-математических наук, профессор, главный научный сотрудник, заведующий Лабораторией конструирования и оптимизации программ, Институт систем информатики имени А. П. Ершова СО РАН, профессор, Новосибирский государственный университет, Новосибирск

Victor Nikolaevich KASYANOV – Dr. Sc., Full Professor, Chief Researcher, Head of the Program Construction and Optimization Laboratory, A. P. Ershov Institute of Informatics Systems SB RAS, Professor, Novosibirsk State University, Novosibirsk

email: kvn@iis.nsk.su

ORCID 0000-0002-4899-9429

КАСЬЯНОВА Елена Викторовна – кандидат физико-математических наук, доцент, почетный работник науки и высоких технологий Российской Федерации, старший научный сотрудник, Институт систем информатики имени А. П. Ершова СО РАН, доцент, Новосибирский государственный университет, Новосибирск

KASYANOVA Elena Victorovna – Ph.D., Associate Professor, Honored Worker of Science and High Technologies of the Russian Federation, Senior Researcher, A. P. Ershov Institute of Informatics Systems SB RAS, Associate Professor, Novosibirsk State University, Novosibirsk

email: kev@iis.nsk.su

ORCID 0000-0002-3412-0997

Материал поступил в редакцию 10 января 2026 года

УДК 004.93

ИНТЕЛЛЕКТУАЛЬНЫЙ СЕРВИС МУЛЬТИМОДАЛЬНОГО НЕЙРОСЕТЕВОГО МОНИТОРИНГА ОБЛАСТИ НАБЛЮДЕНИЯ

Р. Р. Миннеахметов[0009-0007-8551-1393]

Казанский (Приволжский) федеральный университет, г. Казань, Россия
razil0071999@gmail.com

Аннотация

Представлен подход к разработке интеллектуального сервиса мультимодального мониторинга области наблюдения с использованием больших нейросетевых моделей. Предлагаемое решение способно анализировать разнородные данные: видеопотоки, сигналы датчиков окружающей среды (температура, влажность и пр.) и журналы событий – для получения целостной картины происходящего. В качестве основных инструментов задействованы крупные языковые и визуальные модели (например, LLaMA, MiniCPM-V и др.), развернутые локально с помощью платформы Ollama, что обеспечивает автономную и безопасную обработку информации без необходимости передачи данных на удаленные сервера. Разработан прототип системы, работающий в офлайн-режиме и способный выявлять критические ситуации, аномальные отклонения от нормы и контекстно значимые события в наблюдаемой зоне. Описана методика формирования тестовых сценариев и проведения качественной оценки работы модели по метрикам F1-мера, Precision, Recall. Результаты экспериментов подтвердили применимость мультимодальных моделей для решения задач мониторинга: прототип успешно распознает сложные паттерны поведения и демонстрирует потенциал больших моделей в построении адаптивных и масштабируемых систем наблюдения.

Ключевые слова*: интеллектуальный сервис, мультимодальный мониторинг, Ollama, большие языковые модели, отслеживание активностей, видеоаналитика, искусственный интеллект.*

ВВЕДЕНИЕ

В современном мире наблюдается стремительный рост объема данных, генерируемых различными сенсорами, системами видеонаблюдения и другими устройствами интернета вещей. Это стимулирует развитие интеллектуальных систем, которые все чаще применяются для анализа поведенческих и ситуационных паттернов в реальном времени. Одним из перспективных направлений в этой области является мультимодальный мониторинг – анализ информации, поступающей одновременно из различных источников (видео, датчики, логи и т. д.) [1]. Благодаря объединению разнородных данных такой подход позволяет получить более полную и достоверную картину происходящего за счет перекрестной верификации сведений из разных модальностей. Под областью наблюдения в контексте настоящей работы понимается ограниченное пространство (физическое или логическое), в котором ведется автоматизированное отслеживание активности. Это может быть помещение, коридор, производственный участок или виртуальная зона, которая контролируется с помощью видеокамер, сенсоров либо систем логирования событий.

Традиционные системы безопасности и мониторинга, как правило, основаны на сигнатурном анализе и ручной настройке правил срабатывания. Они эффективно выявляют известные угрозы, но могут не замечать новые или нетипичные ситуации. В отличие от таких подходов, интеллектуальный сервис, предлагаемый в настоящей работе, использует возможности больших нейросетевых моделей, как языковых, так и визуальных, для автономного анализа происходящих событий. Большие языковые модели (Large Language Models, LLM) и современные модели компьютерного зрения (Vision-модели) демонстрируют высокую эффективность в самых различных задачах: обработке естественного языка, распознавании образов, анализе временных рядов и т. д. Их применение в системах мониторинга позволяет автоматизировать распознавание сложных поведенческих паттернов и потенциально опасных действий, что ранее требовало значительных вычислительных ресурсов и вмешательства человека [2]. В проактивных системах кибербезопасности похожие методы уже используются для поиска аномалий, не выявляемых стандартными средствами защиты [3]. Под ак-

тивностью в общем случае понимается любое значимое изменение в наблюдаемой зоне – будь то событие, зафиксированное системой логирования, либо действие, зафиксированное на видеокамере.

Настоящая работа направлена на создание прототипа интеллектуальной системы мониторинга, способной локально (на персональном устройстве) анализировать мультимодальные данные активности и выявлять важные события. Особое внимание уделено архитектуре и идее системы, основанной на внедрении больших предобученных моделей для анализа нескольких типов данных одновременно, с акцентом на автономность и безопасность обработки. Предварительные результаты работы были представлены в виде доклада на научной конференции «Научный сервис в сети Интернет» [4]; в настоящей статье эти материалы существенно расширены и углублены. Более подробно рассмотрены структура предлагаемого решения, используемые модели и методы, экспериментальные сценарии и полученные результаты.

1. ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

Для эффективного отслеживания активностей в сложных условиях требуется анализировать информацию из различных источников: видеопотоки, системные логи, показания датчиков, а также данные, генерируемые пользователями (User Generated Content). Современные решения предлагают использовать для этого предобученные большие нейросетевые модели, способные извлекать значимые паттерны из разнородных входных данных [5]. В области компьютерного зрения широко применяются глубокие сверточные нейронные сети и трансформеры Vision Transformer для распознавания действий на видео и классификации поведения людей в режиме реального времени [6]. Например, решения на базе CNN и Vision Transformer успешно идентифицируют различные виды человеческой активности на видеозаписях (ходьба, бег, падение и т. п.) и могут обнаруживать отклонения от нормы [7].

Для анализа последовательностей сигналов носимых и окружающих сенсоров (акселерометров, гироскопов, датчиков среды) применяют рекуррентные архитектуры (LSTM/GRU) или трансформеры, обученные на больших массивах данных о движениях людей. Такие модели способны выявлять характерные по-

следовательности сигналов, соответствующие определенным видам активности, и обнаруживать нетипичные паттерны, сигнализирующие о возможном инциденте [7]. В частности, в задачах распознавания человеческой активности (Human Activity Recognition, HAR) по данным носимых устройств большие нейросети достигли заметных успехов [8]. Они позволяют в реальном времени отслеживать показатели движений и состояния здоровья, генерируя уведомления при выявлении опасных событий (например, резкое падение человека, приступ аритмии и т. д.).

Для текстовых данных (таких как журналы событий, протоколы и отчеты) все шире используются большие языковые модели трансформерного типа. Они способны интерпретировать последовательности записей как связный текст и по контексту выявлять аномалии или критические события [9]. В сфере кибербезопасности подобные модели анализируют сетевые логи и сообщения, распознавая характерные предвестники атак и киберинцидентов, что повышает оперативность реагирования на угрозы [10].

Сочетание мультимодального анализа данных на базе больших моделей уже находит применение в различных прикладных областях. В промышленности системы компьютерного зрения на основе глубоких нейросетей контролируют соблюдение техники безопасности на производстве например автоматически обнаруживают отсутствие каски или спецодежды у рабочего [11]. Анализ вибраций и других сенсорных данных станков с помощью рекуррентных нейросетей позволяет реализовать предиктивное обслуживание оборудования (predictive maintenance), выявляя отклонения в работе механизмов и предотвращая аварии [12]. В сфере «умных» домов крупные модели мониторят повседневную активность жильцов для повышения комфорта и безопасности: так, по данным камер и датчиков движения можно определить, что пожилой человек упал, и автоматически вызвать помощь [13]. Носимые фитнес-устройства с интегрированными моделями HAR отслеживают физическую активность и состояние здоровья пользователя, сигнализируя при обнаружении аномалий (например, чрезмерно длительной неподвижности или аритмии) [8]. В системах общественной безопасности нейросетевые алгоритмы видеоаналитики способны распознавать подозрительные действия в режиме реального времени: оставленные без присмотра предметы, агрессивное поведение в толпе, тем самым помогая предот-

вращать правонарушения и инциденты [10]. Еще одним направлением применения крупных моделей являются медицина и здоровье: обработка потоков данных от носимых сенсоров и даже анализ речи/текста пациентов (записи сессий, соцсети) с помощью LLM дают возможность выявлять признаки стресса, депрессии или ухудшения физического состояния на ранних стадиях [2, 13].

Таким образом, достижения последних лет демонстрируют универсальность и эффективность больших нейросетевых моделей в задачах мониторинга: от производственных цехов до домашней обстановки они позволяют повысить качество наблюдения и снизить влияние человеческого фактора. Вместе с тем многие существующие решения либо сфокусированы на одной модальности данных, либо требуют значительных ресурсов и предварительной настройки под конкретные сценарии. Актуальной задачей остается разработка единой интеллектуальной системы, способной интегрировать несколько источников данных и автоматически выявлять сложные ситуации без заранее прописанных правил. В следующем разделе формально описана постановка задачи для такого сервиса.

2. ПОСТАНОВКА ЗАДАЧИ

Цель настоящего исследования состояла в следующем: разработать прототип интеллектуальной системы мониторинга, способной локально в автономном режиме анализировать мультимодальные данные активности (видеоизображения, показания сенсоров, текстовые логи) и своевременно обнаруживать потенциально опасные или аномальные ситуации. В цель работы входила также оценка применимости крупных предобученных нейросетевых моделей для отслеживания различных видов активности в реальных сценариях и выработки соответствующей реакции на выявленные события.

Для достижения этой цели решались следующие задачи: во-первых, реализовать локальное развертывание современных больших моделей (языковых и визуальных) и обеспечить их совместную работу с различными типами входных данных; во-вторых, разработать набор тестовых сценариев, имитирующих типичные ситуации в области наблюдения (чрезвычайные происшествия, нештатные события и штатный режим), чтобы проверить работоспособность си-

стемы; в-третьих, провести сравнительную оценку нескольких моделей по точности распознавания ситуаций и производительности (времени отклика) и на этой основе определить оптимальные решения и узкие места прототипа.

Отметим, что хотя мультимодальные системы теоретически могут включать анализ звука и речи, в рамках настоящей работы аудиомодальность не рассматривается. Это обусловлено, с одной стороны, ограниченной поддержкой аудиовходов в большинстве доступных LLM- и Vision-моделей (на момент исследования), а с другой – отсутствием звуковых данных во многих системах видеонаблюдения (звук обычно не записывается). Тем не менее заложенная архитектура сервиса допускает расширение за счет подключения дополнительных модальностей, включая звук или биометрические датчики, при наличии соответствующих моделей и аппаратуры.

3. ЛОКАЛЬНОЕ РАЗВЕРТЫВАНИЕ МОДЕЛЕЙ С ПОМОЩЬЮ OLLAMA

Для выполнения поставленной задачи было решено использовать локальное развертывание больших нейросетевых моделей, что обеспечивает автономность и конфиденциальность обработки данных. В прототипе использован инструмент Ollama – легковесная платформа, позволяющая запускать различные предобученные LLM- и Vision-модели на персональном компьютере и взаимодействовать с ними через простой интерфейс. Ollama поддерживает современные архитектуры моделей (семейства LLaMA, Mistral и др.) и предоставляет гибкий REST API для их интеграции [14]. Одним из ключевых преимуществ Ollama является возможность полностью локальной работы: все вычисления происходят на стороне пользователя, без отправки входных данных (например, видеокадров или логов) на удаленные серверы. Это особенно важно при работе с чувствительной информацией, требующей соблюдения политики безопасности и приватности [15].

Взаимодействие с моделью в Ollama осуществляется путем отправки HTTP-запросов на локальный сервер (по умолчанию – порт 11434). Запрос формируется в формате JSON и включает обязательные поля:

- **model** – идентификатор выбранной модели (название веса LLM/Vision-модели, загруженной в Ollama);
- **prompt** – текст инструкции или вопрос, передаваемый модели;

- **temperature** – параметр стохастичности генерации (0 – детерминированный вывод, 1 – максимально разнообразный вывод);
- **format** – требуемый формат ответа (например, "text" для обычного текста или "json" для структурированного вывода);
- **stream** – режим выдачи результата (при значении true ответ возвращается по мере генерации, при false – единым блоком) [15].

Правильное составление промпта имеет решающее значение для получения корректного ответа модели. Если запрос сформулирован нечетко или двусмысленно, даже самая мощная модель может выдать неверный или неуместный результат, что снизит качество работы всей системы [16]. В рамках прототипа особое внимание уделялось тому, чтобы промпт ясно описывал модельную задачу: например, содержал инструкции проанализировать конкретные данные и выдать ответ в требуемом формате (структурированном виде). Для удобства интеграции была использована официальная Python-библиотека Ollama [17], предоставляющий высокоуровневые функции для отправки запросов и получения ответов от локального сервера (см. рис. 1 и 2).

```
{
 "model": "llama3",
 "prompt": "Опишите роль нейросетей в современных производственных системах.",
 "temperature": 0.7,
 "format": "json",
 "stream": false
}
```

Рис. 1. Фрагмент запроса к Ollama

```
import ollama
response = ollama.generate(
    model='llama3',
    prompt='Назовите ключевые принципы устойчивости нейронных сетей.',
    options={
        'temperature': 0.5,
        'format': 'json',
        'stream': False
    }
)
print(response['response'])
```

Рис. 2. Запрос к Ollama в Python

На рис. 3 представлен полученный результат, представляющий собой словарь, содержащий поля с метаданными, а также поле response, содержащее сгенерированный моделью текст.


```
{
 "model": "llama3",
 "created_at": "2025-03-24T12:34:56Z",
 "response": "Ключевыми принципами устойчивости нейронных
сетей являются способность к обобщению, толерантность к
шуму, адаптивность и интерпретируемость архитектуры.",
 "done": true
}
```


Рис. 3. Ответ модели

Кроме того, предусмотрена возможность включения дополнительных параметров, позволяющих более тонко настраивать поведение модели:

- top_p – параметр выборки по вероятностному порогу (nucleus sampling);
- num_ctx – максимальное количество токенов контекста;
- repeat_penalty – штраф за повторение одинаковых токенов;
- stop – список токенов-стопов, при достижении которых генерация прекращается [15].

4. ОБРАБОТКА МУЛЬТИМОДАЛЬНЫХ ДАННЫХ

Сервис построен по модульному принципу, где различные типы входных данных преобразуются в удобный для модели вид и объединяются в рамках единого запроса. Общая архитектура прототипа включает следующие компоненты:

- **видео:** периодически из видеопотока (IP-камеры наблюдения или видеозаписи) извлекаются кадры-изображения, которые затем могут быть поданы на вход модели;
- **сенсоры:** показания датчиков (например, температуры и влажности воздуха) агрегируются за небольшие интервалы времени и представляются в текстовом формате. Для эксперимента значения датчиков моделировались: были заданы нормальные и аномальные условия (повышение температуры, снижение влажности как индикатор возможного возгорания);
- **логи:** из внешних систем безопасности или контроля доступа берутся записи журнала событий за недавний промежуток времени. Эти текстовые записи

включают отметки времени (в формате ISO 8601 [18]) и описание произошедших событий (например, срабатывание пожарной сигнализации, отключение датчика и т. д.). Для испытаний был подготовлен образец такого лога (например, фрагмент журнала системы контроля и управления доступом (СКУД)), пригодный для анализа моделью.

Все перечисленные выше данные формируются в единый промпт для модели. Таким образом, на вход модели поступает комплексная информация: одновременно и изображение с камеры, и соответствующие этому моменту показания сенсоров, и текстовые сообщения от других систем. Модель должна на основе всех вводных данных сформировать вывод о ситуации в наблюдаемой зоне. Благодаря использованию мультимодальных возможностей LLM (в частности, моделей, умеющих работать с визуальной информацией) вся аналитика выполняется единым интеллектуальным модулем – без необходимости раздельной обработки каждым источником и последующего объединения результатов. Это упрощает архитектуру и позволяет модели самой учитывать взаимосвязи между различными модальностями данных.

5. ЭКСПЕРИМЕНТАЛЬНАЯ МЕТОДИКА

Для проверки работоспособности прототипа и оценки эффективности разных моделей была разработана методика тестирования на основе нескольких сценариев. Общий процесс эксперимента состоит из трех этапов.

Этап 1. Подготовка тестовых данных. На первом этапе для каждой модальности были сформированы контрольные наборы данных, имитирующие ситуации в области наблюдения. В качестве видеоданных использовались изображения, сгенерированные нейросетью (модель OpenAI ChatGPT-4o-mini [19]), это позволило варьировать содержимое кадров (наличие людей, обстановка) и одновременно избежать использования реальных снимков. Для датчиков были заданы типичные ряды значений: в нормальных условиях – температура ~22 °C и влажность ~45%, в аномальном случае – резкое повышение температуры (до ~60 °C) и понижение влажности (< 20%) как признак возгорания. Кроме того, был подготовлен текстовый лог из системы безопасности: каждая запись

содержала поле timestamp (время события) и поле event (описание самого события) (рис. 4). Такой лог имитировал внешние сигналы, дополняющие данные датчиков и видео.

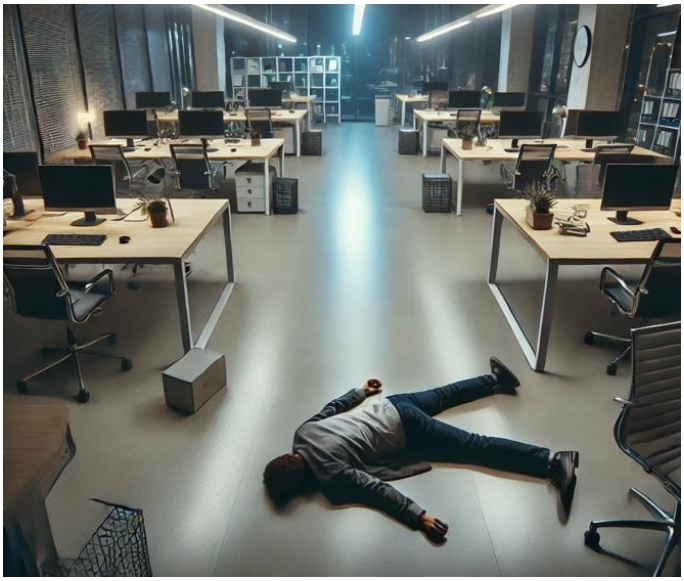


Рис. 4. Сгенерированное фото с камеры видеонаблюдения. Человек упал. Зафиксирована аварийная ситуация.

Рис. 5. Сгенерированное фото с камеры видеонаблюдения. Пустой коридор в офисе. Система видеонаблюдения не обнаружила нарушений.

```
{
  "timestamp": "2025-05-15T12:00:00Z",
  "event": "Fire alarm triggered at Sector 7"
}
```

Рис. 6. Лог с системы безопасности. В данном примере поле timestamp означает время события в формате ISO 8601 [23], а event – описание самого события.

Этап 2. Выбор моделей и сценарии анализа. Для решения задачи были отобраны шесть мультимоделей, доступные в библиотеке Ollama: gemma3:12b [20], llava:13b [21], llama3.2-vision:11b [22], minicpm-v:8b [23], qwen2.5vl:7b [24], mistral-small3.2:24b [25]. Эти модели выбраны исходя из популярности и способности работать с изображениями наравне с текстом [14]. Каждая модель тестировалась на одном и том же наборе из четырех сценариев, заранее подготовленных на этапе 1. Каждый сценарий представлял собой комбинацию данных различных модальностей, соответствующих определенной ситуации.

Сценарий 1: «Человек упал». Видеокадр (условно рис. 4) содержит изображение человека, лежащего на полу без сознания; значения датчиков (рис. 6) находятся в нормальном диапазоне (нет признаков возгорания или других аномалий). Ожидается, что модель, проанализировав картинку, распознает факт падения человека и сформирует вывод о критической ситуации (необходима помощь).

Сценарий 2: «Пожар с людьми». Камера зафиксировала в помещении присутствие людей (рис. 4, на изображении видны люди); датчики (рис. 7) показывают аномальные значения – высокая температура, низкая влажность; в логе присутствует запись о срабатывании пожарной сигнализации. Модель должна учесть все источники: по логам понять, что произошел пожар, по датчикам – подтверждение возгорания, по видео – наличие людей. Ожидаемый вывод: критическая ситуация, в помещении пожар и находятся люди, требуется немедленная реакция.

Сценарий 3: «Пожар без людей». На видеокадре (рис. 5) изображено пустое помещение или коридор; датчики (рис. 7) также сигнализируют о пожаре (высокая температура, сухость), но из логов нет сведений о присутствии людей.

В этом случае модель должна сообщить о пожаре, подчеркнув отсутствие людей (тем не менее ситуация все равно критическая, требует вмешательства, например пожаротушения, но эвакуации людей не требуется).

Сценарий 4: «Штатный режим». Изображение камеры (рис. 5) – пустой коридор, все показатели датчиков в норме (используется тот же набор, что и в сценарии 1, рис. 6), внешних сигналов нет. Это контрольный сценарий благополучного состояния, на который модель не должна выдавать тревожную реакцию (ожидается, что система подтвердит отсутствие подозрительных событий).

Для автоматизации тестирования был написан скрипт на Python, который последовательно подставлял данные каждого сценария в промпт и опрашивал каждую из выбранных моделей через API Ollama. Скрипт измерял время выполнения запроса для каждой модели и сохранял ответы. Чтобы добиться воспроизводимых результатов, параметр temperature для моделей устанавливался равным 0 (детерминированная генерация), а формат ответа задавался как JSON с двумя полями: need_help (логический флаг, сигнализирует о необходимости реагирования) и message (текстовое описание ситуации от лица модели). Таким образом, от каждой модели в каждом сценарии получался структурированный ответ (рис. 7).

```
{
  "need_help": true,
  "message": "Detected an unconscious person, emergency assistance required."
}
```

Рис. 7. Пример структурированного ответа от моделей.

Этап 3. Оценка результатов. На заключительном этапе проводилась оценка качества ответов моделей. Для каждой модели и каждого сценария заранее известен правильный ответ (требуется реакция или нет, корректное описание ситуации). Мы трактовали задачу как бинарную классификацию сценариев на требующие (критические) и нетребующие вмешательства (нормальные). На этой основе вычислялись стандартные метрики: точность (Precision), полнота (Recall) и сводная F1-мера для каждого набора ответов [26, 27]. Кроме того, для практической значимости сравнивалось среднее время отклика различных мо-

делей. В табл. 1 приведены суммарные показатели качества классификации ситуаций для каждой модели, в табл. 2 – среднее время ответа модели на один сценарий.

5. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

По итогам экспериментов получены количественные оценки точности работы мультимодальных моделей в задаче мониторинга активности (выявления критических ситуаций). Табл. 1 демонстрирует сравнение методов с помощью метрик F1, Precision и Recall для шести моделей. Табл. 2 содержит данные о производительности – время, затрачиваемое моделями на обработку одного сценария в среднем (в секундах).

Табл. 1. Качество определения критической ситуации
(средние значения метрик по результатам 4-х сценариев)

Модель	Precision	Recall	F1-Score
gemma3:12b	1.00	0.67	0.80
llava:13b	0.00	0.00	0.00
minicpm-v:8b	1.00	0.50	0.67
qwen2.5vl:7b	0.00	0.00	0.00
mistral-small3.2:24b	1.00	0.67	0.80
llama3.2-vision:11b	1.00	0.50	0.67

Табл. 2. Время отклика моделей на один сценарий (в секундах)

Модель	Сценарий 1	Сценарий 2	Сценарий 3	Сценарий 4
gemma3:12b	17.06	6.42	5.89	6.11
llava:13b	21.74	12.20	10.61	9.32
minicpm-v:8b	5.87	1.84	3.61	1.49
qwen2.5vl:7b	20.41	18.20	17.97	18.37
mistral-small3.2:24b	43.54	40.89	37.45	37.91
llama3.2-vision:11b	31.92	29.33	28.70	29.91

Таким образом, из полученных результатов видно, лучшими оказались gemma3:12b и mistral-small3.2: они правильно отреагировали на три из четырех сценариев, что подтверждается наибольшим значением F1 = 0.8. Кроме того,

они не допустили ни одного ложного срабатывания (Precision = 1.00), хотя и пропустили один из критических сценариев (Recall = 0.67). Модели minicpm-v:8b и llama3.2-vision:11b также продемонстрировали вполне приемлемую точность (F1 = 0.67), без ложных тревог, но с пропуском части инцидентов. Наихудший результат оказался у моделей llava:13b и qwen2.5vl:7b – они не смогли корректно идентифицировать ни одного сценария (все выходные ответы были ошибочными), о чем говорят нулевые значения метрик. Вероятно, модели оказались наименее подходящими для подобных комплексных запросов, возможно, из-за ограниченной специализации или недостаточной обучения для интеграции различных типов данных.

Что касается скорости работы, здесь лидирует облегченная модель minicpm-v:8b – ее среднее время отклика в простых сценариях составляло порядка 1.7 с (сценарии 2 и 4), и даже в более сложных ситуациях (сценарии 1 и 3) она укладывалась в 6 с. Модель gemma3:12b показывала стабильное время ответа около 10–20 с на сценарий, что быстрее тяжелых mistral-small3.2:24b и llama3.2-vision (на отдельных задачах время доходило до 45 с). Модели llava:13b и qwen2.5vl:7b в целом работали сравнительно быстро (до 21 с), однако их низкие точности делают скорость несущественным фактором. Следует отметить, что все модели запускались на одной локальной машине, поэтому абсолютизировать приведенные цифры не стоит – при развертывании на производительном оборудовании время реакции может быть значительно снижено.

В целом эксперимент подтвердил возможность применения больших мультимодальных моделей для мониторинга: по крайней мере две из проверенных моделей (gemma3:12b и minicpm-v:8b) сумели обнаружить большинство заданных событий, правильно интерпретировав и совместив информацию из разных источников. Это весьма обнадеживающий результат, учитывая, что модели не проходили специального обучения под наши сценарии, а использовались «как есть». Таким образом, нейросети, предобученные на больших данных, в сочетании с грамотной инженерией промптов могут успешно решать задачи интеллектуального анализа ситуации.

Однако эксперимент выявил и ряд ограничений текущего прототипа. Во-первых, качество вывода значительно варьируется от модели к модели: выбор подходящей архитектуры критически влияет на точность. Модели, лучше настроенные на визуально-текстовый ввод (например, gemma3:12b), показали

высокую результативность, тогда как другие оказались неприменимы в данном виде. Во-вторых, производительность системы пока оставляет желать лучшего – время отклика в десятки секунд неприемлемо для ряда практических сценариев (например, для систем реального времени, где счет может идти на секунды). Это частично связано с использованием больших моделей (12–13 млрд параметров) на CPU; ускорение возможно при переходе на GPU-версии или при оптимизации моделей (квантование, аппаратное ускорение). В-третьих, прототип был протестирован на ограниченном наборе синтетических данных. Отметим, что выбранные сценарии были приближены к реальным условиям, на практике могут возникать более сложные обстановки, шумы и непредусмотренные комбинации событий, где поведение модели потребует дополнительной проверки.

Тем не менее применимость в реальных условиях представляется вполне вероятной после доработки системы. Одним из преимуществ предложенного подхода является его гибкость: путем замены или обновления модели в Ollama можно улучшить показатели без кардинальной переработки всей системы. Кроме того, локальное исполнение гарантирует, что чувствительные видеоданные и логи не покидают пределов устройства/локальной сети – это важно для организаций, предъявляющих строгие требования к безопасности данных (например, на производствах с режимом секретности или в медицинских учреждениях). Автономность решения означает и независимость от сетевой инфраструктуры: мониторинг не прервется даже при остановится доступа к Интернету или облаку.

ЗАКЛЮЧЕНИЕ

Разработанный прототип интеллектуального сервиса мультимодального мониторинга продемонстрировал перспективность применения больших предобученных нейросетевых моделей в системах отслеживания активности. В ходе экспериментов показано, что современные модели способны интегрированно анализировать данные разных типов (изображения, числовые показатели, текстовые события) и успешно выявлять сложные ситуации, ранее обнаруживаемые лишь человеком или узкоспециализированными алгоритмами. Использование локального фреймворка Ollama позволило запускать LLM-модели непосредственно на месте сбора данных, обеспечивая автономную работу системы

и защиту информации. Полученные результаты подтверждают работоспособность подхода: наиболее точная модель (gemma3:12b) правильно распознала 75% сценариев, а более быстрая minicpm-v:8b при незначительном снижении полноты также может считаться успешной.

Вместе с тем проведенное исследование выявило и направления для дальнейшей работы. Одной из первоочередных задач является **повышение быстродействия**: планируется оптимизировать модели (например, за счет квантования до меньшей разрядности, использования версий «LoRA» или дистилляции знаний) и протестировать их на аппаратном ускорителе, чтобы добиться сокращения времени реакции до приемлемых величин. Еще одно перспективное направление – это обогащение интеллектуального анализа с помощью **онтологической поддержки сценариев**. Введение семантической модели предметной области (онтологии событий и объектов) могло бы помочь интерпретировать ответы модели и уменьшить вероятность ошибок, особенно в нетипичных случаях. Кроме того, интеграция дополнительных модальностей (например, аудио, как обсуждалось выше) расширит возможности мониторинга: звук и речь могут предоставить ценные сведения о происходящем (крики, шумы аварий и пр.).

Настоящая работа выполнялась в рамках инициативного исследования, без прямого привлечения сторонних организаций. Первичное внедрение прототипа планируется осуществить на учебно-исследовательских стендах Казанского федерального университета, что позволит собрать дополнительную обратную связь и улучшить систему. В перспективе доработанное решение может быть опробовано в условиях, приближенных к промышленным, например в лабораториях или в рамках пилотного проекта на предприятии, заинтересованном в интеллектуальных системах безопасности. Таким образом, разработанный сервис представляет собой шаг вперед к созданию универсальных мультимодальных средств мониторинга, объединяющих достижения в области больших моделей с практическими требованиями автономности и безопасности.

СПИСОК ЛИТЕРАТУРЫ

1. *Onsu M.A., Lohan P., Kantarci B., Syed A., Andrews M., Kennedy S.* Leveraging Multimodal-LLMs Assisted by Instance Segmentation for Intelligent Traffic Monitoring [Электронный ресурс] // arXiv, 2025.
URL: https://arxiv.org/abs/2502.11304 (дата обращения: 15.05.2025).

2. *Ferrara E.* Large Language Models for Wearable Sensor-Based Human Activity Recognition, Health Monitoring, and Behavioral Modeling // Sensors. 2024. Vol. 24, No. 15. Article 5045.

3. *Suh S., Rey V.F., Lukowicz P.* Tasked: Transformer-based adversarial learning for human activity recognition using wearable sensors // Knowledge-Based Systems. 2023. Vol. 260. Article 110143.

4. Научный сервис в сети Интернет: труды XXVI Всероссийской научной конференции (22–25 сентября 2025 г., онлайн). // М.: ИПМ им. М.В. Келдыша, 2025.

5. *Nath N.D., Behzadan A.H., Paal S.G.* Deep learning for site safety: Real-time detection of personal protective equipment // Automation in Construction. 2020. Vol. 112. Article 103085.

6. *Gupta S.* Deep learning-based human activity recognition using wearable sensor data // International Journal of Information Management Data Insights. 2021. Vol. 1. Article 100046.

7. *Uçar A., Karakoşe M., Kırımça N.* Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends // Applied Sciences. 2024. Vol. 14, No. 2. Article 898.

8. *Wu Z., Zhao J., Shen H.* Smart home automation based on human activity recognition: A survey // Future Generation Computer Systems. 2023. Vol. 137. P. 41–57.

9. *Han S., Yuan S., Trabelsi M.* LogGPT: Log Anomaly Detection via GPT [Электронный ресурс] // arXiv. 2023. URL: https://arxiv.org/pdf/2309.14482 (дата обращения: 15.05.2025).

10. *Sharma R., Patel N.* Deep learning-based anomaly detection in surveillance videos // Journal of Visual Communication and Image Representation. 2022. Vol. 86. Article 103624.

11. *Özüağ S., Ertuğrul Ö.* Enhanced Occupational Safety in Agricultural Machinery Factories: Artificial Intelligence-Driven Helmet Detection Using Transfer Learning and Majority Voting // Applied Sciences. 2024. Vol. 14. Article 11278. https://doi.org/10.3390/app142311278

12. *Li X., Chen Y., Hu L.* Real-time workplace activity recognition using deep learning models // IEEE Transactions on Industrial Informatics. 2023. Vol. 19, No. 2. P. 1520–1532.

13. *Wu Z., Zhao J., Shen H.* Smart home automation based on human activity recognition: A survey // Future Generation Computer Systems. 2023. Vol. 137. P. 41–57.

14. Ollama: [Электронный ресурс].
URL: https://ollama.com/ (дата обращения: 30.03.2025).

15. Ollama API Documentation: [Электронный ресурс].
URL: https://github.com/ollama/ollama/blob/main/docs/api.md (дата обращения: 30.03.2025).

16. *Sahoo P., Singh A.K., Saha S., Jain V., Mondal S., Chadha A.* A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications [Электронный ресурс] // arXiv. 2024. URL: https://arxiv.org/pdf/2402.07927 (дата обращения: 15.05.2025).

17. Ollama Python Library: [Электронный ресурс].
URL: https://github.com/ollama/ollama-python (дата обращения: 30.03.2025).

18. ISO 8601-1:2019 Standard: [Электронный ресурс].
URL: https://www.iso.org/obp/ui/#iso:std:iso:8601:-1:ed-1:v1:en (дата обращения: 30.03.2025).

19. OpenAI ChatGPT-4o-mini: [Электронный ресурс].
URL: https://chatgpt.com/ (дата обращения: 30.03.2025).

20. Ollama gemma3:12b Model: [Электронный ресурс].
URL: https://ollama.com/library/gemma3:12b (дата обращения: 30.03.2025).

21. Ollama llava:13b Model: [Электронный ресурс].
URL: https://ollama.com/library/llava:13b (дата обращения: 30.03.2025).

22. Ollama llama3.2-vision:11b Model: [Электронный ресурс].
URL: https://ollama.com/library/llama3.2-vision (дата обращения: 30.03.2025).

23. Ollama minicpm-v:8b Model: [Электронный ресурс].
URL: https://ollama.com/library/minicpm-v (дата обращения: 30.03.2025).

24. Ollama qwen2.5vl:7b Model: [Электронный ресурс].
URL: https://ollama.com/library/qwen2.5vl (дата обращения: 16.01.2026).

25. Ollama mistral-small3.2 Model: [Электронный ресурс].
URL: https://ollama.com/library/mistral-small3.2 (дата обращения: 16.01.2026).

26. *Hand D.J., Christen P.* F*: an interpretable transformation of the F-measure // Journal of Classification. 2021. Vol. 38, No. 1. P. 3–17.

27. Scikit Learn F1-Score: [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (дата обращения: 30.03.2025).

INTELLIGENT MULTIMODAL NEURAL NETWORK MONITORING SERVICE FOR THE SURVEILLANCE AREA

R. R. Minneakmetov[0009-0007-8551-1393]

Kazan (Volga Region) Federal University, Kazan, Russia

razil0071999@gmail.com

Abstract

The article presents an approach to the development of an intelligent multimodal monitoring service for the surveillance area using large neural network models. The proposed solution is capable of analyzing heterogeneous data – video streams, environmental sensor signals (temperature, humidity, etc.), and event logs – to obtain a complete picture of what is happening. The main tools used are large language and visual models (for example, LLaMA, MiniCPM-V, etc.) deployed locally using the Ollama platform, which provides autonomous and secure information processing without the need to transfer data to the cloud. A prototype system has been developed that works offline and is capable of detecting critical situations, abnormal deviations from the norm and contextually significant events in the observed area. The method of forming test scenarios and conducting a qualitative assessment of the model's performance using the metrics F1-measure, Precision, Recall on a set of various situations is described. The experimental results confirm the applicability of multimodal models for monitoring tasks: the prototype successfully recognizes complex patterns of behavior and demonstrates the potential of large models in building adaptive and scalable surveillance systems.

Keywords*: intelligent service, multimodal monitoring, Ollama, Large Language Models, activity tracking, video analytics, artificial intelligence.*

REFERENCES

1. *Onsu M.A., Lohan P., Kantarci B., Syed A., Andrews M., Kennedy S.* Leveraging Multimodal Large Language Models Assisted by Instance Segmentation for Intelligent Traffic Monitoring [Electronic resource] // arXiv. 2025. Available at: https://arxiv.org/abs/2502.11304 (accessed: 15.05.2025).

2. *Ferrara E.* Large Language Models for Wearable Sensor-Based Human Activity Recognition, Health Monitoring, and Behavioral Modeling // Sensors. 2024. Vol. 24, No. 15. Article 5045.

3. *Suh S., Rey V.F., Lukowicz P.* Tasked: Transformer-Based Adversarial Learning for Human Activity Recognition Using Wearable Sensors // Knowledge-Based Systems. 2023. Vol. 260. Article 110143.

4. Nauchnyy servis v seti Internet: trudy XXVI Vserossiyskoy nauchnoy konferentsii (September 22–25, 2025, online). Moscow: Keldysh Institute of Applied Mathematics, 2025 (in press).

5. *Nath N.D., Behzadan A.H., Paal S.G.* Deep Learning for Site Safety: Real-Time Detection of Personal Protective Equipment // Automation in Construction. 2020. Vol. 112. Article 103085.

6. *Gupta S.* Deep Learning-Based Human Activity Recognition Using Wearable Sensor Data // International Journal of Information Management Data Insights. 2021. Vol. 1. Article 100046.

7. *Uçar A., Karakoşe M., Kırımça N.* Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends // Applied Sciences. 2024. Vol. 14, No. 2. Article 898.

8. *Wu Z., Zhao J., Shen H.* Smart Home Automation Based on Human Activity Recognition: A Survey // Future Generation Computer Systems. 2023. Vol. 137. P. 41–57.

9. *Han S., Yuan S., Trabelsi M.* LogGPT: Log Anomaly Detection via GPT [Electronic resource] // arXiv. 2023. Available at: https://arxiv.org/pdf/2309.14482 (accessed: 15.05.2025).

10. *Sharma R., Patel N.* Deep Learning-Based Anomaly Detection in Surveillance Videos // Journal of Visual Communication and Image Representation. 2022. Vol. 86. Article 103624.

11. *Özüağ S., Ertuğrul Ö.* Enhanced Occupational Safety in Agricultural Machinery Factories: Artificial Intelligence-Driven Helmet Detection Using Transfer Learning and Majority Voting // Applied Sciences. 2024. Vol. 14. Article 11278. https://doi.org/10.3390/app142311278.

12. *Li X., Chen Y., Hu L.* Real-Time Workplace Activity Recognition Using Deep Learning Models // IEEE Transactions on Industrial Informatics. 2023. Vol. 19, No. 2. P. 1520–1532.

13. *Wu Z., Zhao J., Shen H.* Smart Home Automation Based on Human Activity Recognition: A Survey // Future Generation Computer Systems. 2023. Vol. 137. P. 41–57.

14. Ollama [Electronic resource]. Available at: https://ollama.com/ (accessed: 30.03.2025).

15. Ollama API Documentation [Electronic resource]. Available at: https://github.com/ollama/ollama/blob/main/docs/api.md (accessed: 30.03.2025).

16. *Sahoo P., Singh A.K., Saha S., Jain V., Mondal S., Chadha A.* A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications [Electronic resource] // arXiv. 2024. Available at: https://arxiv.org/pdf/2402.07927 (accessed: 15.05.2025).

17. Ollama Python Library [Electronic resource]. Available at: https://github.com/ollama/ollama-python (accessed: 30.03.2025).

18. ISO 8601-1:2019 Standard [Electronic resource]. Available at: https://www.iso.org/obp/ui/#iso:std:iso:8601:-1:ed-1:v1:en (accessed: 30.03.2025).

19. OpenAI ChatGPT-4o-mini [Electronic resource]. Available at: https://chatgpt.com/ (accessed: 30.03.2025).

20. Ollama Gemma3:12B Model [Electronic resource]. Available at: https://ollama.com/library/gemma3:12b (accessed: 30.03.2025).

21. Ollama LLaVA:13B Model [Electronic resource]. Available at: https://ollama.com/library/llava:13b (accessed: 30.03.2025).

22. Ollama Llama3.2-Vision:11B Model [Electronic resource]. Available at: https://ollama.com/library/llama3.2-vision (accessed: 30.03.2025).

23. Ollama MiniCPM-V:8B Model [Electronic resource]. Available at: https://ollama.com/library/minicpm-v (accessed: 30.03.2025).

24. Ollama Qwen2.5-VL:7B Model [Electronic resource]. Available at: https://ollama.com/library/qwen2.5vl (accessed: 16.01.2026).

25. Ollama Mistral-Small-3.2 Model [Electronic resource]. Available at: https://ollama.com/library/mistral-small3.2 (accessed: 16.01.2026).

26. *Hand D.J., Christen P.* F*: An Interpretable Transformation of the Measure // Journal of Classification. 2021. Vol. 38, No. 1. P. 3–17.

27. Scikit-learn F1-Score [Electronic resource]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (accessed: 30.03.2025).

СВЕДЕНИЯ ОБ АВТОРЕ

МИННЕАХМЕТОВ Разиль Рустемович – магистр программной инженерии, аспирант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Razil MINNEAKHMETOV – Master of Software Engineering, PhD student at the Institute of Information Technology and Intelligent Systems, Kazan (Volga Region) Federal University. Current scientific interests: artificial intelligence, large neural models, recommender systems, cloud computing, internet of things.

email: razil0071999@gmail.com

ORCID: 0009-0007-8551-1393

Материал поступил в редакцию 16 января 2026 года

УДК 004.451

РЕАЛИЗАЦИЯ ОДНОГО РЕШЕНИЯ ПРИ ПЕРЕХОДЕ С CENTOS НА RED OS ДЛЯ КЛАСТЕРА ВЫСОКОЙ ДОСТУПНОСТИ

Г. М. Михайлов[1] [0000-0002-4535-7180], Н. П. Тучкова[2] [0000-0001-5357-9640], А. М. Чернецов[3] [0000-0001-7655-2395]

[1,2]*Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия*

[3]*Национальный исследовательский университет «МЭИ», г. Москва, Россия*

[1]gmihaylov@frccsc.ru, [2]ntuchkova@frccsc.ru, [3]chernetsovam@mpei.ru

Аннотация

Представлен краткий аналитический обзор популярных отечественных дистрибутивов операционных систем, разработанных в рамках реализации задач технологической независимости в области программного обеспечения и средств телекоммуникации. Описано одно из решений перехода с системы CentOS на систему РЕД ОС (RED OS) для кластера высокой доступности на базе Pacemaker и распределенной файловой системы DRBD, обеспечившего работу сайта организации и сервера баз данных MySQL.

***Ключевые слова**: импортозамещение, РЕД ОС-сертифицированная, Pacemaker.*

ВВЕДЕНИЕ

Краткую историю развития отечественных операционных систем (ОС) можно условно начинать с февраля 2000 года, когда появилась первая ОС Альт Линукс [1]. Вслед за ней последовали известные теперь Астра Линукс (2009 г.) [2], РЕД ОС (2017 г.) [3] и др. Целевой потребностью указанных программных продуктов являлись специализированные поставки для разработки систем ограниченного доступа. В этом перечне отсутствуют специализированные системы для нужд Минобороны типа МСВСфера (1999 г.). Важно отметить, что каждая из основных отечественных ОС, занимающих большую часть современного рынка, имеет свою историю. Все указанные системы имели различные сертификации соответствующих контрольных органов. Появление

Указа Президента Российской Федерации от 30.03.2022 № 166 [4] оказало огромное влияние на развитие этого сегмента отечественных информационных технологий. Заметим, что Указом Президента РФ от 01.05.2022 (редакция от 13.06.2024 [5]) запрещено использовать иностранное программное обеспечение (ПО) на значимых объектах критических информационных инфраструктур, а также средства защиты информации из некоторых государств.

К большому сожалению, в настоящее время в научной литературе наблюдается значительный дефицит работ по тематике импортозамещения операционных систем. Одна из работ, посвященная вопросам настройки и использования операционной системы «Альт» для образовательного процесса, представлена в [6].

ОСОБЕННОСТИ ОС CentOS

CentOS (от англ. Community Enterprise Operating System) – это дистрибутив Linux, основанный на коммерческом Red Hat Enterprise Linux компании Red Hat. Долгое время компания RedHat поддерживала стабильные выпуски, имеющие достаточно большой срок поддержки (до 7 лет) и пользующиеся большой популярностью у конечных пользователей и разработчиков. Однако в 2021 г. разработчиком было принято решение о смене модели поддержки, в результате чего ОС CentOS 8 прекратила обновление, а следующая версия CentOS Stream превратилась в «платформу тестирования». В такой модели все непроверенные и нетестированные обновления выходили сразу после создания «как есть», что исключило возможность использования их в разрабатываемых масштабных проектах. В настоящее время CentOS 7.9 – это ОС с устаревшей версией ядра Linux 3.10 и без обновлений безопасности. На начало 2024 г., по данным [7], доля CentOS 7.9 составляла более 26% установок ОС Linux, поэтому переход на современные версии ОС Linux стала важной и актуальной задачей.

ОСОБЕННОСТИ ОПЕРАЦИОННЫХ СИСТЕМ АЛЬТ ЛИНУКС, АСТРА ЛИНУКС, РЕД ОС

Отечественные ОС были разработаны в период локализации программного обеспечения, появления русскоязычных вариантов и аналогов известных продуктов для работы с базами данных. Среди них с начала 2010-х годов развивались системы ОС Альт Линукс (англ. аббревиатура Alt Linux), Астра Линукс (англ. аббревиатура Astra Linux) и далее (по порядку выхода) РЕД ОС (англ. аббревиатура RED OS).

Рассмотрим некоторые особенности перехода со свободно распространяемых версий Linux (CentOS, Debian, Ubuntu) на отечественные. Сразу следует отметить, что изначально ОС РЕД ОС создана на базе Centos, Astra Linux – на базе Debian, а Alt Linux – на базе Mandriva. При этом заметим, что через какое-то время после начала разработки все ОС обзавелись собственным ядром, не привязанным к системе-родителю. Процесс обновления ядра ОС на более новую версию происходит совершенно независимо. Тем не менее исторически наблюдается некоторое отставание между версиями в СПО (свободное программное обеспечение) и отечественными дистрибутивами. Для сертифицированных версий отставание составляет еще больше: в пределах двух-трех лет.

На момент начала настоящей работы самое стабильное ядро Linux имело версию 6.6.120, в популярных зарубежных дистрибутивах поддерживались ядра версий от 6.1 до 6.6. В табл. 1 представлена актуальная информация о версиях ядра, доступных в дистрибутивах отечественных ОС.

Табл. 1. Версии ядра, доступные в дистрибутивах отечественных ОС

Дистрибутив	Версия ядра
Astra Linux 1.8	6.1 (LTS) и 6.6
RED OS 8	6.6.31
Alt Linux	6.12 (LTS)
Astra Linux Special Edition	6.1
RED OS 8-сертифицированный	6.12.21
Alt Linux-сертифицированный	6.12 (LTS)

Ядро версии 6.6 следует использовать для поддержки нового оборудования. Ядро версии 6.1 (LTS) применяют для стабильных релизов, на что стоит опираться разработчикам производителей ПО и средств защиты информации (СЗИ).

С точки зрения разработчика любая из этих отечественных ОС позволяет решать широкий круг стандартных задач, как и любая Linux-подобная ОС. Какие-либо особенности возникают только для специализированных задач.

Сравнив обобщенно между собой выпуски указанных ОС, можно отметить следующее: Астра Линукс – более специализированное ПО для критических систем, которое обеспечивает максимальный уровень защиты. РЕД ОС имеет более простой интерфейс для Windows-пользователей, переходящих на нее. Альт Линукс позволяет организовать максимальную гибкость.

Если рассматривать более общую задачу перехода с ОС семейства Windows на ОС Linux, то часть графических приложений может быть легко перенесена с использованием эмулятора Wine [8]. Есть приложения (например, написанные на Powershell), которые запускаются сразу. Есть ряд приложений, которые целесообразно перевести/написать/разрабатывать на кроссплатформенной среде, например, на языке Qt [9].

ПЕРЕХОД ОТ CentOS К РЕД ОС

Рассмотрим конкретную задачу перехода: с ОС CentOS 7.9 на РЕД ОС 7.3. В работе [10] была рассмотрена «Информационная система приемной комиссии НИУ «МЭИ» (ИСПК)», основанная на клиент-серверной технологии в виде веб-приложения. Типовыми элементами ИСПК являются веб-сервер Apache и СУБД Mysql Community Edition. Отметим также, что ИСПК для повышения надежности была развернута с использованием технологий Pacemaker и DRBD (Distributed Replicated Block Device – распределенное реплицируемое блочное устройство) [11, 12].

Поскольку обработка персональных данных требует работы со специальными версиями ПО, необходимо выбирать решения, сертифицированные ФСТЭК. Такими могут быть как внешние решения (СЗИ, например, Dallas Lock [13] или SecretNet Studio [14]), так и специальные версии

операционных систем. Для Ред ОС такой сертифицированной версией на настоящее время является Ред ОС 7.3.

Как известно, использование некоторого ПО, в частности MySQL Server всех выпусков, кроме Community Edition, ограничено. Для замены рекомендуется использовать MariaDB. В случае ИСПК необходимости такой замены нет, так как Community Edition уже присутствует и используется в системе.

Для переноса узла Pacemaker в целом были выполнены следующие действия:

1) сохранение конфигурации кластера средствами Pacemaker;
2) подготовка чистого сервера (форматирование узла, изменение типа ФС для ОС);
3) развертывание Ред ОС 7.3 сертифицированная на узле;
4) установка необходимых пакетов для работы приложений;
5) отключение SELinux;
6) установка Pacemaker, создание узла из сохраненной конфигурации.

В качестве первого эксперимента был собран кластер Pacemaker из двух узлов, в котором первый узел был сделан на базе CentOS 7, а второй – на базе РЕД ОС 7.3-сертифицированная. При этом работа кластера не прерывалась, его ресурсы были полностью доступны. Было обеспечено взаимодействие между узлами на разных системах – CentOS 7 и РЕД ОС 7.3-сертифицированная. В результате было установлено, что версии Pacemaker в CentOS и РЕД ОС хотя и различны, но достаточно совместимы, чтобы провести миграцию без особых проблем. С DRBD также не возникло сложностей, за исключением того, что из-за различных версий компонента Corosync пришлось вручную пересобрать часть конфигурации на новом узле, сделав ее идентичной сохраненной в п. 1.

Добавим для уточнения, что Corosync – это программный продукт, который позволяет создавать единый кластер из нескольких аппаратных или виртуальных серверов. Corosync отслеживает и передает состояние всех участников (нод) в кластере. Этот продукт позволяет:

- мониторить статус приложений;
- оповещать приложения о смене активной ноды в кластере;
- отправлять идентичные сообщения процессам на всех нодах;

- предоставлять доступ к общей базе данных с конфигурацией и статистикой;
- отправлять уведомления об изменениях, произведенных в базе.

Во втором эксперименте узлы «поменялись местами», и прошла замена ОС второго узла. В результате в оба узла кластера Pacemaker была загружена отечественная РЕД ОС 7.3-сертифицированная. На этом запланированная работа была завершена, и обновленная система ИСПК в рамках импортозмещения вошла в режим надлежащего функционирования.

ЗАКЛЮЧЕНИЕ

Учитывая необходимость переноса многочисленного ПО в разных предметных областях на отечественные разработки, в настоящее время стала крайне актуальной задача импортозамещения программного обеспечения. Есть ряд приложений, которые легко или достаточно просто переносятся на отечественное ПО. Эти приложения чаще всего уже заранее реализованы на СПО. В качестве примеров можно привести сервер СУБД Mysql, веб-сервер Apache, кластер высокой доступности (Pacemaker). В то же время перенос большинства ПО, особенно с проприетарных ОС, обоснованно представляет большую проблему. В общегосударственном масштабе планируемые решения регулируются приказом Минкомсвязи [15].

Благодарности

Работа выполнена в рамках исполнения темы FFNG-2024-0003.

СПИСОК ЛИТЕРАТУРЫ

1. Альт Линукс домашняя страница. URL: https://www.basealt.ru/alt-server.

2. Домашняя страница ОС Астра Линукс. URL: https://astralinux.ru/os/

3. Домашняя страница ОС РЕД ОС. URL: https://redos.red-soft.ru/.

4. Указ Президента Российской Федерации от 30.03.2022 № 166 «О мерах по обеспечению технологической независимости и безопасности критической информационной инфраструктуры Российской Федерации». http://publication.pravo.gov.ru/Document/View/0001202203300001.

5. Указ Президента РФ от 01.05.2022 № 250 — Редакция от 13.06.2024 — Контур.Норматив. «О дополнительных мерах по обеспечению информационной безопасности Российской Федерации».

URL: https://normativ.kontur.ru/document?moduleId=1&documentId=
473111&ysclid =miws3qe2aq231571505.

6. *Стрелков Н.О.* Настройка и использование операционной системы «Альт» для образовательного процесса // Информатизация инженерного образования: Материалы VII Международной научно-практической конференции, Москва, 16–19 апреля 2024 года. Москва: Федеральное государственное бюджетное образовательное учреждение высшего образования Национальный исследовательский университет МЭИ, 2024. С. 55–60. EDN JYPFST.

7. Статистика использования Centos 7.
URL: https://www.lansweeper.com/blog/eol/centos-linux-end-of-life/.

8. Эмулятор Wine. URL: https://www.winehq.org.

9. Кросс-платформенная среда Qt.
URL: https://doc.qt.io/qt-6/reference-overview.html.

10. *Васьковский А.А., Крупин Г.В., Наумова Ю.Д., Титов Д.А., Чернецов А.М*. Специфика организации информационной системы приема поступающих в крупных вузах // Вестник МЭИ. 2020. № 2. С. 106–112.
https://doi.org/10.24160/1993-6982-2020-2-106-112

11. *Михайлов Г.М., Жижченко М.А., Чернецов А.М.* Применение Pacemaker для повышения надежности доступа к критическим данным // Научный сервис в сети Интернет: труды XXIII Всероссийской научной конференции (20–23 сентября 2021 г., онлайн). М.: ИПМ им. М.В. Келдыша, 2021. С. 228–235.
https://doi.org/10.20948/abrau-2021-16

12. *Chernetsov A., Shamayeva O., Mikhailov G.* Improve the Reliability of the Organization’s Resources to Support Remote Education Services // VI International Conference on Information Technologies in Engineering Education (Inforino), Moscow, Russian Federation, 2022, pp. 1–4.
https://doi.org/10.1109/Inforino53888.2022.9783000

13. Домашняя страница СЗИ DalasLock Linux.
URL: https://dallaslock.ru/products/szi-nsd-dallas-lock-linux/.

14. Домашняя страница СЗИ SecretNet Studio.
URL: https://www.securitycode.ru/products/szi_secret_net/?tab=system.

15. Приказ Минкомсвязи России от 01.04.2015 N 96 Об утверждении плана импортозамещения программного обеспечения.
URL: https://digital.gov.ru/documents/prikaz-minkomsvyazi-rossii-ob-utverzhdenii-plana-importozameshheniya-programmnogo-obespecheniya.

IMPLEMENTATION OF ONE SOLUTION WHEN MIGRATING FROM CENTOS TO RED OS FOR A HIGH AVAILABILITY CLUSTER

G. M. Mikhaylov[1] [0000-0002-4535-7180], N. P. Tuchkova[2] [0000-0001-5357-9640], A. M. Chernetsov[3] [0000-0001-7655-2395]

[1, 2]*Federal Research Center "Computer Science and Control" RAS, Moscow, Russia*
[3]*National Research University "MPEI", Moscow, Russia*

[1]gmickail@ccas.ru, [2]ntuchkova@frccsc.ru,[3]chernetsovam@mpei.ru

Abstract

This paper presents a brief overview of popular domestic OS distributions developed as part of the implementation of import substitution tasks in the field of software and telecommunications. One of the solutions for the transition from CentOS to RED OS is presented for a High Availability cluster based on Pacemaker and the DRBD distributed file system, which ensures the operation of the organization's website and MySQL database server.

Keywords*: import substitution, RED OS Certified, Pacemaker.*

REFERENCES

1. Alt Linux Homepage.URL: https://www.basealt.ru/alt-server.

2. OS Astra Linux Homepage. URL: https://astralinux.ru/os/.

3. OS RED OS Homepage. URL: https://redos.red-soft.ru/.

4. Decree of the President of the Russian Federation of March 30, 2022, No. 166 "On measures to ensure the technological independence and security of the critical information infrastructure of the Russian Federation."
http://publication.pravo.gov.ru/Document/View/0001202203300001.

5. Decree of the President of the Russian Federation No. 250 of May 1, 2022, as amended on June 13, 2024, Kontur.Normativ". On Additional Measures to Ensure the Information Security of the Russian Federation."
URL: https://normativ.kontur.ru/document?moduleId=1&documentId=473111&ysclid=miws3qe2aq231571505.

6. *Strelkov N.O.* Setting up and using the Alt operating system for the educational process // Information Technologies in Engineering Education: Proceedings of the VII International scientific and practical conference, Moscow, April 16–19, 2024. Moscow: MPEI, 2024. P. 55–60. EDN JYPFST.

7. Centos 7 usage statistics.
URL: https://www.lansweeper.com/blog/eol/centos-linux-end-of-life/.

8. Wine Emulator. URL: https://www.winehq.org.

9. Cross-platform environment Qt.
URL: https://doc.qt.io/qt-6/reference-overview.html.

10. *Vaskovsky A.A., Krupin G.V., Naumova Yu.D., Titov D.A., Chernetsov A.M.* Specifics of organizing an information system for admission of applicants to large universities // Vestnik MPEI. 2020. № 2. P. 106–112.
https://doi.org/10.24160/1993-6982-2020-2-106-112

11. *Mikhailov G.M., Zhizhchenko M.A., Cherneczov A.M*. Using Pacemaker to Improve Reliability of Access to Critical Data // Scientific Service on the Internet: Proceedings of the XXIII All-Russian Scientific Conference (September 20–23, 2021, online). Moscow: Keldysh Institute of Applied Mathematics, 2021. P. 228–235.
https://doi.org/10.20948/abrau-2021-16

12. *Chernetsov A., Shamayeva O., Mikhailov G.* Improve the Reliability of the Organization's Resources to Support Remote Education Services // VI International Conference on Information Technologies in Engineering Education (Inforino), Moscow, Russian Federation, 2022, pp. 1–4.
https://doi.org/10.1109/Inforino53888.2022.9783000

13. DallasLock Linux Homepage.
URL: https://dallaslock.ru/products/szi-nsd-dallas-lock-linux/.

14. SecretNet Studio Homepage.
URL: https://www.securitycode.ru/products/szi_secret_net/?tab=system.

15. Order of the Ministry of Communications of Russia dated April 1, 2015, No. 96, On Approval of the Software Import Substitution Plan.
URL: https://digital.gov.ru/documents/prikaz-minkomsvyazi-rossii-ob-utverzhdenii-plana-importozameshheniya-programmnogo-obespecheniya.

СВЕДЕНИЯ ОБ АВТОРАХ

МИХАЙЛОВ Гурий Михайлович – кандидат физ-мат. наук, гл. специалист отдела 11 Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук.

Gury Mikhailovich Mikhaylov – Candidate of Physical and MathematicalSciences, Chief Specialist of Department 11 at the A. A. Dorodnitsyn Computing Center, Federal Research Center "Informatics and Management" of the Russian Academy of Sciences.

email: GMihaylov@frccsc.ru

ORCID: 0000-0002-4535-7180

ТУЧКОВА Наталия Павловна – кандидат физ-мат. наук, старший научный сотрудник отдела 11 Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, окончила ВМиК МГУ им. М. В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – Senior Researcher at the A. A. Dorodnitsyn Computing Center, Federal Research Center "Computer Science and Control" of the RAS, PhD in physics with a math degree, graduated from CS Faculty of Lomonosov MSU. The expert in the field of algorithmic languages and information technologies.

email: NTuchkova@frccsc.ru

ORCID: 0000-0001-5357-9640

ЧЕРНЕЦОВ Андрей Михайлович– кандидат технических наук, доцент; доцент кафедры Прикладной математики и искусственного интеллекта Национального исследовательского университета «МЭИ».

Andrey Mikhailovich CHERNETSOV – Candidate of Technical Sciences, Associate Professor; Associate Professor at the Department of Applied Mathematics and Artificial Intelligence, National Research University "MPEI".

email: chernetsovam@mpei.ru

ORCID: 0000-0001-7655-2395

Материал поступил в редакцию 14 января 2026 года

УДК 01+002.53

ПЕРЕЧЕНЬ ЖУРНАЛОВ ВАК И ДРУГИЕ РОССИЙСКИЕ ИНДЕКСЫ

Т. А. Полилова[0000-0003-4628-3205]

Институт прикладной математики им. М.В. Келдыша РАН, г. Москва, Россия
polilova@keldysh.ru

Аннотация

В соответствии с требованием Высшей аттестационной комиссии (ВАК) метаданные выпусков журналов из Перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук (Перечень ВАК) уже более 20 лет регулярно размещаются в Российском индексе научного цитирования (РИНЦ) в библиографической базе eLibrary.ru. С марта 2023 г. редакции журналов из Перечня ВАК по рекомендации ВАК начали размещать сведения о выпусках журналов за 2022 г. в базу данных «Российские научные журналы (РНЖ)», созданной Российским научно-исследовательским институтом экономики, политики и права в научно-технической сфере. В апреле 2025 г. приказом Минобрнауки РФ было добавлено новое требование — для журналов из Перечня ВАК наряду с регистрацией в РИНЦ eLibrary.ru требуется регистрация в информационной системе (ИС) «Метафора», разработанной Российским центром научной информации (РЦНИ). Журналам из Перечня ВАК рекомендовано регулярно передавать в ИС «Метафора» метаданные вышедших выпусков журналов через специально организованные интерфейсы. Какую роль выполняют базы РНЖ и ИС «Метафора» в инфраструктуре научных публикаций?

РЦНИ, помимо развития ИС «Метафора», по поручению Правительства РФ выполняет функцию оператора «Белого списка» (БС) научных изданий. «Белый список» в 2023 г. сформировала Межведомственная рабочая группа (МРГ) Минобрнауки РФ. «Белый список» предлагается использовать для мониторинга и оценки публикационной активности российских ученых. В БС изначально было включено около 29 тыс. англоязычных международных журналов и около 1000 русскоязычных журналов из базы Russian Science Citation Index (RSCI). В сентябре

2025 г. русскоязычная часть БС значительно расширилась за счет включения в него журналов из Перечня ВАК. Хотелось бы получить от идеологов БС развернутую информацию о том, как будут корреспондироваться уровни журналов «Белого списка» (У1, У2, У3, У4) и категории журналов Перечня ВАК (К1, К2, К3)?

***Ключевые слова**: перечень ВАК, РИНЦ, eLibrary.ru, база российских журналов РНЖ, информационная система «Метафора», «Белый список».*

ВВЕДЕНИЕ

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, (далее — Перечень) был учрежден Высшей аттестационной комиссией (ВАК) при Министерстве науки и высшего образования (Минобрнауки) РФ. В действующем Положении «О порядке присуждения ученых степеней» ВАК, утвержденном постановлением Правительства Российской Федерации от 24.09.2013 г. № 842, содержится пункт, обязывающий соискателя ученой степени кандидата наук или доктора наук публиковать статьи с результатами диссертационных исследований в журналах из Перечня ВАК.

Кратко представим историю формирования Перечня [1, 2]. Первый вариант Перечня, сформированный в основном силами экспертных советов ВАК, появился в начале 2000-х годов. Долгое время Перечень ВАК существовал в виде текстового документа, который можно назвать «бумажным» индексом и прародителем будущих информационных систем. Актуальный Перечень размещался на сайте ВАК в разделе, содержащем справочные материалы. Кроме того, можно было обратиться к поисковому сервису Яндекс, который выдавал адреса десятков сайтов с разнообразными версиями Перечня ВАК. Однако понять, на каком сайте размещена актуальная версия Перечня, было довольно сложно.

Перечень ВАК пополнялся новыми журналами, но делалось это несистематически, время от времени.

В 2015 г. ВАК изменила правила формирования Перечня. Редакция журнала могла инициативно подать заявку на включение журнала в Перечень,

включая в эту заявку довольно представительный объем сведений о журнале и выполняя определенные требования. Перечислим основные, наиболее существенные из них.

Прежде всего журнал должен быть зарегистрирован в Роскомнадзоре как средство массовой информации. Научное издание должно иметь международный стандартный номер сериального издания (ISSN). В состав редакционной коллегии журнала должны входить ведущие специалисты (преимущественно кандидаты и доктора наук), внесшие значительный вклад в развитие соответствующей области знаний.

Редакция журнала должна осуществлять рецензирование всех материалов, поступающих в редакцию. Рецензирование осуществляется как членами редколлегии, так и внешними экспертами.

Журнал должен быть зарегистрирован в Российском индексе научного цитирования (РИНЦ) в eLibrary.ru и регулярно поставлять туда метаданные статей (возможно, полные тексты) выходящих выпусков.

Кроме того, журнал должен иметь сайт для открытого опубликования сведений о журнале, редакционной коллегии, издательской политике, тематике журнала, а также правил оформления статей, порядка рецензирования статей и т. д.

В 2018 г. ВАК в своем письме рекомендовала редакциям журналов из Перечня указывать на сайте (в анкете) журнала специальности номенклатуры ВАК, к которым относятся статьи, публикуемые в журнале.

В 2021 г. были изданы нормативные документы ВАК о вступлении в действие новой номенклатуры специальностей. В соответствии с рекомендациями Президиума ВАК в июне 2022 г. редакции журналов, входящих в Перечень, подали в ВАК сведения о привязке тематики журналов к специальностям новой номенклатуры.

В феврале-марте 2023 г. ВАК запустила кампанию по сбору подробной информации о выпусках журналов из Перечня за 2022 г., распространив официальное письмо от 10 февраля 2023 г. № 4/3-разн «О заполнении данных в личных кабинетах журналов Перечня ВАК» [3]. Сбор информации проводился в специальной базе «Российские научные журналы» (РНЖ), созданной

Российским научно-исследовательским институтом экономики, политики и права в научно-технической сфере (РИЭПП) [4].

БАЗА «РОССИЙСКИЕ НАУЧНЫЕ ЖУРНАЛЫ»

Цель создания базы научных журналов РНЖ декларируется на сайте разработчика https://rng.riep.ru/. База РНЖ позволяет:

— обеспечивать учет и мониторинг научных журналов;

— осуществлять работу с редакторами по улучшению качественных параметров научных журналов;

— выполнять детальный анализ публикационной активности;

— выявлять точки концентрации результатов научных исследований для определения наиболее перспективных направлений.

Редакции журналов через личный кабинет должны были внести в РНЖ сведения о вышедших статьях и их авторах. На рис. 1 представлена начальная страница сайта проекта РНЖ, содержащая форму для входа в личный кабинет ответственного лица, выполняющего передачу сведений о выпусках журнала в базу РНЖ.

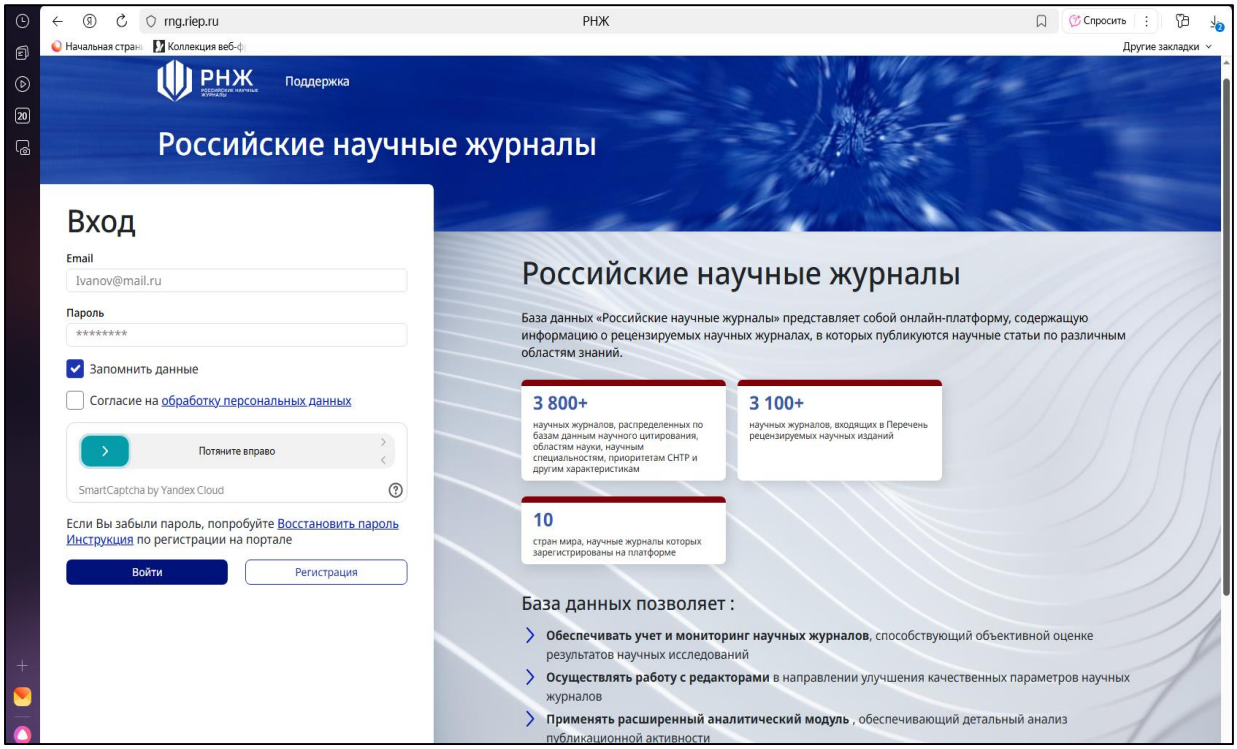


Рис. 1. Сайт проекта «Российские научные журналы» https://rng.riep.ru/. Вход в личный кабинет.

Особенность структуры данных базы РНЖ состоит в том, что *каждая журнальная статья*, опубликованная в журнале в 2022 г., могла быть привязанной к специальностям новой номенклатуры ВАК. Кроме того, редакции журналов получили возможность передать более подробную информацию об авторах статей, в частности, можно было ввести ученую степень и ученое звание авторов.

Таким образом, появилась специализированная информационная система РНЖ, содержащая метаданные журналов, входящих в Перечень ВАК. В базу РНЖ были импортированы из eLibrary.ru библиометрические показатели этих журналов.

База РНЖ была использована экспертными группами ВАК для проведения экспертной оценки журналов, входящих в Перечень, по методике, утвержденной Письмом ВАК РФ от 6 декабря 2022 г. № 02-1198 «О Перечне рецензируемых научных изданий» [5]. В этом документе ВАК приведен список всех научных журналов, входящих в Перечень, с указанием категорий К1, К2, К3. Группа К1 с наилучшими показателями составляет 25% всех журналов, группа К2

— 50%, группа К3 — 25% журналов. Перечень, в соответствии с приведенным в письме ВАК списком, содержал 2587 наименований журналов.

Кроме того, в этом письме отмечается, что журналы, входящие в международные базы данных Web of Science, Scopus, PubMed, MathSciNet, zbMATH, Chemical Abstracts, Springer и GeoRef, а также журналы из базы Russian Science Citation Index (RSCI), приравниваются к изданиям категории К1. Уточним, что база RSCI включает ведущие российские научные журналы.

Кроме того, в указанном письме сообщается, что методика экспертной оценки журналов включает в себя две составляющие: количественную, основанную на общепринятых наукометрических показателях, и экспертную. Эксперты, анализируя качественные показатели журнала, имели возможность оценить такие характеристики журнала, как:

— качество и научный уровень статей;

— соответствие содержания журнала заявленным специальностям номенклатуры ВАК;

— авторитетность организации-учредителя;

— состав редакционной коллегии;

— эффективность рецензирования;

— профессиональный уровень (респектабельность) авторов.

База РНЖ журналов ВАК имела два интерфейса:

— интерфейс представителя журнала для ввода сведений о журнале и опубликованных статьях;

— интерфейс для члена экспертной группы ВАК, проводящего оценку журналов.

В работах [1, 2] было высказано предположение, что в ближайшее время в РНЖ может появиться интерфейс, ориентированный на соискателей ученых степеней и членов диссертационных советов, принимающих к защите диссертации. Такой интерфейс позволил бы и соискателям, и диссертационным советам с помощью запросов к базе РНЖ контролировать выполнение требования о публикациях в журналах Перечня ВАК.

Кроме того, в [1, 2] было сформулировано пожелание, чтобы РНЖ и РИНЦ eLibrary.ru наладили конструктивное взаимодействие, гармонизировали собираемые метаданные и регулярно обменивались библиометрическими

данными научных журналов. Процесс передачи данных в эти смежные библиографические базы должен проходить автоматически, реализуя разумный принцип «одного окна». В этом случае, в частности, редакциям журналов в 2023 г. не пришлось бы фактически дублировать ввод метаданных в базу РНЖ: все данные, необходимые для РНЖ, могли бы импортироваться из РИНЦ eLibrary.ru.

К сожалению, высказанные пожелания не были услышаны, сотрудничество так и не было установлено.

В мае 2025 г. от РНЖ в адрес издателей журналов пришло электронное письмо об актуализации сведений о журналах и загрузке метаданных выпусков, вышедших в 2023–2025 гг. Данная рассылка по электронной почте не была поддержана официальным документом, принятым руководством ВАК. Тем самым возникла «стратегическая неопределенность», дающая основание сомневаться в целесообразности дальнейшей загрузки в РНЖ метаданных выпусков журналов.

ПРИКАЗ МИНОБРНАУКИ ОБ ИЗМЕНЕНИИ ТРЕБОВАНИЙ К ЖУРНАЛАМ

В апреле 2025 г. произошло заметное событие, касающееся, в том числе, редакций научных журналов из Перечня ВАК. Минобрнауки РФ приказом № 337 от 11 апреля 2025 г. [6–8] внесло изменения в «Правила формирования перечня рецензируемых научных изданий, в которых должны публиковаться результаты диссертационных исследований» (рис. 2).

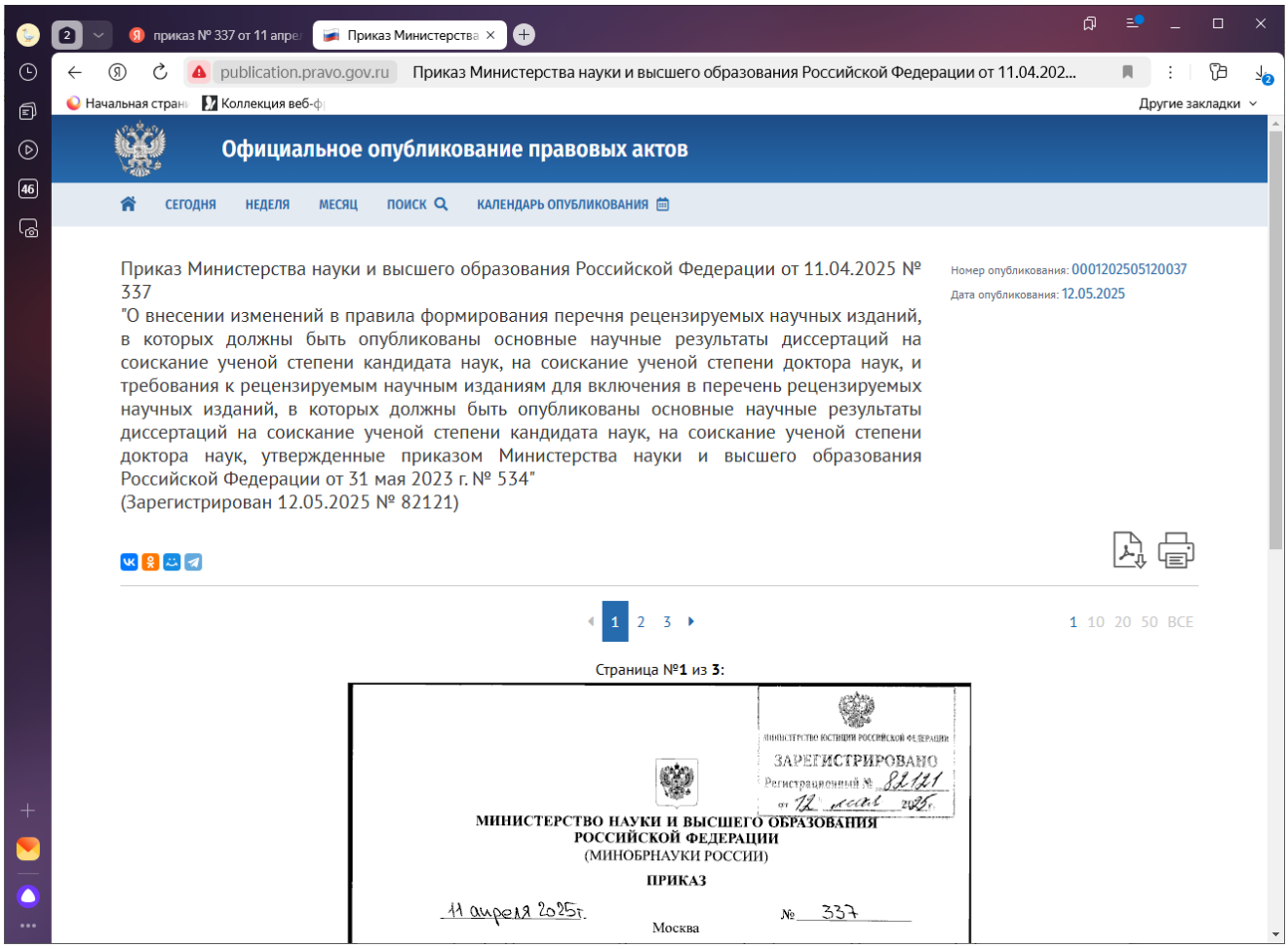


Рис. 2. Приказ Минобрнауки РФ № 337 от 11 апреля 2025 г. на официальном сайте опубликования правовых актов http://publication.pravo.gov.ru/document/0001202505120037.

Данный приказ утверждает новую редакцию п. 11 из списка требований к журналам, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней. Этот пункт звучит теперь так:

«Издание должно быть зарегистрировано в Российском индексе научного цитирования (РИНЦ), Российском центре научной информации (РЦНИ) и/или в другой системе научного цитирования, определяемой международными договорами Российской Федерации и/или рекомендациями Комиссии, и предоставлять в данные системы научного цитирования информацию об опубликованных научных статьях в трехмесячный срок со дня выпуска соответствующего номера издания».

Обсудим эту формулировку с точки зрения формальных правил русского языка. Конструкция «и/или» подразумевает, что возможны несколько вариантов трактовки текста.

Вариант 1. Случай применения предлога «и» предписывает регистрировать издания в РИНЦ, РЦНИ, а также в другой системе научного цитирования, определяемой международными договорами Российской Федерации и рекомендациями ВАК. Регистрация в двух перечисленных базах (РИНЦ, РЦНИ) и подразумеваемой международной базе является, следуя рассматриваемой логике приказа, обязательной.

Жаль, что в приказе не указаны конкретные системы научного цитирования, которые определены «международными договорами Российской Федерации и/или рекомендациями Комиссии». Попробуем обратиться к искусственному интеллекту (ИИ) Яндекса с вопросом, о каких системах научного цитирования может идти речь. Такой способ нахождения нужной информации стал достаточно часто использоваться в деловых отношениях. Уже есть опыт успешного применения ИИ в судебных делах, в которых участникам приходится опираться на сложную систему нормативных документов.

Зададим в Яндексе вопрос: «Что такое система научного цитирования, определяемая международными договорами Российской Федерации», включим поиск с Алисой. В результате анализа доступных материалов и неких рассуждений Алиса выдала весьма длинный текст, в котором внимание можно сосредоточить на следующем фрагменте текста:

«Таким образом, система научного цитирования в РФ сочетает национальные (РИНЦ, RSCI) и международные (Web of Science, Scopus) инструменты, регулируемые как внутренними приказами, так и международными соглашениями. Это позволяет унифицировать критерии оценки научной значимости публикаций в глобальном контексте».

Полученный ответ недостаточно конструктивен и не снимает всех возникающих вопросов. Но, безусловно, нацеливает читателя приказа № 337 от 11 апреля 2025 г. на обязательную регистрацию в базах Web of Science и Scopus, а также, возможно, в других международных базах.

Вариант 2. Продолжим анализировать формулировку пункта 11 требований, вводимых приказом № 337 от 11 апреля 2025 г. Приведем

следующую трактовку формулировки приказа, когда в варианте «и/или» используем предлог «или». В этом случае можно трактовать приказ так, что регистрироваться можно: или в РИНЦ, или в РЦНИ, или в любой международной базе, определяемой международными договорами с РФ. Следовательно, если издание зарегистрировано в РИНЦ, то в свете новой редакции правил формирования Перечня никаких других регистраций в дополнительных базах не требуется.

Вариант 3. Следующий вариант трактовки приказа при использовании предлога «или» таков: регистрироваться нужно в РИНЦ, а также либо в РЦНИ, либо в международной базе.

Однако вслед за этим приказом выходит Постановление ВАК № 13/4-разн от 16 мая 2025 г., в котором явно рекомендовано всем журналам, включенным в Перечень ВАК, до 1 ноября 2025 г. зарегистрироваться в базе РЦНИ. Вспомним, что регистрация в РИНЦ eLibrary.ru уже давно является обязательной, и теперь обязательной становится регистрация в двух базах: РИНЦ eLibrary.ru и РЦНИ. Таким образом, своим постановлением ВАК исключает вариант 2, рассмотренный выше. Обратим внимание, что упомянутое Постановление ВАК не предписывает регистрироваться в международной базе, что, вообще говоря, может искажать суть Приказа № 337 от 11 апреля 2025 г., поскольку явно не исключен рассмотренный выше вариант 1. Однако заметно прибавляется уверенности в том, что случай, когда журнал регистрируется и в РИНЦ, и в РЦНИ, но не регистрируется в международной базе, вполне удовлетворяет требованиям ВАК.

ЖУРНАЛЫ «БЕЛОГО СПИСКА» НА САЙТЕ РЦНИ

Что представляет собой организация «Российский центр научной информации» [9] (РЦНИ — бывший РФФИ)? На сайте РЦНИ размещена информация о том, что по поручению Правительства РФ РЦНИ развивает «Национальную платформу периодических научных изданий», выполняет функцию оператора национальной подписки и оператора «Белого списка» (далее – БС) научных изданий.

Напомним, что такое БС журналов. «Белый список» сформировала Межведомственная рабочая группа Минобрнауки (МРГ) с участием создаваемых

экспертных рабочих групп. В Межведомственную рабочую группу входят около 30 человек, в том числе представители:

— общественно-экспертного совета по национальному проекту «Наука и университеты»;

— Российской академии наук;

— РЦНИ.

«Белый список» предполагается использовать для мониторинга и оценки публикационной активности российских ученых. В БС изначально было около 29 тыс. англоязычных международных журналов и около 1 тыс. русскоязычных журналов из базы Russian Science Citation Index (RSCI). Напомним, что созданная база RSCI включает российские научные журналы, отобранные в ходе реализации проекта компании Clarivate Analytics и электронной библиотеки elibrary.ru. На старте в 2015 г. основной целью проекта RSCI было размещение журналов из базы RSCI на платформе Web of Science [10, 11].

В 2023 г. в методике категорирования (ранжирования) журналов, утвержденной на уровне Минобрнауки, перечислены следующие основные требования к журналам БС [12]:

— в журнале должны публиковаться преимущественно научные статьи с предварительным рецензированием материалов, поступающих в редакцию;

— журнал должен иметь веб-сайт с доступными метаданными публикаций, включая списки процитированных источников;

— журнал должен присваивать публикациям цифровой идентификатор doi или российский аналог, который должен обеспечивать прямой доступ к метаданным публикаций;

— должна обеспечиваться возможность свободного бесплатного использования метаданных с открытой лицензией;

— статьи из журнала должны иметь цитирования в публикациях других журналов БС;

— главный редактор и большинство членов редколлегии журнала должны иметь публикации в ведущих журналах БС (помимо рассматриваемого журнала) за последние годы.

«Белый список», вообще говоря, открыт для изменений. Журналы, не удовлетворяющие предъявляемым требованиям, могут быть исключены

(постоянно или временно) из БС. Как попасть в БС российскому журналу? На сайте РЦНИ в разделе «Часто задаваемые вопросы» [13] можно ознакомиться со следующей информацией:

«Текущая версия БС уже утверждена, в настоящее время Межведомственной рабочей группой по формированию и актуализации БС научных журналов разрабатывается порядок его актуализации и отбора новых журналов для включения в список. Прием заявок на включение журналов в БС будет открыт после утверждения вышеуказанных документов».

РАНЖИРОВАНИЕ ЖУРНАЛОВ БС

Журналы БС ранжируются по четырем уровням У1, У2, У3, У4 согласно методике, принятой Рабочей группой. Сведения о журналах в БС регулярно актуализируются и дополняются новыми показателями метрик, дополнительной аналитической и другой информацией, агрегируемой из более чем 30 международных баз данных и сервисов [14].

Высказывается мнение, что уровни БС примерно соответствуют квартилям Scopus. Уровень У1 БС соответствует 1–2-му квартилю Web of Science [15]. Возможно, в отношении англоязычных журналов, издающихся за рубежом, такое соотнесение оправдано. Вопрос может возникать в отношении русскоязычных журналов из коллекции RSCI. Известно, что в международных базах индексируются в основном англоязычные версии российских журналов. Всегда ли российские журналы, в частности, ведущие в своих областях, оправдано *не попадают* в высокие квартили в западных индексах, где, вероятно, не учитываются цитирования из русскоязычных публикаций? Получают ли ведущие российские журналы из коллекции RSCI адекватные показатели уровня в БС?

Уровень журнала в БС указан в карточке журнала. На сайте РЦНИ имеется форма для поиска журнала, в которой можно создать поисковый запрос, указав название интересующего журнала. Можно отобрать группу журналов, удовлетворяющую заданным условиям поиска. На рис. 3 показаны результаты поиска журналов БС, входящих в базы Scopus и DBLP, поскольку в поисковом запросе эти две базы указаны (отмечены галочкой) в условиях поиска.

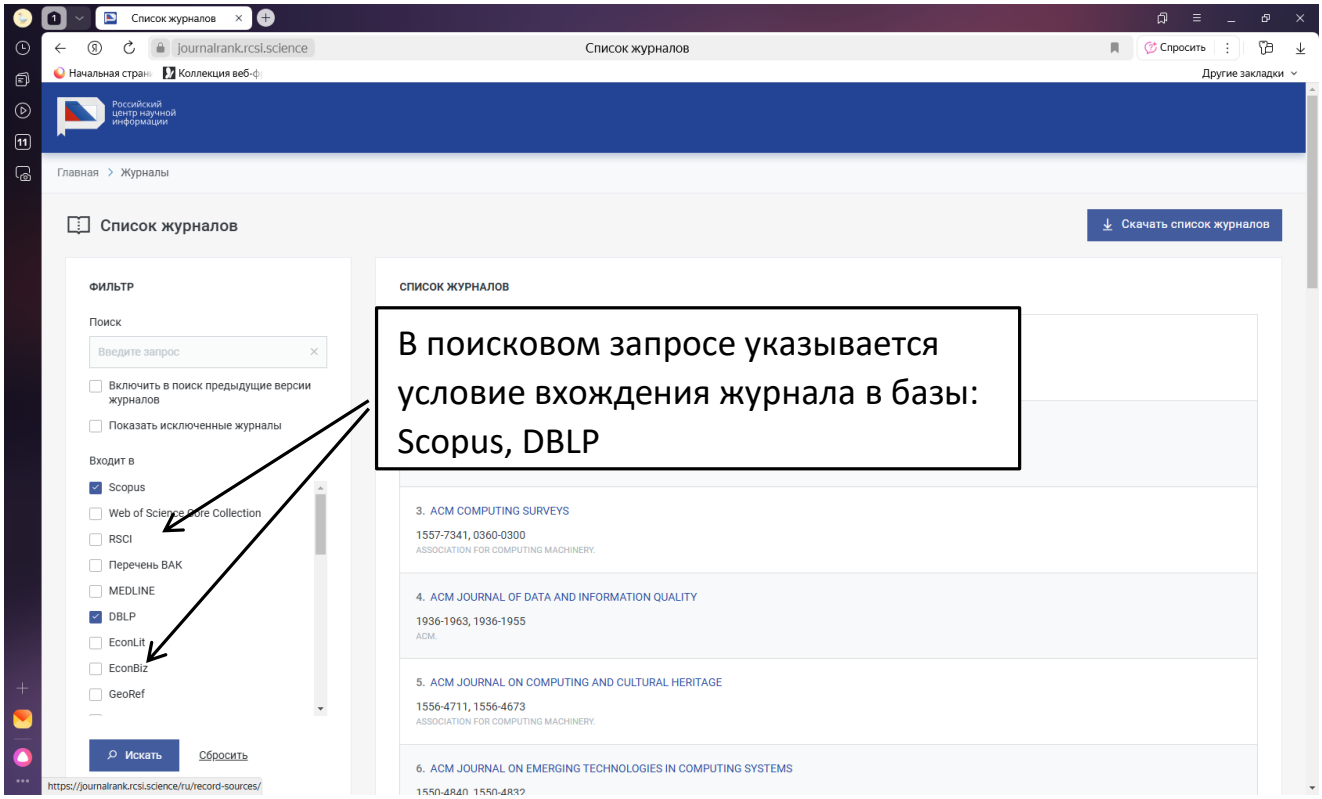


Рис. 3. «Белый список». Форма с запросом для выбора группы журналов.

Какая информация выдается пользователю по выбранному журналу? Рассмотрим, например, данные журнала ACM COMPUTING SURVEYS, издающегося в США на английском языке. Этот журнал отнесен к уровню У1 в БС.

В профиле журнала указаны следующие данные:

— название журнала;

— ISSN (ссылка активная, при переходе открывается карточка журнала на портале ISSN);

— базы, в которых индексируется журнал;

— уровень журнала согласно методике БС;

— другая основная информация журнала.

На рис. 4 в разделе «Анализ» показаны квартили журнала ACM COMPUTING SURVEYS в Scopus за 2015–2024 гг. По двум тематическим категориям классификатора ASLG (All Science Journal Classification) в двух рейтингах *CiteScore* и *SJR* журнал входит в 1-й квартиль (этот квартиль выделен зеленым цветом).

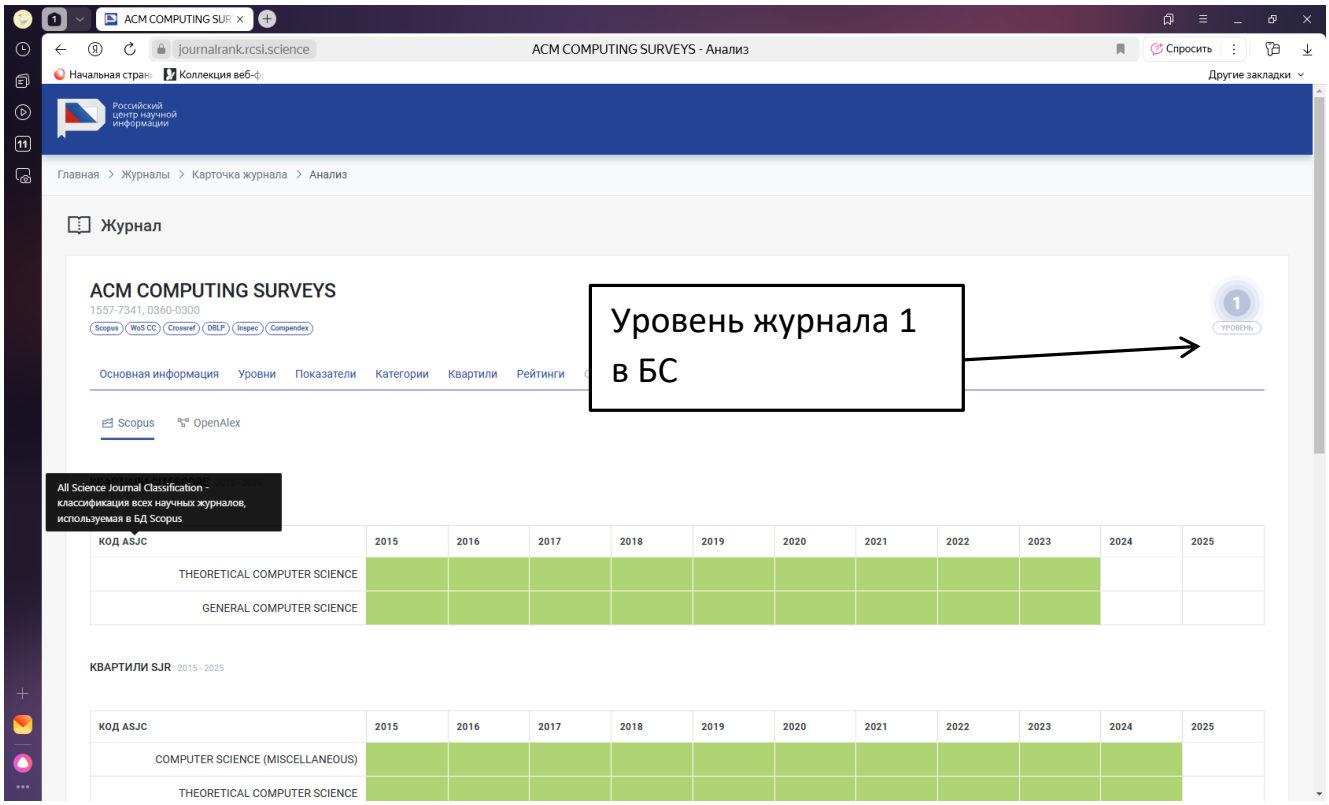


Рис. 4. «Белый список». В разделе «Анализ» показаны квартили журнала ACM COMPUTING SURVEYS в Scopus за 2015-2024 гг. в рейтингах *CiteScore* и *SJR*.

Посмотрим параметры еще одного журнала — ACM COMMUNICATIONS IN COMPUTER ALGEBRA. Этот журнал, издающийся в США на английском языке, имеет уровень У4 в БС. На вкладке «Анализ» можно увидеть показатели журнала в базе Scopus за 2015–2023 гг. (рис. 5) по двум тематическим категориям классификатора ASLG в рейтингах *CiteScore* и *SJR*. Красным цветом обозначен квартиль Q4, оранжевым цветом — Q3.

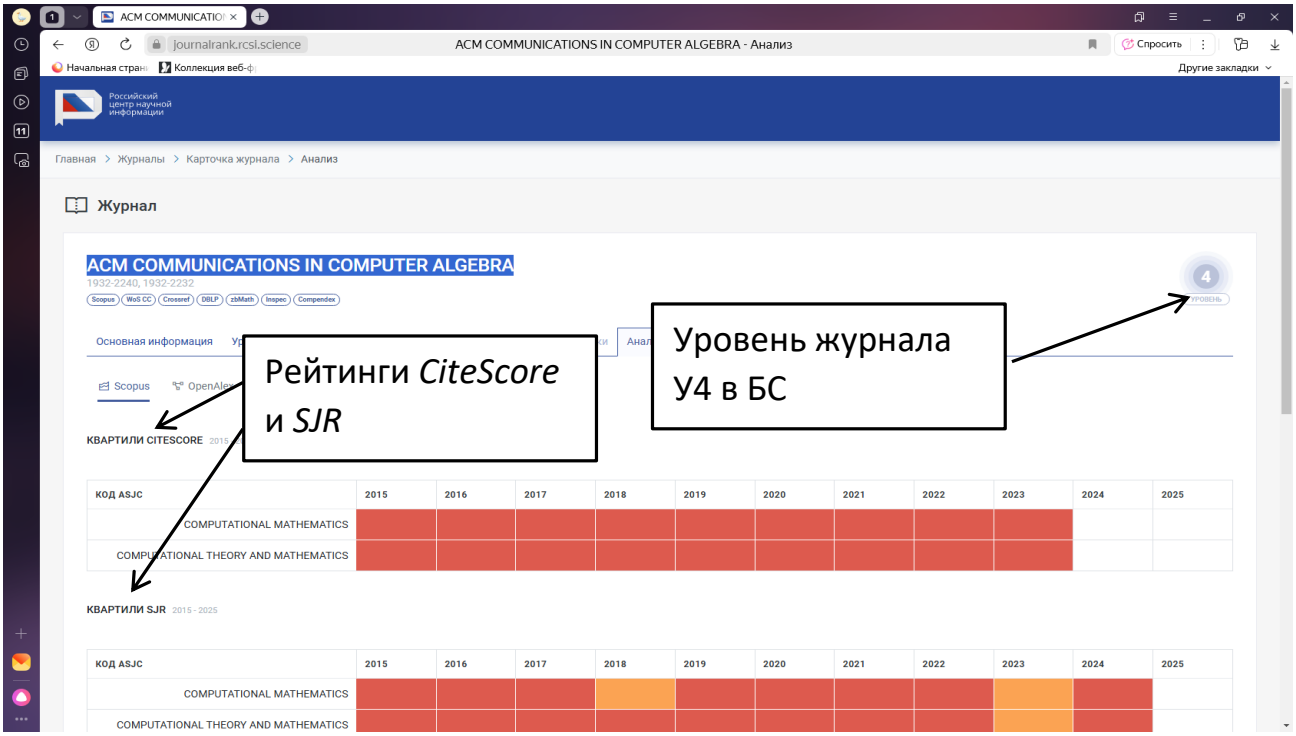


Рис. 5. «Белый список». В разделе «Анализ» показаны квартили журнала ACM COMMUNICATIONS IN COMPUTER ALGEBRA в Scopus за 2015–2024 гг. в рейтингах *CiteScore* и *SJR*.

Найдем в БС русскоязычный журнал «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ», который входит в базу RSCI. Карточка журнала представлена на рис. 6. На вкладке с основной информацией указаны:

— название журнала;

— ISSN;

— язык и страна выпуска;

— базы, в которых индексируется журнал.

Поскольку русскоязычный журнал «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ» в русской версии не входит в Scopus, нет аналитики из этой базы в БС. На вкладке «Анализ» нет информации о показателях рейтингов журнала. Но журнал входит в РИНЦ eLibrary.ru, и на странице «Анализ» могли бы появиться данные о цитированиях и месте этого журнала «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ» в рейтинге Science Index РИНЦ eLibrary.ru. К сожалению, этих сведений пользователь базы БС не видит.

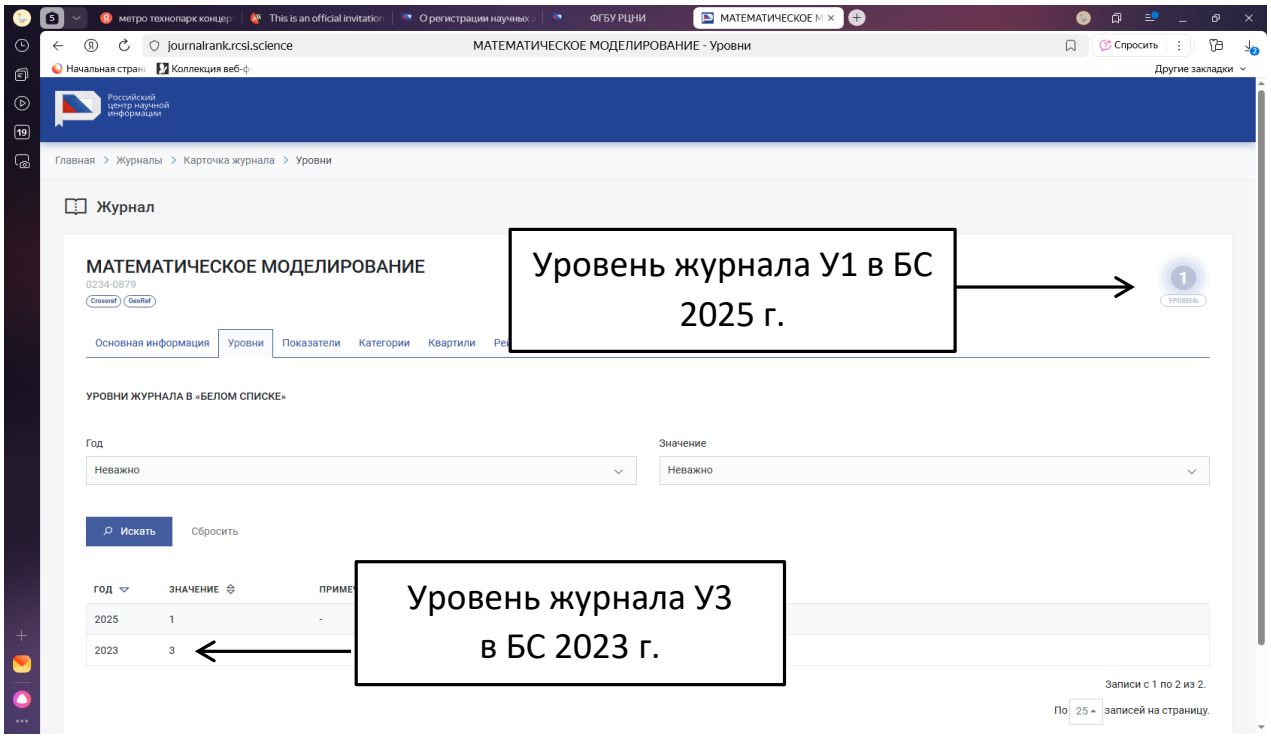


Рис. 6. «Белый список». Страница с основной информацией русскоязычного журнала «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ».

Журнал «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ» в БС версии 2025 г. имеет уровень У1 (иными словами — квартиль 1), который указан в правом верхнем углу. Здесь требуются пояснения. Обратим внимание, что в левом нижнем углу присутствует следующее указание:

2025 год — 1;

2023 год — 3.

Это означает, что в 2025 г. журнал имеет уровень У1. В 2023 г. журнал был отнесен к уровню У3, что является довольно скромным показателем для этого журнала. Если посмотреть показатели журнала «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ» в РИНЦ eLibrary.ru, то увидим другую картину (рис. 7).

Журнал входил в 6-й процентиль в общем рейтинге Science Index РИНЦ eLibrary.ru в 2023 г., и это является весьма хорошим показателем. Если бы в РИНЦ eLibrary.ru использовалось такое понятие, как «квартиль», то журнал попал бы в 1-й квартиль РИНЦ eLibrary.ru. Напомним, что общий рейтинг Science Index РИНЦ нормализован по тематике журналов, то есть при расчете рейтинга использовались средневзвешенные коэффициенты для журналов разной тематики.

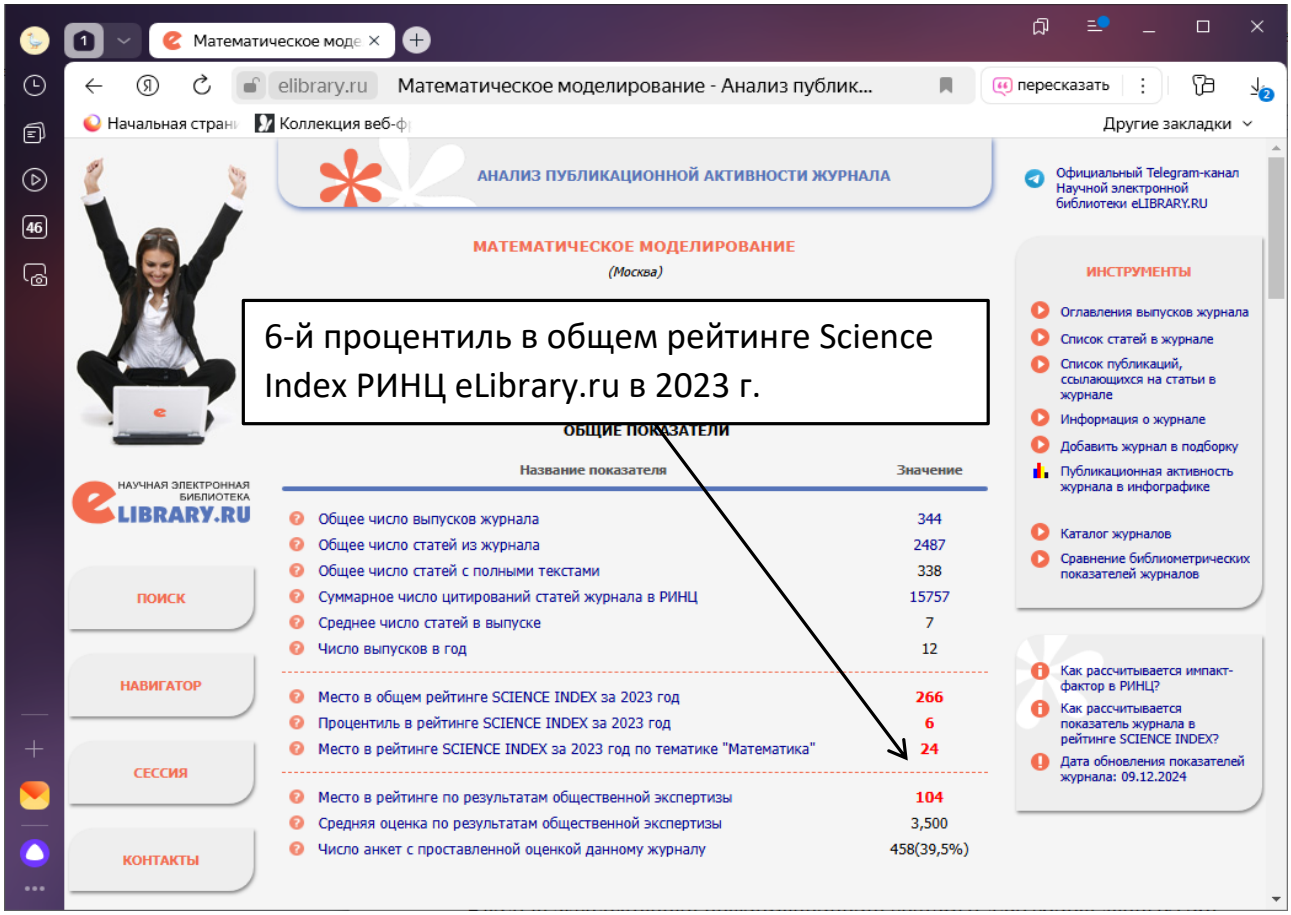


Рис. 7. Журнал «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ» в РИНЦ eLibrary.ru. Раздел «Анализ публикационной активности журнала» (по состоянию на май 2025 г.).

Почему в 2025 г. уровень журнала изменился и стал У1? Неужели в БС изменились требования к журналам? Или у журнала «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ» в 2025 г. резко улучшились показатели цитирования, на которых основаны процедуры привязки журналов к уровням БС? Причина, видимо, в другом.

В середине 2025 г. в БС административным решением были включены около 2.5 тысяч журналов из Перечня ВАК. Вспомним, что журнал «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ» входит в базу RSCI. Журналу из RSCI автоматически присваивается высшая категория Перечня ВАК — К1. Если журнал из К1 получает в БС версии 2023 г. уровень У3, то какие уровни должны получить журналы категории К2 и К3 Перечня ВАК?

Вероятнее всего, идеологи БС решили не создавать неприятную коллизию и присвоили журналам базы RSCI уровень У1. Скорее всего, такая процедура

была применена не ко всем журналам категории К1 Перечня ВАК. Тем не менее у многих специалистов, читающих документы с описанием критериев отбора и разнообразных формул, призванных придать принципам ранжирования журналов БС «математическую точность», могут возникать неудобные вопросы. Всегда ли справедливы показатели ранжирования российских журналов в БС?

В табл. 1 представлены показатели журналов, занимающих первые десять позиций в тематическом разделе «Математика» в рейтинге Science Index в РИНЦ eLibrary.ru (по состоянию на май 2025 г.), и показатели этих журналов в БС в версиях 2023 и 2025 гг.

Табл. 1. Показатели рейтингов математических журналов, занимающих первые десять позиций в рейтинге Science Index РИНЦ eLibrary.ru в разделе «Математика» (по состоянию на май 2025 г.)

№	Название, характеристики	Рейтинг SI РИНЦ (май 2025 г.)	Уровень в «Белом списке»	
			2023 г.	2025 г.
1	ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ WOS (перев.), Scopus (перев.), RSCI, ВАК, Science Index (1%)	15.802	У2	У1
2	МАТЕМАТИЧЕСКИЙ СБОРНИК WOS (перев.), Scopus (перев.), RSCI, ВАК, Science Index (1%)	15.184	У1	У1
3	ИЗВЕСТИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК. СЕРИЯ МАТЕМАТИЧЕСКАЯ WOS (перев.), Scopus (перев.), RSCI, ВАК, Science Index (1%)	15.168	У1	У1
4	ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ WOS (перев.), Scopus (перев.), RSCI, ВАК, Science Index (1%)	14.840	У1	У1
5	УСПЕХИ МАТЕМАТИЧЕСКИХ НАУК WOS (перев.), Scopus (перев.), RSCI, ВАК, Science Index (1%)	13.635	У1	У1
6	МАТЕМАТИЧЕСКИЕ ЗАМЕТКИ WOS (перев.), Scopus (перев.), RSCI, ВАК, Science Index (1%)	13.537	У2	У1

7	СИБИРСКИЙ МАТЕМАТИЧЕСКИЙ ЖУРНАЛ WOS (перев.), Scopus (перев.), RSCI, ВАК, Science Index (1%)	13.262	У2	У1
8	LOBACHEVSKII JOURNAL OF MATHEMATICS WOS, Scopus, RSCI, ВАК, Science Index (2%)	12.960	У2	—
9	ПРИКЛАДНАЯ МАТЕМАТИКА И МЕХАНИКА RSCI, ВАК, Science Index (2%)	12.830	У3	У1
10	REGULAR AND CHAOTIC DYNAMICS WOS, Scopus, RSCI, ВАК, Science Index (2%)	11.984	У1	У1

В первом столбце таблице 1 указано место журнала в рейтинге Science Index РИНЦ eLibrary.ru в разделе «Математика». Во втором столбце показано название журнала. Кроме того, наряду с названием указаны базы, в которых проиндексирован журнал, а также процентиль в общем рейтинге Science Index РИНЦ eLibrary.ru. В третьем столбе таблицы представлены *значения* показателей рейтинга Science Index РИНЦ eLibrary.ru, в четвертом и пятом столбцах — уровень (квартиль) каждого журнала в БС в версиях 2023 г. и 2025 гг.

Как можно заметить, все 10 журналов входят в Перечень ВАК. Поскольку эти журналы одновременно входят в RSCI, они все отнесены к категории К1. Но в БС половина из этих журналов в 2023 г. была отнесена к уровню У1, в половина — к уровню У2 и У3.

Рассмотрим дополнительно параметры двух соседних журналов из табл. 1, отнесенных к разным уровням БС в 2023 г.

У русскоязычного «ЖУРНАЛА ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ», который занимает 1-е место в рейтинге Science Index РИНЦ eLibrary.ru в разделе «Математика» и входит в 1-й процентиль общего рейтинга, уровень журнала в БС 2023 г. — У2.

Русскоязычный журнал «МАТЕМАТИЧЕСКИЙ СБОРНИК» (2-е место в рейтинге Science Index РИНЦ eLibrary.ru в разделе «Математика» и 1-й процентиль общего рейтинга) имеет уровень в БС 2023 г. — У1.

Насколько близко расположились два журнала, занимающие 1-е место и 2-е место в рейтинге Science Index РИНЦ eLibrary.ru (в разделе «Математика»), в рейтинге БС 2023 г.? Хотя оба показателя в рейтингах РИНЦ и БС базируются на учете цитирований, ответить на вопрос, почему эти журналы имеют разные уровни в БС 2023 г. (У1 и У2), весьма затруднительно.

Вновь вернемся к вопросу адекватности отнесения журналов к уровням БС в 2023 г. Обратимся к данным англоязычного журнала “REGULAR AND CHAOTIC DYNAMICS”, разместившегося в рейтинге Science Index РИНЦ eLibrary.ru в разделе «Математика» на 10-м месте. В общем рейтинге Science Index РИНЦ eLibrary.ru этот журнал входит во 2-й процентиль. Значение показателя рейтинга этого журнала значительно ниже, чем у «Журнала вычислительной математики и математической физики», занимающего 1-е место в разделе «Математика». Отметим, что показатель рейтинга Science Index РИНЦ основан на подсчете цитирований. Однако в БС 2023 г. журнал “REGULAR AND CHAOTIC DYNAMICS” имеет уровень У1, в то время как «ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ» отнесен к уровню У2.

Журнал “REGULAR AND CHAOTIC DYNAMICS” входит в Scopus, поэтому в БС можно увидеть весьма богатую информацию о журнале.

На вкладке «Анализ» (рис. 8) показано, что журнал входит в разные квартили в Scopus в рейтинге *CiteScore* в зависимости от тематики (по классификатору ASJC). В 2023 г. по тематике “MATHEMATICS (MISCELLANEOUS)” журнал имел 1-й квартиль (выделен зеленым цветом). По двум тематикам APPLIED MATHEMATICS, MATHEMATICAL PHYSICS журнал отнесен к 2-му квартилю (выделен желтым цветом). По трем тематикам MODELING AND SIMULATION, STATISTICAL AND NONLINEAR PHYSICS, MECHANICAL ENGINEERING журнал входит в 3-й квартал (выделен оранжевым цветом).

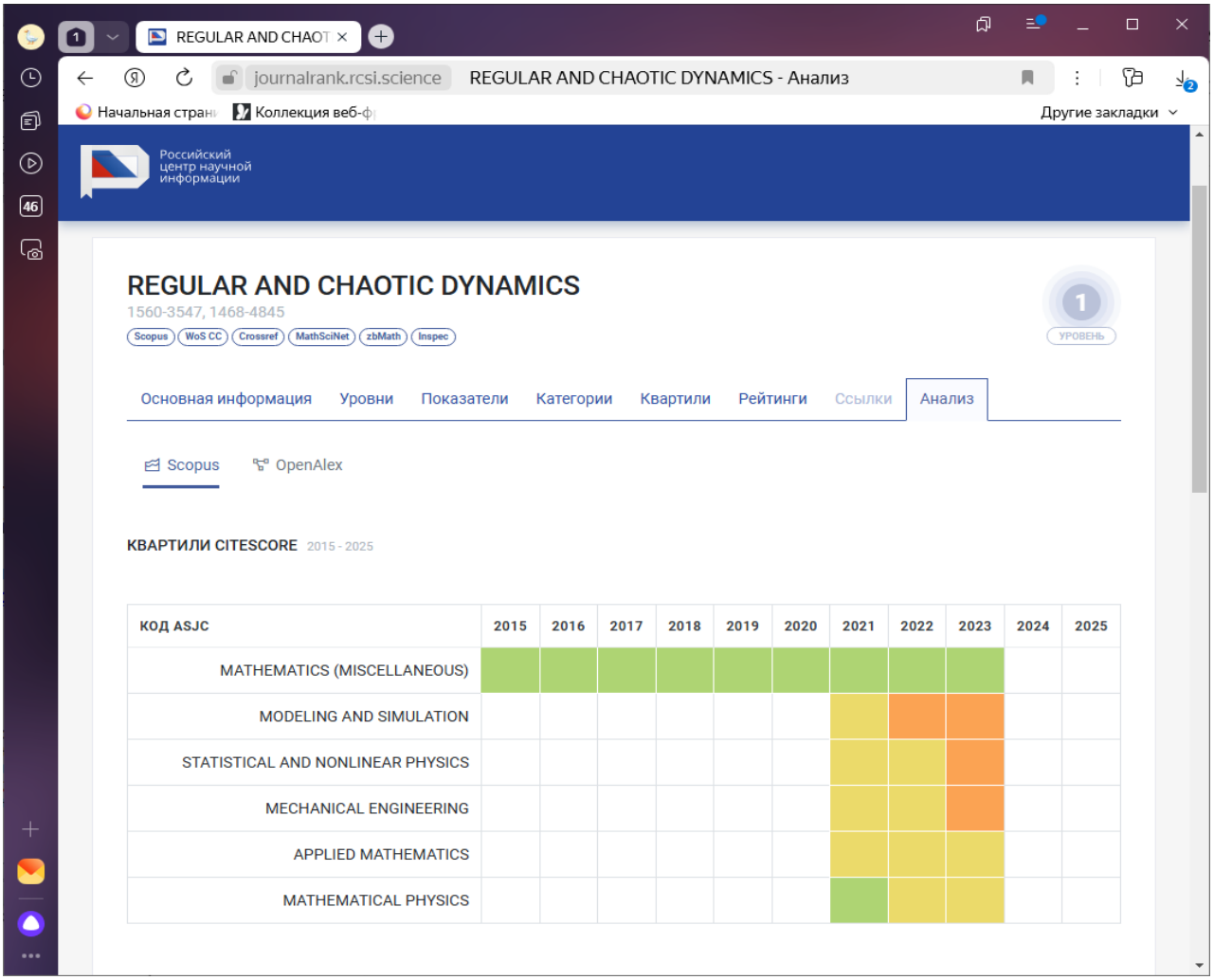


Рис. 8. «Белый список». В разделе «Анализ» показаны квартили журнала REGULAR AND CHAOTIC DYNAMICS в Scopus за 2015-2023 гг. в рейтинге *CiteScore* (данные по состоянию на май 2025 г.)

Обратим внимание, что этот журнал в том же 2023 г. в базе Scopus имел разные позиции в рейтингах *CiteScore* и *SJR*. В табл. 2 показаны квартили журнала в этих двух рейтингах. Показатель рейтинга *CiteScore* отражает среднее количество цитирований статей, показатель рейтинга *SJR* построен с учетом научной значимости цитирующего журнала.

Из табл. 2 следует, что в 1-й квартиль в базе Scopus журнал отнесен только по тематике MATHEMATICS (MISCELLANEOUS) в рейтинге *CiteScore*. По другим тематическим направлениям, а также в рейтинге *SJR* по всем шести направлениям журнал попадает во 2-й и 3-й квартили Scopus.

Табл. 2. Квартили журнала REGULAR AND CHAOTIC DYNAMICS в Scopus за 2015-2023 гг. в рейтинге *CiteScore* и *SJR* (по состоянию на май 2025 г.)

Тематика по ASJC	*Квартиль в CiteScore*	*Квартиль в SJR*
MATHEMATICS (MISCELLANEOUS)	1	2
MODELING AND SIMULATION	3	3
STATISTICAL AND NONLINEAR PHYSICS	3	3
MECHANICAL ENGINEERING	3	2
APPLIED MATHEMATICS	2	3
MATHEMATICAL PHYSICS	2	3

Возникают сомнения, насколько объективно журнал REGULAR AND CHAOTIC DYNAMICS отнесен к уровню У1 в БС в 2023 г. Но таковы алгоритмы системы, с которыми спорить трудно.

Таким образом, анализ показателей рейтингов ведущих русскоязычных математических журналов демонстрирует, что близкие по рейтингу Science Index РИНЦ eLibrary.ru математические журналы ожидаемо попали в одну категорию К1 в Перечне ВАК. В то же время два первых журнала по рейтингу Science Index РИНЦ eLibrary.ru почему-то отнесены к разным уровням БС в версии 2023 г.

На наш взгляд, некоторые ведущие журналы по математике из табл. 1 в 2023 г. оказались явно недооцененными в рейтинге БС с точки зрения русскоязычных авторов, ориентированных на показатели российской библиографической базы РИНЦ eLibrary.ru.

Разумеется, ранжирование ведущих математических российских журналов в БС версии 2025 г., как показано в табл. 1, выглядит более адекватно.

ИНФОРМАЦИОННАЯ СИСТЕМА «МЕТАФОРА»

На сайте РЦНИ сообщается, что, согласно приказу Минобрнауки РФ № 337 от 11 апреля 2025 г. о внесении изменений в правила формирования перечня рецензируемых научных изданий [6], в сентябре 2025 г. РЦНИ вводит в эксплуатацию информационную систему (ИС) «Метафора» для сбора данных о научных журналах и опубликованных статьях. Редакциям научных журналов ИС

«Метафора» предоставляет различные способы передачи информации об опубликованных научных статьях:

— API;

— загрузка XML-файлов в схемах JATS и eLIBRARY.RU (Science Index и Science Space);

— загрузка данных из Национальной платформы периодических научных изданий РЦНИ;

— веб-интерфейс для ручного ввода сведений о публикациях.

Для работы в системе «Метафора» организация учредителя или издателя научного журнала должна быть зарегистрирована в комплексной информационно-аналитической системе (КИАС) РЦНИ. Все уполномоченные участники процесса передачи данных в ИС «Метафора» должны иметь учетную запись в КИАС РЦНИ, а также специальное разрешение (назначение) со стороны руководителя организации. Все действия участников, связанные с вводом карточек изданий и передачей данных в ИС «Метафора», выполняются в личных кабинетах КИАС РЦНИ. Следует отметить, что многие ученые, участвовавшие в предыдущие годы в конкурсах на получение грантов Российского фонда фундаментальных исследований, имеют опыт работы в КИАС. ИС «Метафора» реализована как новая функция системы КИАС на фоне других функций, связанных с обслуживанием заявок на получение грантов и выполнением научных проектов.

На рис. 9 показан интерфейс личного кабинета пользователя КИАС, где функции ИС «Метафора» представлены на отдельной вкладке (ИС «Метафора») наряду с другими вкладками, не имеющими к ИС «Метафора» никакого отношения. Используется также введенный ранее порядок обслуживания заявок, когда после создания и редактирования заявки процедура передачи заявки в систему заканчивается таким специфическим действием, как *подписание заявки*. На рис. 9 показан статус заявки на ввод в систему карточки журнала «Препринты ИПМ им. М.В. Келдыша» — «подписана».

Этот статус позволяет перейти к дальнейшему диалогу ИС «Метафора» с пользователем, связанным с подписанием предлагаемого РЦНИ лицензионного договора на передачу и использование данных (метаданных и полных текстов) выпусков журнала, где Лицензиаром выступает издатель

научного журнала, а Лицензиатом — РЦНИ. Лицензионный договор является договором присоединения и не предполагает индивидуальное обсуждение и коррекцию текста документа — текст договора должен приниматься «как есть». Лицензия предоставляется Лицензиату (РЦНИ) на неисключительной основе.

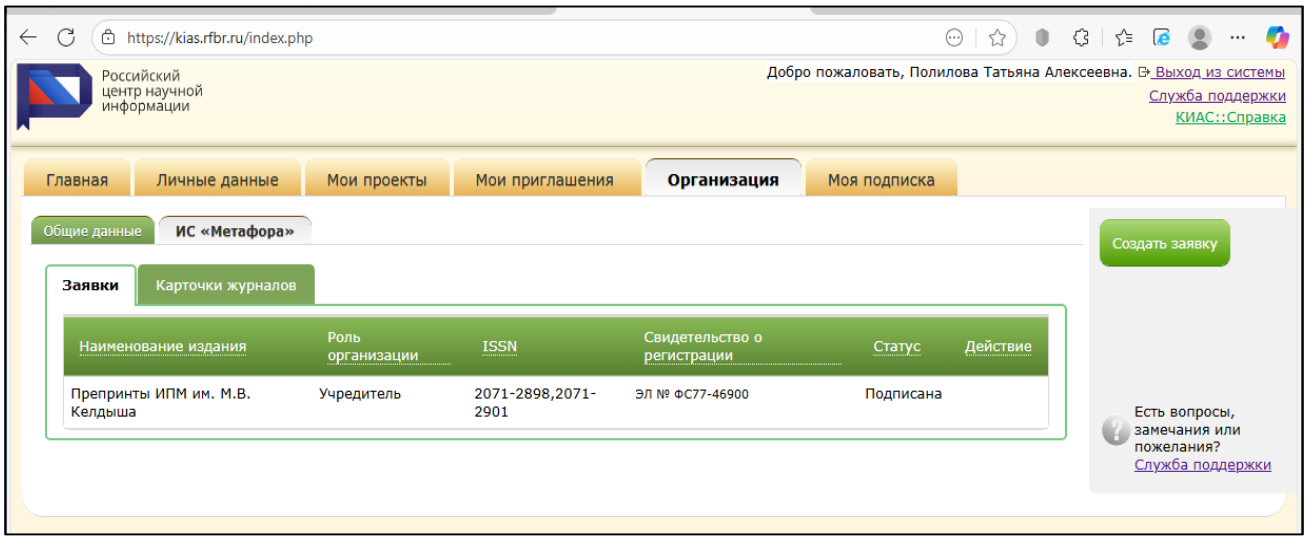


Рис. 9. Личный кабинет в КИАС РЦНИ, обеспечивающий работу с ИС «Метафора».

Приведем список метаданных, которые могут передаваться в ИС «Метафора» согласно лицензионному договору:

– наименование статьи;

– сведения об авторах статьи (в том числе ФИО, ученая степень, ученое звание, место работы, персональные идентификаторы);

– идентификаторы статьи;

– наименование научного издания (журнала) и его идентификаторы;

– сведения об источниках финансирования;

– УДК/ББК либо другие библиотечно-библиографические классификационные и предметные индексы;

– библиографический список;

– ключевые слова произведения/статьи;

– сведения о конфликте интересов авторов;

– сведения о вкладе авторов;

– сведения о датах поступления, принятия рукописи к публикации;

– сведения о дате и причинах ретракции для ретрагированных статей;

– сведения о правах на использование статьи;

– аннотация статьи.

Перечень метаданных достаточно широкий. Но, как указано в лицензии, перечисленные данные передаются только при наличии этих данных. Отсутствие некоторых метаданных не нарушает условия лицензионного договора.

Обратим внимание, что лицензионный договор разрешает использование текстов произведений в целях развития нейросетей, систем с применением моделей машинного обучения и больших языковых моделей. Таким образом, передавая РЦНИ права на использование текстов статей, можно ожидать дальнейшее развитие русскоязычных систем искусственного интеллекта в научной области.

К концу октября 2025 г. более тысячи организаций приступили к работе в ИС «Метафора» и внесли информацию о более чем 1600 журналах.

ЗАКЛЮЧЕНИЕ

В настоящее время в соответствии с требованием ВАК редакции журналов из Перечня ВАК размещают метаданные журналов в российском индексе научного цитирования (РИНЦ) eLibrary.ru. В последние два года появились новые информационные системы, индексирующие российские научные журналы. Наряду с РИНЦ eLibrary.ru, журналы из Перечня ВАК индексируются в базе проекта «Российские научные журналы» (РНЖ) Российского научно-исследовательского института экономики, политики и права (РИЭПП) и в ИС «Метафора» Российского центра научной информации (РЦНИ) с последующим включением сведений о журналах Перечня ВАК в БС. Стоит ли ожидать, что указанные базы РИНЦ eLibrary.ru, РНЖ и «Метафора» наладят оперативное взаимодействие, освобождая редакции журналов от необходимости дублировать ввод данных в каждую из этих баз?

База РНЖ от РИЭПП и «Метафора» от РЦНИ в отношении журналов из Перечня ВАК выполняют одну и ту же функцию — размещают на своих мощностях метаданные (возможно, полные тексты) выпусков журналов из Перечня ВАК. Оправданно ли подобное дублирование? Появятся ли новые возможности у пользователей этих баз, соискателей ученых степеней и

диссертационных советов по сравнению с теми возможностями, которые сейчас предоставляет РИНЦ eLibrary.ru? Редакции научных журналов ждут ответов на поставленные вопросы.

СПИСОК ЛИТЕРАТУРЫ

1. *Полилова Т.А.* Доступ к Перечню ВАК через интерфейс пользователя // Научный сервис в сети Интернет: труды XXV Всероссийской научной конференции (18–21 сентября 2023 г., онлайн). М.: ИПМ им. М.В.Келдыша, 2023. С. 298–307. https://doi.org/10.20948/abrau-2023-34; https://keldysh.ru/abrau/2023/theses/34.pdf

2. *Полилова Т.А.* Перечень ВАК: интерфейс пользователя в базе РНЖ и eLibrary.ru // Электронные библиотеки. 2024. Т. 27. № 1. С. 43–64. https://rdl-journal.ru/article/view/820/885. Перевод: *Polilova T.A.* List of the Higher Attestation Commission: User Interface in the RSJ Database and Elibrary.Ru. Automatic Documentation and Mathematical Linguistics. 2024. Vol. 58 (Suppl 1), S17–S26. https://doi.org/10.3103/S0005105525700074; https://link.springer.com/article/10.3103/S0005105525700074

3. *Письмо* от 10 февраля 2023 г. № 4/3-разн «О заполнении данных в личных кабинетах журналов Перечня ВАК». https://rng.riep.ru/help/recomend.pdf

4. *Письмо* от 10 февраля 2023 г. № 4/3-разн «О заполнении данных в личных кабинетах журналов Перечня ВАК». https://rng.riep.ru/help/recomend.pdf

5. *Письмо* ВАК РФ от 6 декабря 2022 г. № 02-1198 «О Перечне рецензируемых научных изданий». https://www.garant.ru/products/ipo/prime/doc/405821249/?ysclid=lilhqkkqwg60916187

6. *Приказ* Министерства науки и высшего образования Российской Федерации от 11.04.2025 № 337 о внесении изменений в правила формирования перечня рецензируемых научных изданий. http://publication.pravo.gov.ru/document/0001202505120037

7. *О регистрации* научных журналов из перечня ВАК в Российском центре научной информации. https://www.rcsi.science/press-center/news/novosti-organizatsii/o-registratsii-nauchnykh-zhurnalov-iz-perechnya-vak-v-rossiyskom-tsentre-nauchnoy-informatsii/

8. *О регистрации* научных журналов из перечня ВАК в Российском центре научной информации.
https://vk.com/wall-98716205_1963?w=wall-98716205_1963&ysclid=mb9j7zlfct962465086

9. *Российский центр научной информации* (РЦНИ). https://www.rcsi.science/

10. *Russian Science Citation Index*. https://www.elibrary.ru/project_rsci.asp

11. *Российский центр научной информации* (РЦНИ). Список журналов. https://journalrank.rcsi.science/ru/

12. *Утверждены правила* распределения по категориям научных изданий «Белого списка».
https://www.minobrnauki.gov.ru/press-center/news/nauka/68029/

13. *Российский центр научной информации* (РЦНИ). Часто задаваемые вопросы. https://journalrank.rcsi.science/ru/info/

14. *Российский центр научной информации* (РЦНИ). «Белый список» научных журналов.
https://rcsi.science/activity/belyy-spisok/?ysclid=mcn6q2p7p8641242799

15. *Издательский дом «Научное обозрение»*. Где разрешено публиковаться российским ученым?
https://russian-science.info/gde-razresheno-publikovatsya-rossijskim-uchenym-belyj-spisok

LIST OF HIGHER ATTESTATION COMMISSION JOURNALS AND OTHER RUSSIAN INDEXES

T. A. Polilova[0000-0003-4628-3205]

Keldysh Institute of Applied Mathematics, Moscow, Russia

polilova@keldysh.ru

Abstract

In accordance with the requirement of the Higher Attestation Commission (HAC), journal issue data from the 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 (HAC List) have been regularly published in the Russian Science Citation Index (RSCI) in the bibliographic database eLibrary.ru for more than 20 years. In March 2023, the editorial offices of journals from the HAC List, in accordance with the recommendation of the HAC, have post data of 2022 year issues in the Russian Scientific Journals database (RSJ) created by the Russian Scientific Research Institute RIEPP. In April 2025, by order of the Ministry of Science and Higher Education of the Russian Federation, a new requirement was added — for a journal from the HAC List, along with registration in the RISC eLibrary.ru, registration in the Information System (IS) “Metaphora”, developed by the Russian Center for Scientific Information, is required. Journals from the HAC List are recommended to regularly transfer metadata of published issues of journals to the “Metaphora” through specially organized interfaces. What role do the RSJ and “Metaphora” databases play in the infrastructure of scientific publications?

In addition, according to commission of the Government of the Russian Federation, the Russian Center for Scientific Information performs the function of the operator of the “White List” of scientific journals. The “White List” in 2023 was formed by the Interdepartmental Working Group (IWG) of the Ministry of Education and Science of the Russian Federation. The "White List" is supposed to be used to monitor and evaluate the publication activity of Russian scientists. The "White List" currently includes about 29,000 English-language international journals and about 1,000 Russian-language journals from the Russian Science Citation Index (RSCI) database. In 2025, the Russian-language part of the "White List" significantly

expanded due to the inclusion of journals from HAC List into the "White List". We would like to receive detailed information from the ideologists of the "White List" on how the levels (U1, U2, U3, U4) of the “White List” journals and the categories (K1, K2, K3) of journals on the HAC List will correspond?

Keywords*: List of Higher Attestation Commission, RSCI, eLibrary.ru, RSJ database, Information System "Metaphora, "White List".*

REFERENCES

1. *Polilova T.A.* Dostup k Perechniu VAK cherez interfeis polzovatelia // Nauchnyi servis v seti Internet: trudy XXV Vserossiiskoi nauchnoi konferentsii (18–21 sentiabria 2023 g., onlain). M.: IPM im. M.V. Keldysha, 2023. S. 298–307. https://doi.org/10.20948/abrau-2023-34;
https://keldysh.ru/abrau/2023/theses/34.pdf

2. *Polilova T.A.* Perechen VAK: interfeis polzovatelia v baze RNZh i eLibrary.ru // Elektronnye biblioteki. 2024. T. 27. № 1. S. 43–64.
https://rdl-journal.ru/article/view/820/885. Perevod: *Polilova T.A.* List of the Higher Attestation Commission: User Interface in the RSJ Database and Elibrary.Ru. Automatic Documentation and Mathematical Linguistics. 2024. Vol. 58 (Suppl 1), S17–S26. https://doi.org/10.3103/S0005105525700074;
https://link.springer.com/article/10.3103/S0005105525700074

3. *Pismo* ot 10 fevralia 2023 g. № 4/3-razn «O zapolnenii dannykh v lichnykh kabinetakh zhurnalov Perechnia VAK». https://rng.riep.ru/help/recomend.pdf

4. *Pismo* ot 10 fevralia 2023 g. № 4/3-razn «O zapolnenii dannykh v lichnykh kabinetakh zhurnalov Perechnia VAK». https://rng.riep.ru/help/recomend.pdf

5. *Pismo* VAK RF ot 6 dekabria 2022 g. № 02-1198 «O Perechne retsenziruemykh nauchnykh izdanii».
https://www.garant.ru/products/ipo/prime/doc/405821249/?ysclid=lilhqkkqwg60916187

6. *Prikaz* Ministerstva nauki i vysshego obrazovaniia Rossiiskoi Federatsii ot 11.04.2025 № 337 o vnesenii izmenenii v pravila formirovaniia perechnia retsenziruemykh nauchnykh izdanii.
http://publication.pravo.gov.ru/document/0001202505120037

7. *O registratsii* nauchnykh zhurnalov iz perechnia VAK v Rossiiskom tsentre nauchnoi informatsii.
https://www.rcsi.science/press-center/news/novosti-organizatsii/o-registratsii-nauchnykh-zhurnalov-iz-perechnya-vak-v-rossiyskom-tsentre-nauchnoy-informatsii/

8. *O registratsii* nauchnykh zhurnalov iz perechnia VAK v Rossiiskom tsentre nauchnoi informatsii.
https://vk.com/wall-98716205_1963?w=wall-98716205_1963&ysclid=mb9j7zlfct962465086

9. *Rossiiskii tsentr nauchnoi informatsii* (RTsNI). https://www.rcsi.science/

10. *Russian Science Citation Index*. https://www.elibrary.ru/project_rsci.asp

11. Rossiiskii tsentr nauchnoi informatsii (RTsNI). Spisok zhurnalov. https://journalrank.rcsi.science/ru/

12. *Utverzhdeny pravila* raspredeleniia po kategoriiam nauchnykh izdanii “Belogo spiska”. https://www.minobrnauki.gov.ru/press-center/news/nauka/68029/

13. *Rossiiskii tsentr nauchnoi informatsii* (RTsNI).Chasto zadavaemye voprosy. https://journalrank.rcsi.science/ru/info/

14. *Rossiiskii tsentr nauchnoi informatsii* (RTsNI). “Belyi spisok” nauchnykh zhurnalov. https://rcsi.science/activity/belyy-spisok/?ysclid=mcn6q2p7p8641242799

15. *Izdatelskii dom “Nauchnoe obozrenie”*. Gde razresheno publikovatsia rossiiskim uchenym?
https://russian-science.info/gde-razresheno-publikovatsya-rossijskim-uchenym-belyj-spisok

СВЕДЕНИЯ ОБ АВТОРЕ

ПОЛИЛОВА Татьяна Алексеевна – старший научный сотрудник Института прикладной математики им. М.В. Келдыша РАН, доктор физико-математических наук, лауреат Премии Президента РФ в области образования;

Tatyana Alekseevna POLILOVA – Senior Researcher at the Keldysh Institute of Applied Mathematics of the Russian Academy of Sciences.

email: polilova@keldysh.ru.

ORCID: 0000-0003-4628-3205

Материал поступил в редакцию 19 декабря 2025 года

УДК 004.056.5

ИССЛЕДОВАНИЕ АЛГОРИТМОВ ОБРАБОТКИ, ДЕТЕКЦИИ И ЗАЩИТЫ ДАННЫХ С ЦЕЛЬЮ МИНИМИЗАЦИИ ВОЗДЕЙСТВИЯ ВРЕДОНОСНОГО ПО И ФИШИНГОВЫХ АТАК НА ПОЛЬЗОВАТЕЛЕЙ ЦИФРОВЫХ ПЛАТФОРМ

Т. С. Волокитина[1] [0000-0002-5493-447X], **М. О. Таныгин**[2] [0000-0002-4099-1414]

[1, 2]*Юго-Западный государственный университет», г. Курск, Россия*

[1]tativolokitina@gmail.com, [2]tanygin@yandex.ru

Аннотация

Статья посвящена разработке научно-методического аппарата повышения эффективности защиты цифровых платформ от киберугроз путем создания алгоритмов обработки и детекции с учетом когнитивных особенностей пользователей. Предложена концептуальная модель трехэтапной системы защиты, интегрирующая технические механизмы безопасности с когнитивными моделями принятия решений. Разработан алгоритм эвристической детекции на основе машинного обучения Random Forest с анализом 47 признаков, включающих технические характеристики URL и когнитивно-семантические характеристики контента. Создана методика динамической интеграции четырех источников данных об угрозах, сокращающая время реагирования с 12–14 ч. до 2 ч. Предложен алгоритм рекурсивного анализа цепочек перенаправлений глубиной до десяти уровней для обнаружения замаскированных угроз. Экспериментальная валидация на эмпирической базе объемом около миллиона записей подтвердила точность детекции 87% при обработке ста тысяч записей в час. Разработанные решения обеспечивают соответствие требованиям ГОСТ Р 57580.1–2017 и российского законодательства в области защиты персональных данных.

Ключевые слова*: эвристическая детекция угроз, машинное обучение, когнитивная безопасность, фишинговые атаки, социальная инженерия, защита данных, интеграция источников угроз.*

ВВЕДЕНИЕ

Стремительное развитие цифровых платформ в Российской Федерации сопровождается критическим ростом киберугроз, эксплуатирующих когнитивные уязвимости пользователей. По данным МВД России, в 2023 г. зарегистрировано свыше 50 тыс. преступлений в сфере информационных технологий, что на 47% превышает показатели предыдущего периода [1, 2]. Особую опасность представляют фишинговые атаки и вредоносное программное обеспечение, распространяемые через социальные сети, где 78% атак осуществляется путем манипуляции восприятием и доверием пользователей [3, 4].

В результате анализа существующих технических решений установлена фундаментальная проблема: традиционные системы защиты, основанные исключительно на черных списках URL и сигнатурном анализе, демонстрируют недостаточную эффективность вследствие игнорирования когнитивно-поведенческих факторов пользователей [5, 6]. Задержки обновления черных списков составляют 12–14 ч. для Google Safe Browsing и 24–48 ч. для реестра Роскомнадзора, в течение которых реализуется 70% кликов пользователей на вредоносные ссылки из-за максимальной когнитивной уязвимости в период новизны угрозы [7].

Когнитивный аспект проблемы усугубляется низкой цифровой грамотностью 60% российских пользователей, которые не способны распознавать признаки социальной инженерии и игнорируют предупреждения систем безопасности вследствие когнитивных искажений восприятия рисков [8–10]. До 70% пользователей игнорируют технические предупреждения из-за когнитивной перегрузки, привычки к игнорированию сообщений и искажений оценки вероятности рисков.

Теоретические основы защиты от фишинга с учетом когнитивных факторов заложены в работах, исследовавших демографические факторы восприимчивости к фишингу и эффективность образовательных интервенций (см., например [11]). Однако эти подходы ориентированы на западную аудиторию и не учитывают специфику когнитивных паттернов российских пользователей социальных сетей, включая эксплуатацию доверия к государственным брендам и омографические атаки с использованием кириллических символов.

В российской науке значительный вклад в развитие методов информационной безопасности внесли исследователи, изучавшие технические аспекты защиты [11–15]. Однако существующие работы недостаточно раскрывают проблемы интеграции технических средств защиты с когнитивно-поведенческими моделями пользователей.

Анализ показывает следующие ограничения: недостаточный учет когнитивных особенностей российских пользователей; слабая интеграция технических средств с когнитивно-поведенческими моделями; отсутствие комплексного анализа многоэтапной структуры защиты; недостаточное внимание к культурной специфике восприятия киберугроз.

Правовые аспекты проблемы регулируются Федеральными законами № 152-ФЗ и № 149-ФЗ, обязывающими операторов цифровых платформ обеспечивать защиту от несанкционированного доступа [15, 16]. Требования ГОСТ Р 57580.1–2017 устанавливают минимальную эффективность защиты не менее 80%, что не достигается большинством платформ при учете когнитивных факторов пользователей [17].

Целью настоящего исследования была разработка научно-методического аппарата для повышения эффективности защиты цифровых платформ от фишинговых атак и вредоносного ПО путем создания алгоритмов обработки и детекции угроз, учитывающих когнитивные особенности пользователей.

Для достижения цели решались следующие задачи: разработка математической модели многоэтапной системы защиты; создание алгоритма эвристической детекции на основе анализа признаков; разработка методики интеграции источников данных; создание алгоритма анализа цепочек перенаправлений; экспериментальная проверка разработанных методов.

КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ МНОГОЭТАПНОЙ СИСТЕМЫ ЗАЩИТЫ С УЧЕТОМ КОГНИТИВНЫХ ФАКТОРОВ

Предложена концептуальная модель функционирования системы защиты цифровой платформы, интегрирующая технические механизмы безопасности с когнитивными моделями принятия решений пользователями. Модель описывает три последовательных этапа обработки угроз.

Этап публикации представляет первую линию защиты, где автоматические фильтры анализируют контент до его появления в ленте пользователей. Когнитивная оценка угрозы пользователями отсутствует, защита осуществляется исключительно техническими средствами: сигнатурный анализ вредоносных URL по черным спискам, эвристическая детекция на основе анализа признаков контента, проверка цифровых подписей. Эффективность этапа определяется коэффициентом обнаружения.

Этап перехода по ссылке характеризуется когнитивным принятием решения пользователем о клике на подозрительную ссылку. Данный этап критически важен с точки зрения когнитивной безопасности, поскольку здесь проявляются факторы доверия к источнику публикации, эмоциональной вовлеченности в контент, автоматизмов принятия решений. Эффективность определяется долей пользователей, отказавшихся от клика без технических предупреждений.

Этап браузерной защиты активируется при попытке перехода на URL, обнаруженный в черных списках браузера. Пользователю выдается предупреждение о риске с описанием потенциальной угрозы. Эффективность этапа определяется долей правильных реакций на предупреждения, зависящей от когнитивных факторов восприятия технических сообщений.

Концептуальная архитектура системы представлена 7 функциональными блоками, соединенными через системную шину для параллельной обработки данных. Блок аппаратного обеспечения содержит процессор с количеством ядер не менее 28, оперативную память объемом не менее 64 ГБ, устройства энергонезависимой памяти объемом не менее 1 ТБ, сетевые интерфейсы пропускной способностью не менее 10 Гбит/с.

Блок сбора данных выполнен с возможностью извлечения публикаций из API с обработкой исключительно публично доступных данных и анонимизацией персональных идентификаторов посредством SHA-256-хеширования в соответствии с требованиями ФЗ № 152. Блок содержит модуль асинхронных HTTP-запросов для параллельного извлечения не менее 100 тыс. записей в час.

Разработана математическая модель временной динамики когнитивной уязвимости пользователей, описывающая зависимость вероятности реализации клика от времени после публикации угрозы. Модель учитывает три ключевых

фактора: новизну контента в ленте, отсутствие технических предупреждений и эмоциональную вовлеченность.

Временная динамика характеризуется экспоненциальным убыванием вероятности клика. В первые два часа после публикации реализуется 25.2% кликов при максимальной когнитивной уязвимости из-за новизны угрозы и отсутствия информации в черных списках. За 12 ч. происходит 70% кликов, что определяет критический временной интервал. После 24 ч. вероятность клика стабилизируется на низком уровне 5–7% вследствие появления информации об угрозе в черных списках браузеров.

Экспериментально установлено, что задержка реагирования систем защиты критически влияет на эффективность. Традиционные системы с временем обновления черных списков 12–14 ч. перехватывают угрозу только после реализации 70% потенциальных кликов, обеспечивая защиту лишь для 30% пользователей. Сокращение времени реагирования до 2 ч. позволяет перехватить угрозу до реализации 74.8% кликов.

Предложен количественный индекс когнитивной уязвимости пользователей CVI, измеряющий степень подверженности социальной инженерии с учетом базовых цифровых навыков и способности правильно реагировать на предупреждения. Индекс вычисляется как произведение двух компонентов: первый представляет базовую когнитивную уязвимость и равен (1−G), где G является уровнем цифровой грамотности; второй компонент выражает поведенческую уязвимость и равен (1−P_реакция/P_эталон).

Уровень цифровой грамотности G измеряется как доля пользователей, преодолевших пороговое значение функциональной цифровой грамотности в стандартизированном тесте по методике ОЭСР, содержащем не менее 20 заданий по категориям: распознавание фишинга, оценка безопасности URL, понимание технологий защиты, практические действия при угрозах. Пороговое значение установлено на уровне не менее 55% правильных ответов. Для российских пользователей цифровых платформ с преобладанием возрастной группы 45+ значение G составляет 0.40 по данным Росстата 2023 г.

Фактическая вероятность правильной реакции измеряется в ходе контролируемых экспериментов с использованием открытых обезличенных данных через

официальный API. Анализ 1625 случаев отображения предупреждений браузеров показал, что только 487 случаев привели к правильной реакции, тогда как 1138 были проигнорированы. Фактическая вероятность составляет 0.30. Эталонная вероятность 0.70 установлена на основе международных исследований [18].

Для российской аудитории индекс CVI вычисляется: первый компонент равен (1−0.40) = 0.60; второй компонент равен (1−0.30/0.70) = 0.57; произведение компонентов дает CVI = 0.34. Значение интерпретируется как средняя когнитивная уязвимость аудитории на верхней границе перехода к высокой уязвимости.

МЕТОДОЛОГИЯ ПРОЕКТИРОВАНИЯ АЛГОРИТМОВ ДЕТЕКЦИИ

Разработан алгоритм эвристической детекции киберугроз на основе Random Forest с комплексным анализом технических и когнитивно-семантических признаков. Алгоритм построен в виде программного модуля, размещенного в оперативной памяти и связанного с процессором через системную шину для обработки потока данных не менее 100 тыс. записей в час.

Модуль извлечения признаков выполняет вычисления на процессоре для извлечения 47 признаков из каждой записи и каждого URL. Признаки разделяются на две категории.

Технические признаки URL включают 32 параметра, характеризующих структурные и сетевые свойства адреса. Длина доменного имени в символах используется для детекции аномально длинных доменов, типичных для фишинга. Наличие IP-адреса вместо буквенного имени домена индицирует попытку скрыть истинного владельца сайта. Возраст домена по данным WHOIS в днях позволяет выявить недавно зарегистрированные домены, характерные для одноразовых фишинговых сайтов. Наличие протокола HTTPS проверяется с учетом того, что его присутствие не гарантирует легитимность сайта.

Количество поддоменов анализируется для обнаружения попыток имитации через размещение названия бренда в поддомене вместо основного домена. Энтропия Шеннона доменного имени вычисляется для детекции случайно сгенерированных доменов. Наличие дефисов в домене проверяется как признак попытки имитации. Наличие символа @ в домене индицирует специфическую атаку с использованием правила парсинга URL браузерами.

Длина пути URL и количество параметров в строке запроса анализируются для выявления аномально сложных адресов с множественными параметрами, используемых для обфускации. Отношение длины домена к общей длине URL вычисляется для детекции URL с избыточно длинным путем. Сходство домена с известными брендами вычисляется через расстояние Левенштейна к списку из 50 брендов российских организаций.

Когнитивно-семантические признаки контента включают 15 параметров, характеризующих психологическое воздействие текста. Наличие ключевых слов социальной инженерии проверяется через список не менее 15 слов: срочно, выигрыш, блокировка, подтверждение, проверка, бесплатно, успей, последний день, ограниченное предложение, кликни, перейди, введи данные, подтверди личность, верни деньги, возврат.

Наличие названий брендов российских государственных организаций и коммерческих компаний проверяется через список не менее 50 брендов: Госуслуги, Сбербанк, ВТБ, Альфа-Банк, Тинькофф, Газпромбанк, Налоговая служба, ФНС, ПФР, МВД, Росреестр, ГИБДД, МФЦ, Почта России, Wildberries, Ozon, Яндекс, ВКонтакте. Анализ выполняется с учетом замены латинских символов на визуально сходные кириллические для детекции омографических атак типа sbe**г**bank.ru с кириллической буквой г вместо латинской r.

Наличие эмоциональных триггеров определяется через список слов, эксплуатирующих страх, жадность, любопытство. Присутствие призывов к немедленному действию проверяется через маркеры временного давления. Длина текстового содержания записи в символах анализируется с учетом того, что фишинговые публикации часто характеризуются краткостью. Количество восклицательных и вопросительных знаков подсчитываются как индикаторы эмоциональной окраски.

Наличие смешения латиницы и кириллицы анализируется для детекции омографических атак, где визуально сходные символы разных алфавитов используются для обмана. Например, в слове сбербанк буква е может быть заменена латинской e, визуально неотличимой, но имеющей другой код.

Модель Random Forest размещается в оперативной памяти объемом около 2 ГБ. Модель представляет собой ансамбль из 300 деревьев решений, каждое

с максимальной глубиной 15 уровней. Обучение модели проводилось на размеченном датасете, содержащем не менее 10 тыс. образцов URL, специфичных для российского сегмента, с балансом вредоносных и легитимных примеров 40:60.

Модуль классификации применяет модель к вектору из 47 признаков и формирует вероятность угрозы в диапазоне от 0 до 1. Классификация выполняется путем голосования деревьев: каждое из 300 деревьев выдает класс 0 для легитимного или 1 для вредоносного, затем вычисляется доля деревьев, проголосовавших за класс 1. Модуль обеспечивает точность классификации не менее 87% при обработке потока данных не менее 100 тыс. записей в час.

Разработана методика интеграции четырех внешних источников черных списков URL для повышения охвата обнаружения угроз и сокращения времени реагирования. Методика реализована в виде блока динамической интеграции источников данных с возможностью асинхронного опроса через сетевые интерфейсы.

Четыре внешних источника данных выбраны на основании анализа их характеристик. Google Safe Browsing обеспечивает наибольший охват и точность 85%, но характеризуется задержкой обновления 12–14 ч. PhishTank представляет краудсорсинговую базу с охватом 70% и задержкой 6–8 ч. OpenPhish обеспечивает охват 65% с задержкой 4–6 ч. Реестр Роскомнадзора охватывает специфичные для России угрозы с охватом 50% и задержкой 24–48 ч.

Весовые коэффициенты источников установлены на основе эмпирической калибровки: Google Safe Browsing – 0.35; PhishTank – 0.25; OpenPhish – 0.20; реестр Роскомнадзора – 0.20. Сумма весовых коэффициентов равна 1. Коэффициенты отражают компромисс между точностью и оперативностью источников.

Модули интерфейсов реализуют асинхронные HTTP-запросы с использованием библиотеки aiohttp для Python. Асинхронность обеспечивает параллельное выполнение четырех запросов одновременно, сокращая общее время проверки URL с потенциальных 8–12 с. до 2–3 с. Каждый модуль выполняет опрос своего источника с интервалом не более 30 мин.

Модуль кеширования результатов записывает результаты проверок в энергонезависимую память. Для каждого URL сохраняется запись в формате: хеш SHA-

256 от URL, временная метка проверки в формате Unix timestamp, 4 бинарных флага результатов от источников. Время жизни записи в кеше составляет 1 ч.

Модуль агрегирования вычисляет агрегированную оценку угрозы как взвешенную сумму бинарных откликов 4 источников по формуле:

$$0.35 \times \text{отклик\_GSB} + 0.25 \times \text{отклик\_PhishTank} +$$
$$+0.20 \times \text{отклик\_OpenPhish} + 0.20 \times \text{отклик\_РКН},$$

где каждый отклик принимает значение 1, если URL обнаружен, или 0, если не обнаружен.

Методика обеспечивает охват обнаружения угроз не менее 90% против 60–70% у отдельных источников за счет покрытия различных сегментов пространства угроз. Сокращение времени реагирования до 2 ч. достигается за счет интервала опроса 30 мин. и интеграции источников с различными задержками обновления.

Разработан алгоритм рекурсивного анализа цепочек перенаправлений для обнаружения замаскированных вредоносных ссылок, скрытых за несколькими уровнями сокращенных URL и HTTP-редиректов. Алгоритм учитывает когнитивные особенности восприятия пользователями сокращенных URL.

Модуль обнаружения систем сокращения URL проверяет каждый URL на принадлежность к известным сервисам путем сравнения доменного имени со списком не менее 50 доменов-сокращателей. Список включает глобальные сервисы: bit.ly, goo.gl, tinyurl.com, ow.ly; специфичные для России: vk.cc, clck.ru; сокращатели социальных платформ: okl.lt для Одноклассников, vk.link для ВКонтакте.

Модуль HTTP HEAD-запросов выполняет запрос типа HEAD к URL для получения финального адреса без загрузки полного контента. HEAD-запрос возвращает только HTTP-заголовки без тела ответа, что экономит пропускную способность. Параметр allow_redirects = True включает автоматическое следование редиректам для получения конечного URL в цепочке.

Процесс раскрытия ограничен тайм-аутом не более 5 с. на 1 URL для предотвращения атак типа бесконечный редирект. Установлена максимальная глубина цепочки (не более 10 перенаправлений) на основе анализа легитимных сокращателей. Превышение лимита глубины индицирует попытку обфускации и классифицируется как подозрительное поведение.

Буфер хранения цепочек размещается в оперативной памяти и фиксирует полную последовательность URL в цепочке от исходного до финального. Для каждого уровня сохраняется URL, HTTP-код ответа и заголовок Location. Зафиксированная цепочка передается через системную шину в блок эвристической детекции, который применяет модель Random Forest к финальному URL в цепочке для классификации угрозы.

Алгоритм обеспечивает эффективность обнаружения замаскированных угроз не менее 78% на тестовой выборке из 500 сокращенных URL с цепочками перенаправлений от 2 до 10 уровней. Сравнительный анализ показал, что системы защиты без анализа цепочек пропускают до 50% замаскированных угроз.

РЕАЛИЗАЦИЯ И ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА

Создан программный комплекс для экспериментальной проверки разработанных алгоритмов на реальных данных цифровой платформы Одноклассники. Программная составляющая зарегистрирована как программа для ЭВМ № 2025683166 от 02.09.2025 под названием OKPHISH, что подтверждает воспроизводимость технического решения.

Программный комплекс реализован на Python 3.9 с использованием библиотек: pandas 1.3 для обработки структурированных данных, scikit-learn 1.0 для реализации алгоритмов машинного обучения, aiohttp 3.8 для асинхронных HTTP-запросов, numpy 1.21 для численных вычислений, nltk 3.6 для лингвистического анализа текста.

Архитектура программного комплекса соответствует блочной структуре устройства. Модуль сбора данных реализован с использованием асинхронного программирования через библиотеку asyncio. Модуль подключается к API Одноклассников по протоколу HTTPS с аутентификацией OAuth 2.0, извлекая публично доступные записи с соблюдением ограничений скорости не более 300 запросов в минуту.

Модуль анонимизации применяет хеширование SHA-256 к персональным идентификаторам с добавлением соли длиной 32 байта в соответствии с требованиями ФЗ № 152-ФЗ. Соль генерируется единожды при инициализации системы и сохраняется в защищенном хранилище ключей операционной системы.

Модуль извлечения признаков реализован с использованием векторизованных операций numpy для обеспечения производительности. Технические признаки URL извлекаются через парсинг компонентов адреса библиотекой urllib.parse. Когнитивно-семантические признаки контента извлекаются через токенизацию текста библиотекой nltk с последующим поиском ключевых слов.

Модель Random Forest загружается из сериализованного файла формата pickle при инициализации для размещения в оперативной памяти. Обучение модели проводилось на выделенном вычислительном кластере с использованием процедуры кросс-валидации по 5 блокам. Финальная модель демонстрирует точность 87% на независимой тестовой выборке.

Экспериментальная валидация разработанных алгоритмов проводилась на эмпирической базе, полученной из открытых обезличенных данных пользовательских записей социальной сети Одноклассники через официальный API. Платформа выбрана как репрезентативный пример российской социальной сети с аудиторией 20 млн активных пользователей в месяц.

Период наблюдения составил с 01.02.2024 по 10.10.2025, общей продолжительностью 20 месяцев. Объем эмпирической выборки составил 1 млн записей, включающих 500 тыс. публикаций, 300 тыс. комментариев, 200 тыс. кликов по ссылкам с временными метками взаимодействий.

Обработка осуществлялась исключительно с публично доступными записями, размещенными пользователями в открытом доступе без ограничений видимости. Персональные идентификаторы необратимо хешировались алгоритмом SHA-256 с добавлением соли перед обработкой, что исключает возможность деанонимизации.

Идентификация вредоносных URL осуществлялась через комбинацию методов. Первичная верификация проводилась через проверку в черных списках Google Safe Browsing, PhishTank, OpenPhish и реестре Роскомнадзора. Дополнительная верификация выполнялась с помощью ручного экспертного анализа выборки из 500 URL специалистами в области информационной безопасности. Из массива 1 млн записей идентифицировано 5 тыс. записей с подтвержденными вредоносными URL.

Экспериментальная валидация подтвердила достижение заявленных технических характеристик. Точность детекции угроз блоком эвристической детекции составила 87% на независимой тестовой выборке объемом 5 тыс. URL, включающей 2500 вредоносных и 2500 легитимных адресов. Полнота обнаружения составила 85%, что означает корректную идентификацию 2125 вредоносных URL из 2500.

Сравнительный анализ с четырьмя распространенными системами защиты на той же тестовой выборке показал превосходство разработанного алгоритма. Google Safe Browsing продемонстрировал точность 85%, PhishTank 79%, Yandex Safe Browsing 83%, Kaspersky URL Advisor 84%. Разработанный алгоритм превосходит ближайшего конкурента на 2% по точности и на 5% по полноте обнаружения.

Статистическая значимость различий с системой Google Safe Browsing проверена при помощи теста χ^2. Значение статистики составило 12.4 с. p-значением 0.002, что подтверждает статистически значимое превосходство разработанного алгоритма на уровне достоверности, равном 98%.

Разработанный алгоритм имеет явное преимущество для угроз, специфичных для российского сегмента. На подвыборке из 500 URL с имитацией российских брендов точность составила 92% против 78% у Google Safe Browsing. Это объясняется включением признаков анализа кириллических омографических атак типа sbe**r**bank.ru.

Время реагирования на новые угрозы составило 2 ч. благодаря интеграции источников с асинхронным опросом каждые 30 мин. и блока эвристической детекции, работающего независимо от черных списков. Для сравнения: аналоги демонстрируют время реагирования 12–14 ч. для Google Safe Browsing и 24–48 ч. для реестра Роскомнадзора.

Охват обнаружения угроз составил 90% за счет интеграции четырех источников данных против 60–70% у отдельных источников. Эффективность обнаружения замаскированных угроз блоком анализа цепочек перенаправлений составила 78% на выборке из 500 сокращенных URL с цепочками от 2 до 10 уровней.

Анализ паттернов взаимодействия пользователей с вредоносным контентом выявил критическую роль когнитивных факторов. Из 5 тыс. выявленных угроз

системы фильтрации заблокировали 2 тыс. угроз, обеспечив коэффициент обнаружения 40%. Оставшиеся 3 тыс. угроз прошли фильтры и появились в лентах пользователей.

1350 угроз привели к реализованным кликам пользователей, преодолевших все уровни технической защиты. Анализ показал доминирующую роль доверия к источнику: 743 клика (55%) произошли через контент от пользователей в списке друзей, 135 кликов (10%) от участников тех же социальных групп, 472 клика (35%) от незнакомых пользователей.

Фактор воспринимаемой срочности играл роль в 405 кликах (30% от общего числа). Публикации с явными маркерами временного давления демонстрировали на 40% более высокую эффективность. Социальное подтверждение влияло на 270 кликов (20%). Критическим порогом оказалось наличие минимум 15–20 позитивных реакций.

Анализ реакции на предупреждения браузеров выявил критически низкий уровень эффективности защитных сообщений. Из 1200 угроз, обнаруженных браузерами и сопровожденных предупреждениями, только 360 пользователей (30%) прекратили попытку перехода. Остальные 840 пользователей (70%) проигнорировали предупреждения.

Демографический анализ показал значительную вариацию уязвимости. Пользователи старше 50 лет составляли 540 жертв (40%) при доле в общей аудитории 30%. Жители населенных пунктов с населением менее 100 тыс. человек составляли 877 жертв (65%) при доле в общей аудитории 45%.

ЗАКЛЮЧЕНИЕ

Решена важная научно-техническая задача повышения эффективности защиты цифровых платформ от фишинговых атак и вредоносного ПО путем разработки алгоритмов обработки и детекции угроз с учетом когнитивных особенностей пользователей и обеспечения соответствия требованиям российского законодательства.

Основные научные результаты включают разработку концептуальной модели трехэтапной системы защиты, интегрирующей технические механизмы безопасности с когнитивными моделями принятия решений пользователями. Модель учитывает временные характеристики когнитивной обработки информации.

Предложен алгоритм эвристической детекции на основе Random Forest с комплексным анализом 47 технических и когнитивно-семантических признаков. Алгоритм обеспечивает точность классификации 87% при обработке 100 тыс. записей в час. Показана высокая эффективность алгоритма для обнаружения угроз, специфичных для российского сегмента, установлена точность 92% против 78% у глобальных систем защиты.

Разработана методика динамической интеграции 4 разнородных источников данных об угрозах, позволяющая увеличить охват обнаружения до 90% и сократить время реагирования с 12–14 ч. до 2 ч. Сокращение времени реагирования критически важно с учетом временных характеристик когнитивной уязвимости: 70% кликов происходит в первые 12 ч. после публикации угрозы.

Создан алгоритм рекурсивного анализа цепочек перенаправлений глубиной до 10 уровней с эффективностью 78%, учитывающий когнитивные особенности восприятия сокращенных URL. Алгоритм обнаруживает замаскированные угрозы, которые пропускаются системами без анализа цепочек в 50% случаев.

Обоснована система показателей эффективности защиты, интегрирующая технические метрики с когнитивно-поведенческими индикаторами. Система адаптирована к требованиям российского законодательства ФЗ № 152-ФЗ, ФЗ № 149-ФЗ и ГОСТ Р 57580.1–2017.

Практическая значимость результатов подтверждена экспериментальной проверкой на реальных данных социальной сети Одноклассники объемом 1 млн записей за период 20 месяцев. Внедрение разработанных алгоритмов позволяет повысить эффективность защиты с 67% базового уровня до 80%, что соответствует требованиям ГОСТ Р 57580.1–2017.

Когнитивный анализ 1350 успешных кибератак выявил доминирующую роль доверия к источнику информации, определяющего 55% инцидентов. Критически низкая реакция пользователей на предупреждения браузеров на уровне

30% указывает на проблему привыкания к предупреждениям и необходимость разработки адаптивных форматов коммуникации рисков.

Направления дальнейших исследований включают адаптацию разработанных алгоритмов к другим российским социальным платформам с учетом специфики их аудиторий. Перспективным является исследование влияния культурных и возрастных факторов на когнитивную уязвимость для разработки более точных моделей сегментации пользователей.

СПИСОК ЛИТЕРАТУРЫ

1. *Селиверстов В.В., Корчагин С.А.* Анализ актуальности и состояния современных фишинг-атак на объекты критической информационной инфраструктуры // Инженерный вестник Дона. 2024. № 6 (114). С. 17.

2. *Group-IB.* Отчет о киберугрозах в России за 2023 год: анализ трендов и прогнозы. М.: Group-IB, 2024. 89 с.

3. *Kaspersky Lab.* Развитие киберугроз в 2023 году: статистика и аналитика инцидентов информационной безопасности. М.: Лаборатория Касперского, 2024. 156 с.

4. *Русских Е.И.* Прошлое, настоящее и будущее фишинговых атак // ББК 1 Н 34. С. 6015.

5. *Назаров А.К.* Некоторые современные средства защиты от киберугроз // редакционно-издательским советом Краснодарского университета МВД России. С. 76.

6. *Брюханов В.А., Грызунов В.В., Шестаков А.В.* Выявление проблем информационной безопасности методом систематического обзора литературы. 2024.

7. *Токолов А.В.* Социальная инженерия в вопросах обеспечения информационной безопасности // Криминологический журнал. 2024. № 4. С. 175–182.

8. *Горбунова Е.А., Сайкинов В.Е.* Российская Федерация. Проблема фишинга в использовании информационных систем на основе облачных технологий // И74 Информационное общество: современное состояние и перспективы развития: сборник материалов XI международного студенческого форума. Краснодар: КубГАУ, 2018. С. 103.

9. *Сергеев А.Ю., Широкова О.В.* Мошенничество в цифровом обществе в условиях социальных изменений // Цифровая социология. 2023. Т. 6, № 1. С. 59–71.

10.*Мрочко В.Л., Рощина Т.М., Тарасов М.Д.* Обеспечение безопасности в сети Интернет: психолого-педагогические аспекты // Экономические и социально-гуманитарные исследования. 2024. № 3 (43). С. 196–204.

11.*Серік А.С.* Правовые основы предотвращения кибермошенничества: состояние и перспективы развития. 2022.

12.*Швецова Е.Э.* Виды мошенничества в сфере дистанционного банковского обслуживания и способы борьбы с ними // Сборник материалов Всероссийской научной конференции молодых исследователей с международным участием ИНТЕКС-2024. 2024. С. 269–272.

13.*Уваров А.А.* Информационная безопасность граждан России: современное состояние // Lex russica. 2024. Т. 77, № 1 (206). С. 133–143.

14.*Харисова З.И.* Генезис преступности в сфере компьютерной информации и ее детерминанты // Общество, право, государственность: ретроспектива и перспектива. 2025. № 1 (21). С. 57–65.

15.*Битюкова А.Ф.* Направления развития банковских электронных услуг и способы обеспечения их безопасности. 2019.

16.*ГОСТ Р 57580.1-2017.* Безопасность финансовых (банковских) операций. Требования к организации и проведению работ по обеспечению безопасности. М.: Стандартинформ, 2017. 26 с.

17.*Федеральный закон от 27.07.2006 № 152-ФЗ «О персональных данных»* (ред. от 14.07.2022). Доступ из справочно-правовой системы «КонсультантПлюс».

18.*Федеральный закон от 27.07.2006 № 149-ФЗ* «Об информации, информационных технологиях и о защите информации» (ред. от 14.07.2022). Доступ из справочно-правовой системы «КонсультантПлюс».

19.*Sheng S., Holbrook M., Kumaraguru P., Cranor L.F., Downs J.* Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Atlanta, GA, USA, 2010. P. 373-382. https://doi.org/10.1145/1753326.1753383

20.*Guarino N.* Formal ontology, conceptual analysis and knowledge representation // Int. J. of Human Computer Studies. 1995. Vol. 43 (5/6). P. 625–640.

RESEARCH OF DATA PROCESSING, DETECTION AND PROTECTION ALGORITHMS TO MINIMIZE THE IMPACT OF MALWARE AND PHISHING ATTACKS ON USERS OF DIGITAL PLATFORMS

T. S. Volokitina[1] [0000-0002-5493-447X], M. O. Tanygin[2] [0000-0002-4099-1414]

[1,2]Southwest State University, Kursk, Russia

[1]tativolokitina@gmail.com, [2]tanygin@yandex.ru

Abstract

The article is devoted to the development of a scientific and methodological apparatus for improving the effectiveness of protecting digital platforms from cyber threats by creating processing and detection algorithms that take into account the cognitive characteristics of users. A conceptual model of a three-stage protection system is proposed, integrating technical security mechanisms with cognitive decision-making models. A heuristic detection algorithm based on Random Forest machine learning with analysis of 47 features, including technical URL characteristics and cognitive-semantic content characteristics, has been developed. A methodology for dynamic integration of four threat data sources has been created, reducing response time from 12–14 hours to two hours. An algorithm for recursive analysis of redirection chains up to ten levels deep to detect masked threats is proposed. Experimental validation on an empirical base of approximately one million records confirmed detection accuracy of 87% when processing one hundred thousand records per hour. The developed solutions ensure compliance with the requirements of GOST R 57580.1-2017 and Russian legislation in the field of personal data protection.

Keywords: *heuristic threat detection, machine learning, cognitive security, phishing attacks, social engineering, data protection, threat source integration.*

REFERENCES

1. *Seliverstov V.V., Korchagin S.A.* Analysis of the relevance and state of modern phishing attacks on critical information infrastructure objects // Engineering Bulletin of the Don. 2024. No. 6 (114). P. 17.

2. *Group-IB.* Report on cyber threats in Russia for 2023: analysis of trends and forecasts. Moscow: Group-IB, 2024. 89 p.

3. Kaspersky Lab. Development of cyber threats in 2023: statistics and analytics of information security incidents. Moscow: Kaspersky Laboratory, 2024. 156 p.

4. *Russkikh E.I.* Past, present and future of phishing attacks // BBK 1 N 34. P. 6015.

5. *Nazarov A.K.* Some modern means of protection against cyber threats // Editorial and publishing council of the Krasnodar University of the Ministry of Internal Affairs of Russia. P. 76.

6. Bryukhanov V.A., Gryzunov V.V., Shestakov A.V. Identification of information security problems by the method of systematic literature review. 2024.

7. *Tokolov A.V.* Social engineering in information security issues // Criminological Journal. 2024. No. 4. P. 175–182.

8. *Gorbunova E.A., Saykinov V.E.* Russian Federation The problem of phishing in the use of information systems based on cloud technologies // I74 Information Society: current state and development prospects: collection of materials of the XI international student forum. Krasnodar: KubSAU, 2018. P. 103.

9. *Sergeev A.Yu., Shirokova O.V.* Fraud in digital society under conditions of social change // Digital Sociology. 2023. Vol. 6, No. 1. P. 59–71.

10. *Mrochko V.L., Roschina T.M., Tarasov M.D.* Ensuring security on the Internet: psychological and pedagogical aspects // Economic and socio-humanitarian research. 2024. No. 3 (43). P. 196–204.

11. *Serik A.S.* Legal foundations for preventing cybercrime: state and development prospects. 2022.

12. *Shvetsova E.E*. Types of fraud in the field of remote banking and methods of combating them // Collection of materials of the All-Russian scientific conference of young researchers with international participation INTEX-2024. 2024. P. 269–272.

13. Uvarov A.A. Information security of Russian citizens: current state // Lex russica. 2024. Vol. 77, No. 1 (206). P. 133–143.

14. *Kharisova Z.I.* Genesis of crime in the field of computer information and its determinants // Society, law, statehood: retrospective and perspective. 2025. No. 1 (21). P. 57–65.

15. *Bityukova A.F.* Directions for the development of banking electronic services and methods of ensuring their security. 2019.

16. *GOST R 57580.1-2017.* Security of financial (banking) operations. Requirements for the organization and conduct of security work. Moscow: Standartinform, 2017. 26 p.

17. *Federal Law No. 152-FZ of July 27, 2006 "On Personal Data"* (as amended on July 14, 2022). Access from the reference legal system "ConsultantPlus".

18. Federal Law No. 149-FZ of July 27, 2006 "On Information, Information Technologies and Information Protection" (as amended on July 14, 2022). Access from the reference legal system "ConsultantPlus".

19. Sheng S., Holbrook M., Kumaraguru P., Cranor L.F., Downs J. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Atlanta, GA, USA, 2010. P. 373–382. https://doi.org/10.1145/1753326.1753383

20. Guarino N. Formal ontology, conceptual analysis and knowledge representation // Int. J. of Human Computer Studies. 1995. Vol. 43 (5/6). P. 625–640.

СВЕДЕНИЯ ОБ АВТОРАХ

ВОЛОКИТИНА Татьяна Сергеевна, окончила Юго-Западный государственный университет в 2021 г. Аспирант кафедры информационной безопасности Юго-Западного государственного университета. В списке научных трудов более 50 работ в области кибербезопасности и защиты информации.

Tatiana Sergeevna VOLOKITINA graduated from South-Western State University in 2021. She is currently a postgraduate student at the Department of Information Security of South-Western State University. She has authored more than 50 scientific publications in the fields of cybersecurity and information protection.

email: tativolokitina@gmail.com

ORCID: 0000-0002-5493-447X

ТАНЫГИН Максим Олегович, окончил Курский государственный технический университет в 2001 г., д. т. н. (2022). Доцент кафедры информационной безопасности Юго-Западного государственного университета. В списке научных трудов более 100 работ в области информационной безопасности и анализа данных.

Maxim Olegovich TANYGIN graduated from Kursk State Technical University in 2001, Doctor of Technical Sciences (2022). He is an Associate Professorat the Department of Information Security of South-Western State University. He has authored more than 100 scientific publications in the fields of information security and data analysis.

email: tanygin@yandex.ru

ORCID: 0000-0002-4099-1414

Материал проступил в редакцию 12 декабря 2025 года

УДК 004.41+004.9+004.5

РАЗРАБОТКА ЦИФРОВОЙ ПЛАТФОРМЫ СО ВСТРОЕННЫМ 3D-КОНФИГУРАТОРОМ ДЛЯ КАСТОМИЗАЦИИ ОДЕЖДЫ

Е. В. Евдущенко[1] [0000-0003-3692-2587], М. В. Шматко[2] [0000-0002-7255-8885]

[1]*Военный институт (инженерно-технический) Военной академии материально-технического обеспечения, г. Санкт-Петербург, Россия*
[2]*Омский государственный технический университет, г. Омск, Россия*

[1]elena.online_ktilp@mail.ru, [2]marin298@gmail.com

Аннотация

В условиях стремительного роста онлайн-продаж и запроса на персонализацию российский рынок кастомизированной одежды сталкивается с дефицитом технологичных и массово доступных решений. В статье представлены результаты исследовательско-внедренческого проекта по созданию мультибрендовой цифровой платформы со встроенным 3D-конфигуратором, нацеленного на трансформацию цикла предзаказа. Разработка позволяет покупателям интерактивно создавать модели одежды в веб-среде, а дизайнерам – оптимизировать логистику и минимизировать перепроизводство.

Основной научно-технический интерес в работе представляют детально описанная целевая архитектура платформы и масштабируемый конвейер обработки 3D-моделей, обеспечивающий их оптимизацию и корректное отображение в браузере. Дополнительный вклад составляет методика подготовки и оптимизации 3D-моделей одежды для веб-визуализации, формализованная в виде технических требований, которая позволяет обеспечить баланс визуального качества и производительности.

В результате исследования решена задача унификации форматов 3D-моделей одежды от различных дизайнеров в рамках мультибрендовой цифровой платформы (ключевого отличия от существующих монобрендовых решений) и реализована технология кастомизации с возможностью

интерактивного отображения всех видоизменений дизайна на одной экранной форме.

Технологическая состоятельность решения обоснована сравнительным анализом существующих аналогов, анализом рынка по модели PAM-TAM-SAM-SOM и оценкой функциональных требований.

В статье также представлена практическая стратегия внедрения цифровой платформы, что делает ее ценной для исследователей и специалистов, работающих на стыке e-commerce, компьютерной графики и цифровой трансформации бизнес-процессов.

Ключевые слова: *цифровая трансформация, веб-приложение, цифровая платформа, 3D-конфигуратор, 3D-модель, кастомизация одежды, виртуальная примерка, AR-примерка, технологический стек, архитектура, масштабирование, производительность.*

ВВЕДЕНИЕ

Согласно прогнозам, к 2030 г. цифровая трансформация индустрии моды в России и мире обеспечит увеличение доли онлайн-покупок одежды и аксессуаров на 50–68% [1]. В перспективе лидерами отрасли станут компании, которые смогут интегрировать массовые кастомизацию и персонализацию, а также обеспечить производство одежды с минимальными экономическими затратами [2–4]. По мнению специалистов, расширенная цифровая инженерия с привлечением самих покупателей к реалистичному прототипированию моделей одежды не только простимулирует сбыт, но и приведет к преобразованию традиционных производственных компаний, работающих в этой сфере [5–8].

В этом контексте проблема массовой кастомизации одежды может быть решена с помощью нового цифрового продукта, который обеспечит покупателям надлежащую доступность цифровых моделей одежды, а также возможность самостоятельно по своему вкусу создавать их конфигурацию.

Реалистично осуществлять кастомизацию цифровых моделей одежды позволяет 3D-конфигуратор. Это программа для создания и визуализации сложных объектов в 3D-представлении [9, 10]. Сегодня только несколько зарубежных монобрендовых онлайн-магазинов мужской одежды

предоставляют ее своим покупателям для массового использования (lanieri.com и suitsupply.com). У многих локальных российских брендов одежды есть запрос на аналогичные цифровые решения, которые помогут им повысить конкурентоспособность и оптимизировать бизнес-процессы, связанные с продвижением, изготовлением и продажей одежды по предзаказу.

В настоящей работе представлена информация о цифровом стартап-проекте, который был инициирован, исходя из оценки реальных потребностей роста и масштабирования российского сегмента кастомизированной одежды, изготовление которой осуществляется через длительный и экономически затратный цикл предзаказа. На текущий момент нами достигнута промежуточная цель – разработан и протестирован прототип мультибрендовой цифровой платформы со встроенным 3D-конфигуратором уровня TRL 5, определены целевые показатели ее массовой эксплуатации.

ЦИФРОВАЯ ТРАНСФОРМАЦИЯ ИНДУСТРИИ МОДНОЙ ОДЕЖДЫ

Современные цифровые технологии активно трансформируют индустрию модной одежды: исследования, проведенные среди представителей поколений X, Y и Z, подтверждают рост спроса на цифровую одежду, аватаров и NFT [11, 12]. Эта тенденция открывает для дизайнеров новые возможности, позволяя им использовать такие преимущества цифровой одежды, как создание материалов и дизайна, невозможных в физическом мире, снижение затрат на производство, транспортировку и физическое хранение, минимизация экологического ущерба и экономия природных ресурсов.

Пандемия COVID-19 ускорила развитие цифровых платформ для индустрии модной одежды, что привело к росту онлайн-покупок и необходимости адаптации к новым форматам взаимодействия между заказчиками и покупателями [13, 14]. Так, например, в 2023 г. в Лондоне появилась первая виртуальная примерочная ZERO10. Она позволила создателям одежды заявить о себе с помощью технологии дополненной реальности (AR). Еще ранее онлайн-сервис Virtusize предложил покупателям популярного магазина ASOS протестировать новый способ демонстрации размера модели, чтобы они могли выбирать одежду в соответствии с индивидуальными параметрами своей фигуры (см. рис. 1).

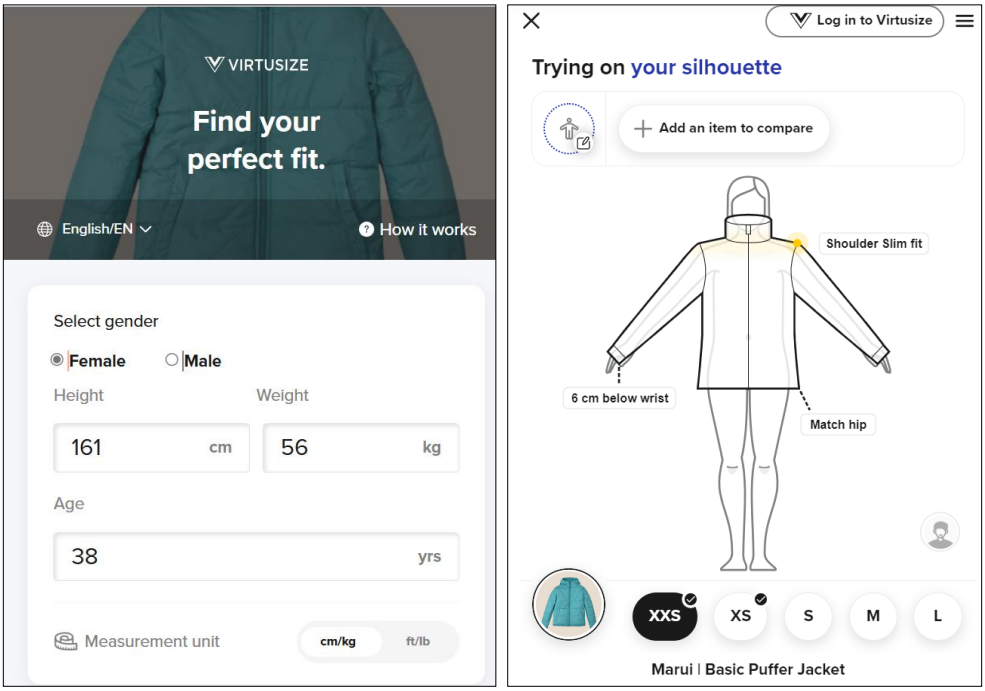


Рис. 1. Определение размера с помощью онлайн-сервиса Virtusize

Известный модный дом The Fabricant сегодня стал лидером в создании собственной виртуальной одежды и NFT, он устраивает онлайн-показы и поддерживает коллаборации с другими крупными модными домами, интегрируя digital-одежду в традиционную индустрию моды. На его цифровой платформе регулярно проводятся нейросетевые исследования, формируется тренд-аналитика [15].

Помимо виртуальной и дополненной реальностей многие крупные компании, создающие одежду, активно применяют технологии искусственного интеллекта. Так, например, японский бренд одежды Uniqlo одним из первых внедрил мобильного помощника Uniqlo IQ для персонализации рекомендаций своим покупателям.

В 2020 г. компания DressX запустила высокотехнологичную цифровую платформу, объединив все возможности цифровой трансформации индустрии моды. Сегодня DressX предлагает своим пользователям множество решений с использованием искусственного интеллекта, дополненной и виртуальной

реальностей, выпускает совместно с люксовыми брендами NFT drops, продает цифровую одежду, которую можно носить даже на игровой платформе Roblox.

Цифровая трансформация российской индустрии модной одежды также связана с появлением мультибрендовых цифровых платформ, таких как Ozon, Wildberries, ЯндексМаркет и Lamoda. Для многих малых и средних российских компаний, специализирующихся на производстве одежды, эти маркетплейсы стали эффективным каналом сбыта и взаимодействия с покупателями [16].

Другой интересный пример – это цифровая платформа Artisant, запущенная дизайнером из Уфы Регионой Турбиной в 2021 г. Она предоставила российским художникам возможность создавать, обменивать и продавать свои уникальные цифровые модели одежды (см. рис. 2).

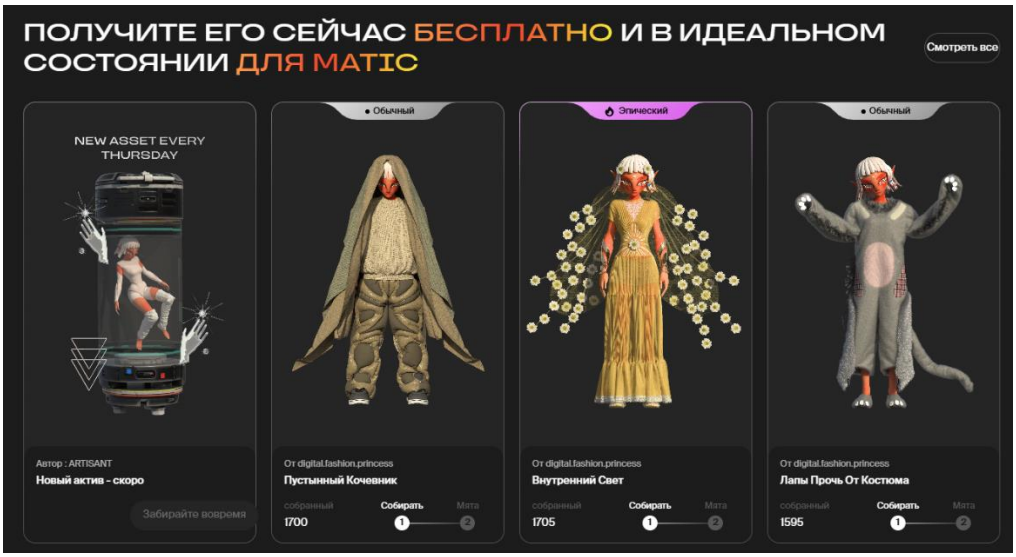


Рис. 2. Digital-образы на цифровой платформе Artisant

Сегмент кастомизированной модной одежды стал активно развиваться с внедрением 2D-, а затем 3D-конфигураторов, позволяющих покупателям собирать индивидуальные образы из цифровых компонентов одежды. Так, например, в онлайн-магазине российского производителя спортивной экипировки Safsport покупатели могут попробовать 2D-конфигуратор для создания необходимого им дизайна одежды, выбора цвета, декоративных элементов и текстуры материала (см. рис. 3).

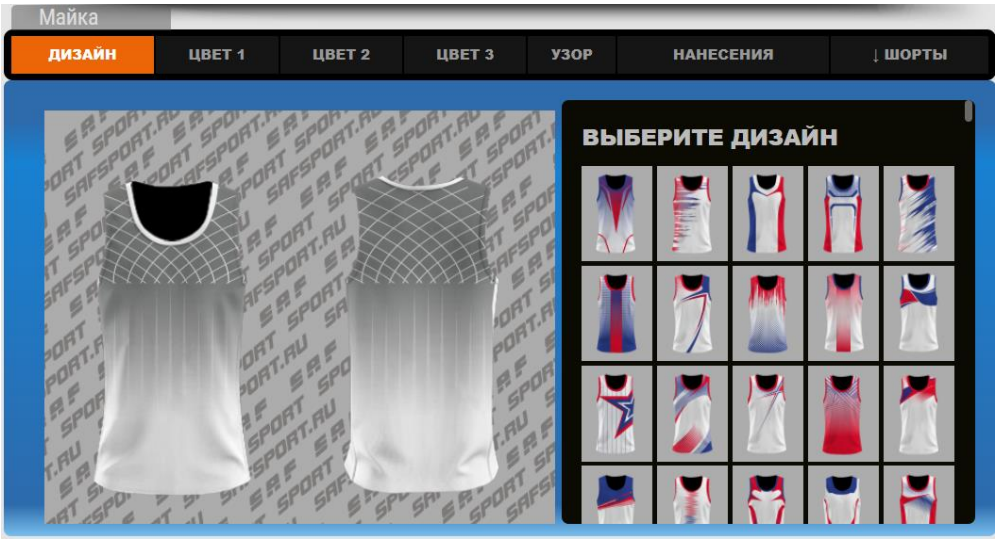


Рис. 3. 2D-конфигуратор онлайн-магазина спортивной экипировки Safsport

Такие известные в мире бренды мужской одежды, как Lanieri и Suitsupply, реализовали в своих онлайн-магазинах 3D-конфигураторы, с помощью которых можно кастомизировать модели пиджаков, рубашек и брюк (см. рис. 4).

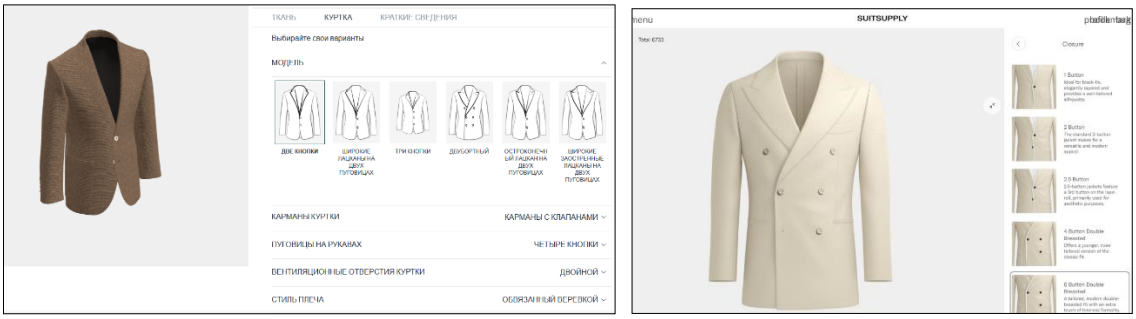


Рис. 4. 3D-конфигураторы мужской одежды на онлайн-магазинах итальянского бренда Lanieri и нидерладского бренда Suitsupply

Практика зарубежных монобрендовых онлайн-магазинов показала, что 3D-визуализация способствует увеличению конверсии (в среднем на 17–20%) и онлайн-продаж одежды (в среднем на 12–13%) [13]. В связи с этим интеграция функциональных возможностей 3D-конфигуратора и мультибрендовой цифровой платформы может оказать положительное влияние на развитие российской индустрии моды и повысить конкурентоспособность локальных брендов одежды.

КОНЦЕПЦИЯ И ФУНКЦИОНАЛЬНОСТЬ МУЛЬТИБРЕНДОВОЙ ЦИФРОВОЙ ПЛАТФОРМЫ

Целью разработки новой мультибрендовой цифровой платформы со встроенным 3D-конфигуратором – является цифровая трансформация ритейла кастомизированной одежды, созданной российскими дизайнерами. В 2024 г. стартап-проект получил финансовую поддержку Фонда содействия инновациями [17].

Основные функциональные преимущества цифровой платформы по отношению к существующим аналогам обеспечиваются 3D-конфигуратором. Он позволяет как дизайнерам, так и покупателям реализовывать технологию кастомизации – создавать индивидуальные образы (виртуальные модели одежды для мужчин, женщин и детей) из ограниченного набора шаблонов (компонентов одежды) и рассматривать их на экране компьютера со всех сторон (см. рис. 5). В перспективе планируется реализовать возможность наглядного просмотра цифровой модели одежды на 3D-аватаре с заданными параметрами роста, размера по линии груди, талии, бедер, а также цвета волос, глаз и пр.

Рис. 5. Пример вариантов кастомизации цифровой модели одежды

У разрабатываемой цифровой платформы две основные категории пользователей, за счет которых планируется осуществлять коммерциализацию стартап-проекта: российские дизайнеры кастомизированной одежды и

покупатели кастомизированной одежды. После внедрения публичной версии этой платформы дизайнеры кастомизированной одежды получат доступ к таким основным функциональным возможностям, как:

- создание личного кабинета юридического лица;
- загрузка 3D-компонентов для кастомизации цифровых моделей одежды;
- создание карточек под модели кастомизированной одежды;
- создание брендового магазина на цифровой платформе;
- продвижение магазина, бренда, коллекции одежды;
- получение заказов от покупателей на индивидуальное изготовление и доставку кастомизированных моделей одежды, а также уведомлений об их оплате;
- отслеживание данных веб-анализа о действиях покупателей (статистики посещения разных веб-страниц магазина, статистики по продажам, отзывам и пр.);
- обращение в службу технической поддержки;
- обращение в службу решения споров.

Покупатели кастомизированной одежды, в свою очередь, получат доступ к другим функциональным возможностям цифровой платформы, а именно:

- создание личного кабинета физического лица;
- поиск, сортировка, фильтрация и просмотр цифровых моделей в каталоге кастомизированной одежды;
- кастомизация цифровых моделей одежды;
- 3D-просмотр цифровых моделей одежды;
- сохранение вариантов своего дизайна моделей одежды;
- оформление заказов на индивидуальное изготовление и доставку кастомизированных моделей одежды, а также отслеживание статуса их выполнения;
- оплата заказа;
- публикация отзывов и вопросов о заказах, моделях, брендах;
- обращение в службу технической поддержки;
- обращение в службу решения споров.

Табл. 1. Оценка рынка мультибрендовой цифровой платформы со встроенным 3D-конфигуратором

Уровень потенциального рынка по модели PAM-TAM-SAM-SOM	Критерий расчета и емкость потенциального рынка
PAM – максимальный объем потенциального рынка	Общее количество российских брендов одежды: около 33 тыс. компаний
TAM – общий объем целевого рынка	Локальные российские бренды с производством одежды в России (в оценку не входят такие национальные бренды, как ZARINA, Ostin, Lime и пр., с производством за рубежом): около 20 тыс. компаний
SAM – доступный объем целевого рынка: доля от общего объема целевого рынка, которая активно использует различные цифровые решения конкурентов (30% от TAM)	Локальные российские бренды с производством одежды по предзаказу: около 6 тыс. компаний
SOM – реально достижимый объем целевого рынка: доля от доступного объема целевого рынка, которая может перейти на более функциональное и (или) технологичное цифровое решение от стартап-проекта (минимум 10% от SAM).	Локальные российские бренды с производством одежды по предзаказу: минимум 600 компаний

Для привлечения дополнительных инвестиций в проект по разработке мультибрендовой цифровой платформы со встроенным 3D-конфигуратором проведена оценка рынка по модели PAM-TAM-SAM-SOM (см. табл. 1). Она позволяет структурировано анализировать рыночный потенциал и разрабатывать стратегии для максимального использования доступных возможностей стартап-проекта.

Коммерциализацию стратап-проекта по разработке мультибрендовой цифровой платформы планируется осуществлять на основе бизнес-модели Freemium (асиметричная):

– для покупателей кастомизированной одежды, которые захотят получить индивидуальный предмет одежды, изготовленный дизайнером по созданной

ими самим цифровой 3D-модели, доступ к функциональным возможностям будет предоставляться бесплатно;

– для российских дизайнеров кастомизированной одежды (локальных российских брендов, дизайнеров в статусе самозанятых, швейных фабрик, ателье и пр.) доступ к функциональным возможностям будет предоставляться на платной основе, а стоимость доступа будет дифференцироваться в зависимости от необходимого им набора услуг.

Для второй категории пользователей, от которой зависит прямое поступление доходов в стартап-проект, ключевыми преимуществами перехода на новую цифровую платформу со встроенным 3D-конфигуратором (например, с маркетплейсов Wildberries, Ozon и пр.) являются:

– уменьшение доли технических остатков за счет перехода на изготовление изделий по предзаказам (решение проблемы перепроизводства и невостребованности отдельных размеров одежды, расцветок и моделей);

– сокращение затрат на разработку, внедрение, обслуживание и продвижение собственных онлайн-магазинов кастомизированной одежды;

– предварительная оценка спроса на разрабатываемые дизайнерами модели одежды (без запуска их в производство), а также снижение количества возвратов за счет 3D-визуализации;

– экономия ресурсов на проведение фотосессий для продвижения новых коллекций одежды.

В случае отсутствия достаточного объема инвестиций у стартап-проекта для выведения на рынок и продвижения цифровой платформы продумана альтернативная бизнес-модель – продажа одному из крупных маркетплейсов программного модуля с 3D-конфигуратором [18]. Его внедрение обеспечит маркетплейсу привлечение новых продавцов и покупателей, предпочитающих индивидуальный подход в одежде, а также позволит уменьшить количество возвратов и снизит нагрузку на логистику.

На основе серии глубинных интервью с дизайнерами, производителями и покупателями кастомизированной одежды были собраны и по модели Кано проанализированы данные о необходимых им дополнительных функциональных возможностях и других атрибутах качества цифровой

платформы [19]. Это позволило определить перспективные направления для ее развития и масштабирования (см. табл. 2).

Табл. 2. Варианты развития мультибрендовой цифровой платформы со встроенным 3D-конфигуратором

Целевые сегменты	Базовый цифровой продукт	Что можно предложить и продать дополнительно (продукты)	Что можно предложить и продать дополнительно (услуги)	Как можно индивидуализировать и повысить цену и ценность
Дизайнеры кастомизированной одежды	Загрузка 3D-компонентов модели одежды под кастомизацию, продажа одежды по предзаказу, создание брендового магазина	Библиотека с 3D-аватарами	Создание 3D-моделей для загрузки в 3D-конфигуратор	Веб-аналитика, кастомизация интерфейса, продвижение бренда или новой коллекции одежды
Покупатели кастомизированной одежды	Создание кастомизированной 3D-модели одежды по своему вкусу, оформление и оплата заказа на ее изготовление	AR-примерка	Создание 3D-аватара по параметрам покупателя	Создание комплекта кастомизированных 3D-моделей одежды в стиле Family look

Помимо расширения функциональных и технологических возможностей цифровой платформы, представленных в табл. 2, планируется также масштабировать ее через привлечение дизайнеров кастомизированной обуви, сумок и аксессуаров. Разработанный 3D-конфигуратор может иметь более широкий спектр применения, например для проектирования одежды специального назначения (военной формы, защитной от неблагоприятных факторов). Он позволит создавать новые 3D-модели специальной одежды в

соответствии с эргономическими и функциональными требованиями, которые могут быть далее исследованы в виртуальной среде до изготовления образцов в материале.

СТЕК ТЕХНОЛОГИЙ РАЗРАБОТКИ ЦИФРОВОЙ ПЛАТФОРМЫ

Опишем технологический стек и целевую архитектуру мультибрендовой цифровой платформы со встроенным 3D-конфигуратором.

Программный продукт разрабатывается как веб-приложение с трехуровневой архитектурой:

1. Клиентская часть (Frontend): реализация осуществляется на React с фреймворком Next.js, использующим гибридный подход к рендерингу: SSR (Server-Side Rendering) для критически важных страниц и SSG (Static Site Generation) для статических страниц, с возможностью клиентской навигации по принципам SPA. При этом для бесшовной интеграции библиотеки с Three.js (основной библиотеки для 3D-конфигуратора) выбрана библиотека @react-three/fiber, а для работы с 3D-моделями в React-окружении – @react-three/drei.
2. Серверная часть (Backend): REST API на Node.js (TypeScript) с использованием фреймворка Nest.js. TypeScript обеспечивает строгую типизацию для DTO, Entities и интерфейсов, что значительно снижает количество runtime-ошибок в бизнес-логике.
3. Уровень данных: для хранения метаданных (пользователи, заказы, проекты) используется реляционная СУБД PostgreSQL; тяжелые файлы (3D-модели, текстуры, ассеты) хранятся в объектном хранилище (S3).

Все подсистемы цифровой платформы объединены с помощью прикладного программного интерфейса. Клиентское приложение взаимодействует с бэкендом через REST API. Бэкенд, в свою очередь, отвечает за бизнес-логику, аутентификацию и авторизацию пользователей, а также управляет данными в PostgreSQL и генерирует pre-signed URLs для безопасной загрузки и выгрузки ассетов из объектного хранилища. Статические файлы и 3D-модели доставляются через CDN.

Представим следующие базовые требования к целевым показателям надежности (Reliability) и отказоустойчивости цифровой платформы.

1. Работа в режиме 24/7.

2. RTO (Recovery Time Objective): целевая архитектура платформы спроектирована для обеспечения высокого уровня доступности с возможностью поэтапного улучшения показателей:

– целевые показатели для массовой эксплуатации включают RTO < 5 мин. при отказе отдельной ноды Kubernetes и RTO < 2 ч. при серьезных сбоях инфраструктуры за счет использования Infrastructure as Code (Terraform) и управляемых сервисов;

– на этапе развития в целевую архитектуру заложен фундамент для достижения RTO < 1 ч. при потере целой зоны доступности (дата-центра). Этот показатель может быть обеспечен переходом на multi-AZ конфигурации всех компонентов (Kubernetes, PostgreSQL) по мере роста требований к отказоустойчивости и повышения рентабельности стартап-проекта.

3. RPO (Recovery Point Objective) определяется причиной сбоя: реализованы механизмы резервного копирования для минимизации потерь данных. RPO для базы данных – не более 15 мин. (за счет транзакционных реплик WAL), для объектного хранилища – не более 24 ч. (за счет ночных снепшотов).

Реализация вышеперечисленных требований включает:

– регулярное бэкапирование (резервное копирование) данных и метаданных;

– раздельное хранение метаданных (СУБД) и бинарных данных (объектное хранилище) для повышения отказоустойчивости и упрощения масштабирования;

– оптимизация хранения данных для минимизации дублирования на уровне логики приложения, а также использование возможностей объектного хранилища для экономии места.

Представим базовые требования к целевым показателям безопасности (Security), которые реализуются в ходе разработки цифровой платформы. Командой разработчиков выбрана модель RBAC (Role-Based Access Control) со следующими ролями:

1) Admin: полные права, управление пользователями и их правами доступа;

2) Designer: права на CRUD-операции с 3D-компонентами и моделями в отведенном пространстве;

3) Customer: права на просмотр, использование 3D-конфигуратора и оформление заказов.

Безопасность обеспечивается за счет:

– идентификации и аутентификации;

– авторизации на основе ролей;

– логирования событий безопасности (Audit Log);

– контроля целостности данных и кодовой базы;

– процедур реагирования на инциденты информационной безопасности.

Развитие стартап-проекта в будущем возможно при условии масштабируемости и производительности (Scalability & Performance) цифровой платформы. Поэтому ее целевая архитектура спроектирована с учетом следующих требований.

1. Горизонтальное масштабирование, которое обеспечивается за счет:

– Stateless-архитектуры бэкенда (сессии вынесены во внешнее хранилище Redis);

– использования балансировщика нагрузки для распределения трафика между инстансами;

– настройки распределенного кэша на основе Redis (для сессий и данных).

2. Вертикальное масштабирование, которое обеспечивается за счет увеличения мощности отдельных серверов (CPU, RAM).

3. Гибкость: возможность поэтапного наращивания функционала и производительности.

4. Интеграция: API-first подход позволяет легко интегрироваться со сторонними системами (CRM, платежные шлюзы).

5. Мониторинг: в целевой архитектуре планируется использование стека Prometheus и Grafana для сбора метрик по приложению, базам данных и системным ресурсам, а также настройки алертов для критических инцидентов.

Реализация 3D-конфигуратора для цифровой платформы основана на следующем технологическом стеке и методологии:

1. Технология: WebGL через высокоуровневую библиотеку Three.js.

2. Форматы: основной формат для 3D-моделей – glTF (рекомендуемый), с поддержкой OBJ, FBX.

3. Конвейер обработки контента:

– создание и симуляция моделей одежды в CLO3D;

– экспорт и оптимизация моделей: дизайнеры вручную экспортируют модели из CLO3D в один из поддерживаемых форматов, затем Node.js-скрипт конвертирует его в оптимизированный glTF, выполняя сжатие текстур и удаление неиспользуемых данных для минимизации веса моделей. (На текущем этапе экспорт из CLO3D выполняется вручную, что является оптимальным для стартап-проекта с небольшим потоком 3D-моделей. Такое решение позволило сфокусировать ресурсы команды разработчиков на создании ядра цифровой платформы и 3D-конфигуратора. Однако оно является операционным и создает задержку обратной связи для дизайнера. В дальнейшем при росте нагрузки процесс планируется автоматизировать с помощью CLO Virtual Fashion SDK для создания прямого экспорта в glTF из среды дизайнера, что обеспечит бесшовный workflow и полный контроль над качеством 3D-контента на всех этапах);

– загрузка готовых ассетов в объектное хранилище платформы.

Отметим преимущества предлагаемого подхода:

– интерактивность: вращение моделей на 360°;

– адаптивность: корректное отображение на разных устройствах;

– производительность: оптимизированный вес моделей и использование аппаратного ускорения.

Инфраструктура разработки (DevOps) цифровой платформы со встроенным 3D-конфигуратором включает:

1) систему контроля версий Git;

2) GitLab CI/CD: используется для автоматизации пайплайна (сборка, запуск юнит- и интеграционных тестов, деплой в тестовые среды);

3) Figma: используется для дизайна интерфейса.

Выбранный стек технологий (JavaScript/TypeScript, React/Next.js, Node.js, Three.js) в сочетании с модульной трехуровневой архитектурой полностью соответствует требованиям стартапа: обеспечивает высокую производительность, простоту поддержки, удобство масштабирования и быструю разработку цифровой платформы силами небольшой команды.

ТРЕБОВАНИЯ К РАЗРАБОТКЕ ЦИФРОВЫХ МОДЕЛЕЙ ОДЕЖДЫ ДЛЯ 3D-КОНФИГУРАТОРА

Настоящие требования описывают процесс подготовки 3D-моделей кастомизированной одежды для загрузки на новую цифровую платформу, использующую технологию WebGL (библиотека Three.js). Соблюдение этих рекомендаций обеспечивает корректное отображение, высокую производительность и быструю загрузку моделей в веб-браузере.

Для понимания требований важно знать процесс обработки модели состоящий из трех этапов:

Этап 1. Создание модели: дизайнер создает и симулирует модель одежды в CLO3D (рекомендуется) или аналогах (Marvelous Designer, Style3D);

Этап 2. Экспорт и оптимизация: дизайнер вручную экспортирует модель из CLO3D в один из поддерживаемых форматов (FBX, OBJ). Далее Node.js-скрипт платформы автоматически конвертирует модель в оптимизированный формат glTF, выполняет сжатие текстур и удаляет неиспользуемые данные;

Этап 3. Публикация: готовый ассет загружается в 3D-конфигуратор.

Важно корректно подготовить модель одежды в CLO3D, чтобы автоматическая конвертация прошла успешно и дала наилучший результат.

Перечислим требования к полигональной сетке (Geometry): автоматическая оптимизация скриптом эффективнее работает с моделями, которые изначально не имеют избыточной детализации. Исходя из этого, рекомендуемое количество полигонов для одной 3D-модели составляет 10000–50000 треугольников. Соблюдение этого требования обеспечивает баланс между качеством и производительностью в веб-браузере.

Ключевой параметр в CLO3D – это настройка «Размер ячейки мэша». Он напрямую влияет на количество полигонов. Рекомендуется использовать значение 10–20 мм. Для наглядности представим на рис. 6 три варианта

настройки полигональной сети в программе для 3D-модели свитшота: 40, 20 и 5 мм.

а) б) в)

Рис. 6. Размер ячейки полигональной сети для цифровой модели одежды: а) 40 мм; б) 20 мм; в) 5 мм.

Размер ячейки 40 мм является максимально допустимым для корректного отображения складок и силуэта. Значения 50 мм и более приводят к сильным искажениям. При этом размер ячейки 5 мм создает избыточное количество полигонов, что значительно увеличивает размер файла и нагрузку на браузер без заметного визуального преимущества при отображении в веб-браузере.

Отметим требования к текстурам и материалам: скрипт автоматически сжимает текстуры, поэтому задача дизайнера – предоставить ему корректные исходные карты. Для достижения оптимального баланса между качеством и весом 3D-модели одежды рекомендуется использовать набор текстур с двумя основными картами:

– Diffuse (цвет/принт);

– Normal (симулирует рельеф).

Сохранять текстуры необходимо в форматах PNG (без потерь) или JPEG (с потерями). Исходное разрешение для загрузки — 1024 x 1024 px. Скрипт автоматически создаст оптимизированные версии (вплоть до 512 px) для различных устройств.

Все компоненты 3D-модели одежды должны иметь корректную UV-развертку. Для этого их необходимо расположить в пределах 0–1 UV-

пространства без наложений. Это критически важно для правильного наложения текстур (см. рис. 7).

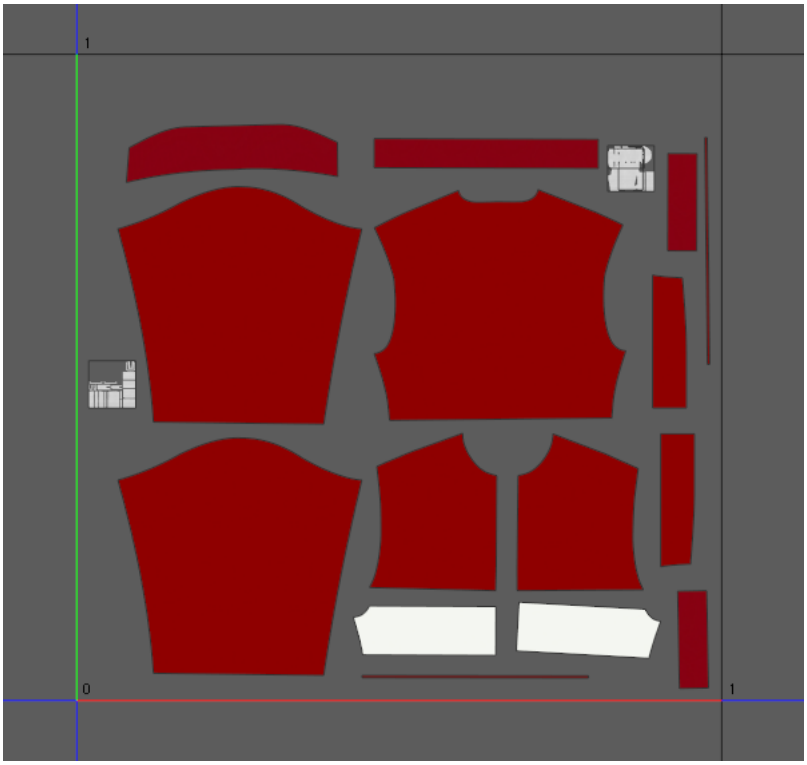


Рис. 7. UV-развертка текстур для 3D-модели одежды

Опишем процедуру экспорта и загрузки 3D-модели одежды на новую цифровую платформу:

1) готовая модель экспортируется из CLO3D в формате FBX (предпочтительно) или OBJ. При этом в настройках экспорта необходимо активировать опцию встраивания текстур (например, Embed Textures), чтобы все карты были включены в файл модели;

2) полученный файл модели (.fbx или .obj) загружается через интерфейс платформы, дизайнеру достаточно следовать подсказкам в интерфейсе загрузчика;

3) после загрузки файла запустится конвейер оптимизации: система автоматически сконвертирует модель в формат glTF и подготовит ее для визуализации в 3D-конфигураторе (см. рис. 8).

Рис. 8. Визуализация изменений цифровой модели платья в 3D-конфигураторе до и после загрузки рукавов и текстурных карт

Создание моделей одежды для 3D-конфигуратора требует от дизайнера понимания не только инструментов 3D-моделирования (CLO3D), но и основ веб-оптимизации. Соблюдение этих требований позволит обеспечить высокое качество визуализации и бесперебойную работу цифровой платформы для всех пользователей.

ИНТЕРФЕЙС ЦИФРОВОЙ ПЛАТФОРМЫ

Эффективное взаимодействие между пользователем и цифровой платформой очень важно. Оно обеспечивается посредством интуитивно-понятного интерфейса. Для его создания использовался специализированный инструмент Figma, который позволил реализовать следующие требования:

– минималистичный дизайн с продуманной информационной архитектурой, исключающей когнитивную перегрузку пользователей;

– интуитивная навигация, оптимизированная для быстрого доступа к целевым разделам;

– разграничение доступа к функционалу для 3D-дизайнера кастомизированной одежды и ее покупателя.

Для дизайна экранных форм, на которых производится кастомизация цифровых моделей одежды, разработана монохроматическая цветовая схема: фон белого и серого цветов обеспечивают концентрацию внимания

пользователей-покупателей на разных вариантах кастомизации, особенно на кастомизации с помощью цвета и принта (см. рис. 9).

Рис. 9. Основные цвета UI-дизайна цифровой платформы со встроенным 3D-конфигуратором

В качестве акцента в UI-дизайне цифровой платформы использован зеленый цвет (см. рис. 10).

Рис. 10. Акцентные цвета UI-дизайна цифровой платформы со встроенным 3D-конфигуратором

Для оформления типографики использованы два шрифта: Merge One и Afacad: Merge – для начертания логотипа цифровой платформы на английском языке, Afacad – для заголовков, кнопок и текста от 24 pt до 14 pt.

Представим часть компонентов из UI-kit, созданных для интерфейса цифровой платформы (рис. 11). UI-kit позволил существенно оптимизировать создание высокодетализированных макетов.

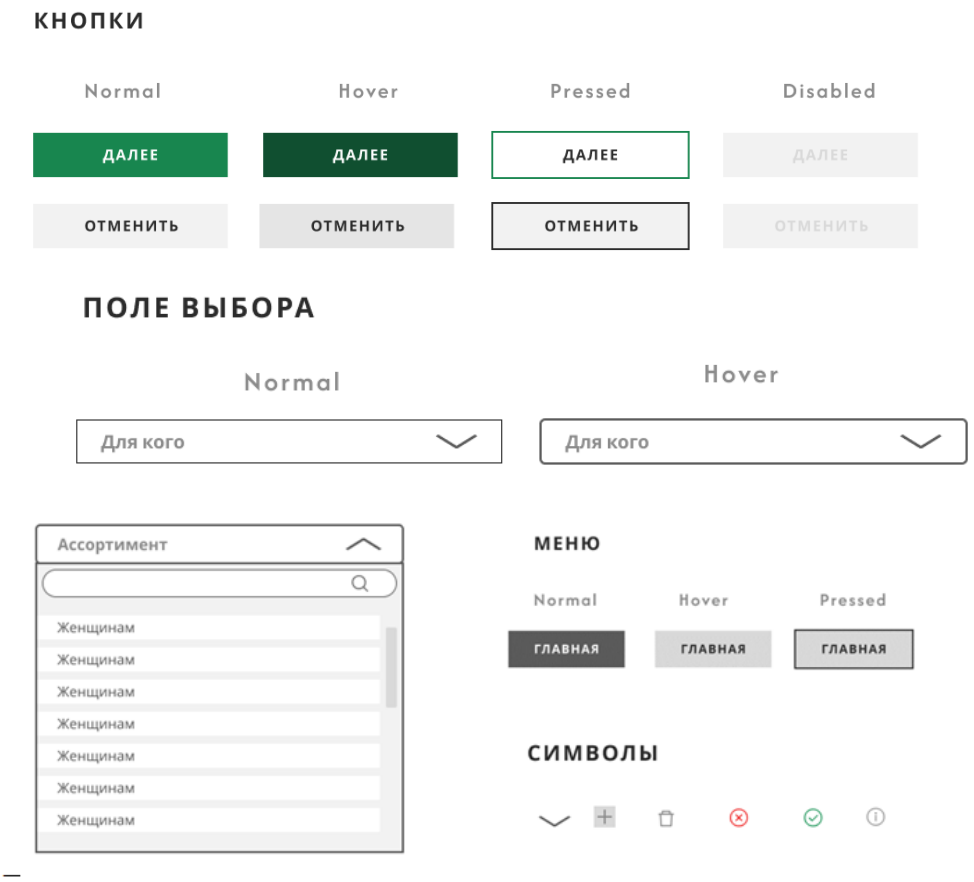


Рис. 11. Часть компонентов из UI-kit цифровой платформы со встроенным 3D-конфигуратором

Продемонстрируем также несколько высокодетализированных макетов экранных форм интерфейса цифровой платформы для различных категорий пользователей (см. рис. 12–14). В ходе их разработки учитывались принципы UI-дизайна, которые позволяют повысить эффективность представления 3D-контента (моделей одежды) как для пользователей-дизайнеров, так и для пользователей-покупателей:

1) для воздействия на пользователей-покупателей приоритет отдан цветовому решению: по сравнению с ним размеры элементов или объектов интерфейса не так важны;

2) обеим категориям пользователей на выбор предоставлено от трех до пяти вариантов элементов интерфейса, карточек товаров, категорий: слишком большой выбор тормозит выполнение задачи и снижает производительность со стороны пользователя;

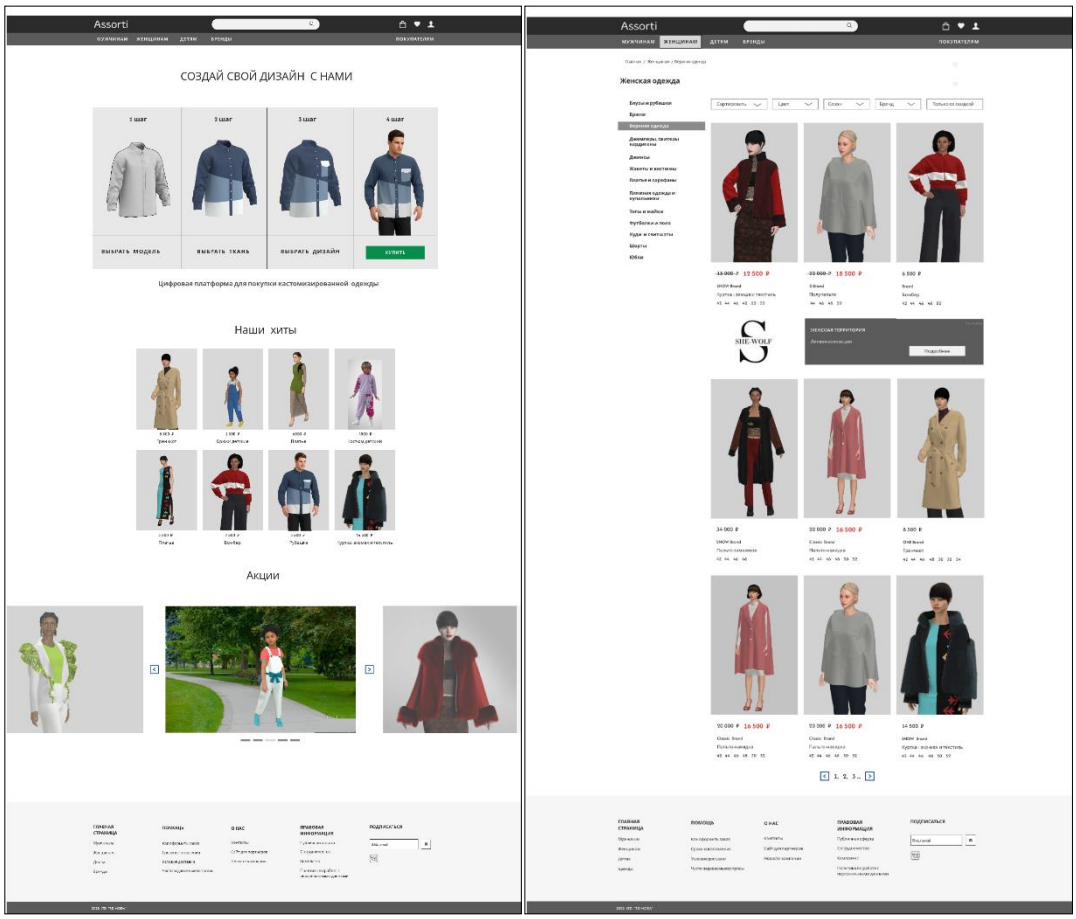


Рис. 12. Макеты главной страницы цифровой платформы и страницы с каталогом женской верхней одежды.

3) фон поддерживает основные элементы навигации и акценты, создавая глубину страниц, предназначенных для кастомизации и стимулирования сбыта моделей одежды;

4) меню интерфейса цифровой платформы расположены по F-форме: слева и сверху для лучшего восприятия различных блоков информации пользователями;

5) визуальное оформление интерфейса цифровой платформы планируется изменять в соответствие с сезонами, принятыми в модной индустрии.

Продемонстрируем также дифференциацию функциональных возможностей, доступных различным категориям пользователей (рис. 13 и 14).

Так, для пользователей-дизайнеров кастомизированной одежды разработаны три экранных формы, предназначенных для формирования

карточки под кастомизированную модель одежды. На первой экранной форме есть поля для ввода и выбора из готового перечня информации о модели одежды: ее категории, названия, артикула, цены, размеров и роста. На ней также можно ввести краткое описание модели одежды и загрузить таблицу размеров от дизайнера (производителя) по представленному образцу. Завершающим этапом работы на этой экранной форме являются просмотр и подтверждение финального варианта карточки под кастомизированную модель одежды, в котором она будет представлена пользователю-покупателю.

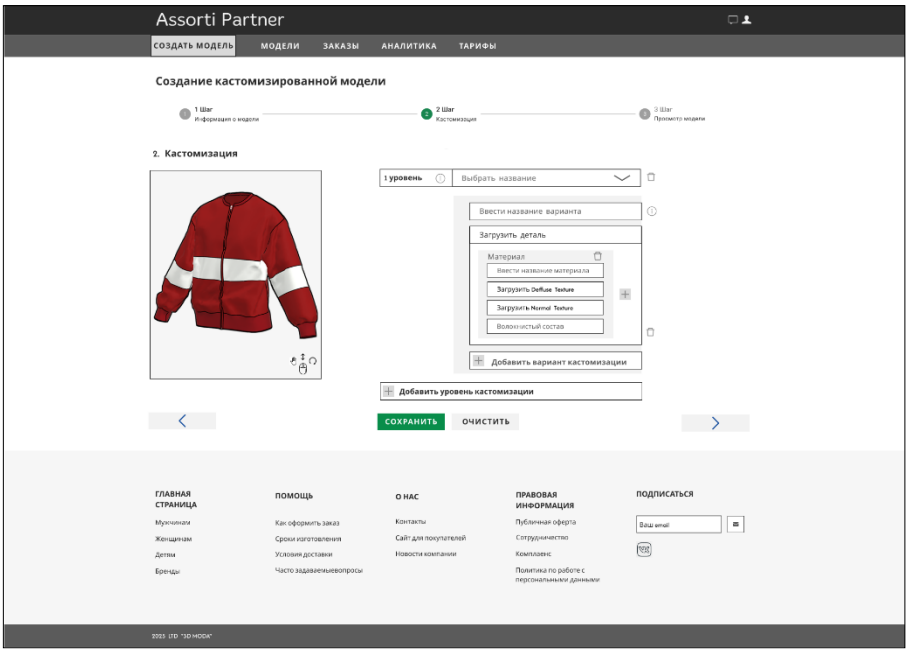


Рис. 13. Макет страницы цифровой платформы для загрузки компонентов 3D-модели одежды и настройки кастомизации пользователем-дизайнером.

Вторая экранная форма интерфейса позволяет реализовать технологию кастомизации цифровой модели одежды с помощью встроенного 3D-конфигуратора (см. рис. 13). Пользователь-дизайнер может выбрать уровень кастомизации, варианты кастомизации, а также загрузить все компоненты и текстуры модели одежды: например, для свитшота это может быть вариативное конструктивное оформление горловины или воротника, полочки или спинки, рукавов, манжетов, пуговиц и молний, печать рисунка или аппликации.

На экранной форме, представленной на рис. 13, также реализовано 3D-окно, отображающее результат всех действий пользователя-дизайнера с компонентами и текстурами 3D-модели.

На третьей экранной форме реализован просмотр всех возможных вариантов кастомизации модели одежды. При необходимости их редактирования пользователь-дизайнер может вернуться на предыдущий шаг. Если он убедился, что все выполнено корректно, он может опубликовать карточку своей модели одежды в соответствующий каталог на цифровой платформе.

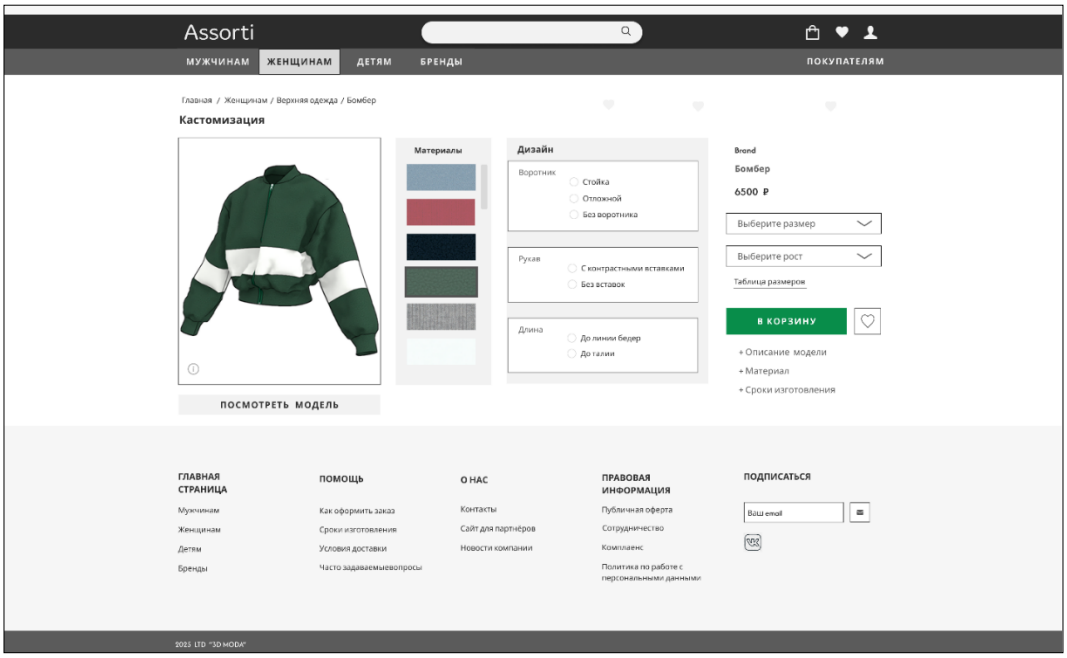


Рис. 14. Макет страницы цифровой платформы для создания кастомизированной модели одежды пользователем-покупателем

Для пользователей-покупателей разработана одна экранная форма, предназначенная для кастомизации моделей одежды. Она позволяет выполнить все действия по формированию заказа и сэкономить время пользователей (см. рис. 14), в то время как у аналогов для кастомизации используется от четырех до семи экранных форм и множества подгружаемых страниц, что увеличивает когнитивную нагрузку на пользователя и снижает конверсию.

После выбора ткани, вариантов дизайна и фурнитуры по своему вкусу пользователь-покупатель может посмотреть в 3D-окне кастомизированную им модель одежды со всех сторон, приблизить или отдалить ее, нажав на соответствующую кнопку.

Итак, интерфейс цифровой платформы реализован с учетом требований снижения когнитивной нагрузки на пользователей и повышения их продуктивности. Он содержит интуитивно-понятные элементы навигации, такие как кнопки целевых действий (CTA), выделенные акцентными цветами, кнопки перемещения между страницами «вперед» и «назад», выполненные размером более 45 px по ширине, и прочие удобные решения для реализации функциональных возможностей встроенного 3D-конфигуратора и взаимодействия между дизайнерами и покупателями кастомизированной одежды.

ЗАКЛЮЧЕНИЕ

Исследованы варианты цифровой трансформации индустрии модной одежды в России и мире, зафиксирован растущий спрос на массовую кастомизацию и изготовление одежды «по запросу» при одновременном отсутствии доступных технологий.

Для решения этой проблемы мы совместили современные веб-технологии с глубоким пониманием отраслевой специфики и предложили доступное для массового пользователя решение, которое имеет не только коммерческий потенциал, но и может оказать значительный трансформационный эффект на всю индустрию. Разрабатываемая цифровая платформа со встроенным 3D-конфигуратором представляет собой практическую реализацию концепции “co-creation” – совместного создания ценности производителем и потребителем.

В методологическом плане осуществлен комплексный анализ рынка по модели PAM-TAM-SAM-SOM. Это позволило валидировать экономические эффекты реализуемого стартап-проекта и обосновать оптимизацию бизнес-процессов для российских дизайнеров и компаний, специализирующихся на изготовлении кастомизированной одежды. Согласно модельным расчетам, внедрение цифровой платформы поможет снизить логистические издержки на 16–18% (прежде всего за счет сокращения возвратов) и уменьшить объемы технических остатков на 24–27%.

Особого внимания заслуживают целевые технологические параметры цифровой платформы, обеспечивающие ее масштабируемость,

производительность и отказоустойчивость. Представлен также реализованный конвейер обработки 3D-моделей, решающий критически важную для веб-среды задачу оптимизации полигональных сеток и текстур без потери визуального качества.

Перспективы технологического развития стартап-проекта мы видим в нескольких направлениях: это внедрение AR-примерки одежды для покупателей, интеграция технологий компьютерного зрения для создания предиктивных моделей спроса, а также разработка системы рекомендаций на основе машинного обучения. Реализация этих направлений развития сделает возможным создание многофункциональной и технологической экосистемы, позволяющей вовлечь не только ключевые, но и смежные сегменты модной индустрии.

Благодарности

Авторы выражают благодарность Федеральному государственному бюджетному учреждению «Фонд содействия развитию малых форм предприятий в научно-технической сфере» (Фонд содействия инновациям) за поддержку стартап-проекта по разработке цифровой платформы со встроенным 3D-конфигуратором, которая предназначена для цифровой трансформации российской индустрии моды.

СПИСОК ЛИТЕРАТУРЫ

1. The Opportunity in Digital Fashion and Avatars Report // BoF Insights. November 2021.

URL: https:/www.businessoff ashion.com/reports/technology/ the-opportunity-in-digital-fashion-and-avatars-report-bof-insights/ (дата обращения: 11.10.2025).

2. *Barata J., Cardoso J., Cunha P.* Mass customization and mass personalization meet at the crossroads of Industry 4.0: A case of augmented digital engineering // Systems Engineering. 2023. No. 26. P. 715–727. https://doi.org/10.1002/sys.21682

3. *Hou S., Gao J., Wang C.* Design for mass customisation, design for manufacturing, and design for supply chain: a review of the literature // IET Collab. Intell. Manuf. 2022. No 4(1). P. 1–16. https://doi.org/10.1049/cim2.12041

4. *Park H., Armstrong C.* Collaborative apparel consumption in the digital sharing economy: An agenda for academic inquiry // Int J Consum Stud. 2017. No. 41. P. 465–474. https://doi.org/10.1111/ijcs.12354

5. *Longo F., Padovano A., Cimmino B., Pinto P.* Towards a mass customization in the fashion industry: An evolutionary decision aid model for apparel product platform design and optimization // Computers & Industrial Engineering. 2021. Vol. 162, 107742. https://doi.org/10.1016/j.cie.2021.107742

6. *Евдущенко Е.В., Ковалева Ю.В.* Концепция проектирования женских платьев на фигуры смежных размеров для увеличения онлайн-продаж // Костюмология. 2020. № 2. Том 5. URL: https://kostumologiya.ru/PDF/12TLKL220.pdf (дата обращения: 19.10.2025).

7. *Qiu Y., Duan H., Xie H., Ding X., Jiao Y.* Design and development of a web-based interactive twin platform for watershed management // Transactions in GIS. 2022. No. 26. P. 1299–1317. https://doi.org/10.1111/tgis.12904

8. *Пудовкина О.Е.* Формирование цифровой экосистемы промышленной кооперации на базе передовых цифровых платформ в условиях реиндустрилизации // Вестник ГУУ. 2020. №9.
https://doi.org/10.26425/1816-4277-2020-9-41-48

9. Сам себе дизайнер: кастомизация кроссовок Louis Vuitton. Тенденции. AR World Luxury Guide // Annarusska.
URL: https://www.annarusska.ru/fashion/trends/sam-sebe-dizayner-kastomizaciya-krossovok-louis-vuitton/ (дата обращения 26.10.2025).

10. *Hwang C., Feng J.* Using 3D virtual fitting room stimuli to enhance older adults' spatial visualization skills // Family and Consumer Sciences Research Journal. 2023. No. 52. P. 5–18. https://doi.org/10.1111/fcsr.12486

11. *Lee H., Xu Y., Porterfield A.* Virtual Fitting Rooms for Online Apparel Shopping: An Exploration of Consumer Perceptions // Family and Consumer Sciences Research Journal. 2022. No. 50. P. 189–204.
https://doi.org/10.1111/fcsr.12428

12. *Galizia F., Bortolini M., Calabrese F.* A cross-sectorial review of industrial best practices and case histories on Industry 4.0 technologies // Systems Engineering. 2023. Vol. 2, No. 6. P. 908–924. https://doi.org/10.1002/sys.21697

13. *Давыденко Е.А., Григорян А.В.* Особенности выстраивания коммуникаций российскими брендами одежды в контексте импортозамещения // Маркетинговые коммуникации. 2023. №1. С. 2–9. https://doi.org/10.36627/2619-1407-2023-1-1-2-9

14. *Dey U., Cheruvu S.* A web-based integrated GUI for 3D modeling, kinematic study, and control of robotic manipulators // Comput Appl Eng Educ. 2020. No. 28. P. 1028–1040. https://doi.org/10.1002/cae.22282

15. *Сахарова Н.А.* Цифровые технологии в дизайне и конструировании одежды // Инновации в текстиле, одежде, обуви (ICTAI-2022). Материалы докладов международной научно-технической конференции, Витебск, Республика Беларусь, 23–24 ноября 2022 года. Витебский государственный технологический университет: 2022. С. 75–79. URL: https://www.researchgate.net/publication/376596872_Cifrovye_tehnologii_v_dizajne_i_konstruirovanii_odezdy (дата обращения 29.10.2025).

16. *Сурай Н.М., Теплая Н.А., Баскаков В.А., Бурланков П.С., Пислегина Н.В.* Маркетплейсы как драйвер развития электронной коммерции // Инновации и инвестиции. 2023. № 5. С. 154–157. URL: https://cyberleninka.ru/article/n/marketpleysy-kak-drayver-razvitiya-elektronnoy-kommertsii/viewer (дата обращения 31.10.2025).

17. Программа «Студенческий стартап»: официальный сайт Фонда содействия инновациям. URL: https://www.fasie.ru/programs/programma-studstartup/ (дата обращения 15.10.2025).

18. . Свидетельство о государственной регистрации программы для ЭВМ № 2025685040 Российская Федерация. Программа – "3D конфигуратор для кастомизации одежды": заявл. 15.08.2025: опубл. 19.09.2025 / Е. В. Евдущенко, В. В. Козлов; заявитель Общество с ограниченной ответственностью "3Д МОДА". URL: https://www.elibrary.ru/item.asp?id=82913375

19. *Ларин С.Н., Нуждин М.Г., Даниелян А.С., Хунузиди Е.И., Нуждин Г.А.* Структурирование функции качества и модель Кано // Известия Тульского государственного университета. Технические науки. 2025. № 1. С. 263-271.

https://doi.org/10.24412/2071-6168-2025-1-263-264.
URL: https://cyberleninka.ru/article/n/strukturirovanie-funktsii-kachestva-i-model-kano (дата обращения: 15.10.2025)

DEVELOPMENT OF A DIGITAL PLATFORM WITH AN INTEGRATED 3D CONFIGURATOR FOR CLOTHING CUSTOMIZATION

E. V. Evdushenko[1] [0000-0003-3692-2587], M. V. Shmatko[2] [0000-0002-7255-8885]

[1]*Military Engineering Institute of the Military Logistics Academy, Saint Petersburg, Russia*

[2]*Omsk State Technical University, Omsk, Russia*

[1]elena.online_ktilp@mail.ru, [2]marin298@gmail.com

Abstract

Amidst the rapid growth of e-commerce and increasing demand for personalization, the Russian market for customized clothing faces a shortage of technological and widely accessible solutions. This paper presents the results of a research and implementation project focused on developing a multi-brand digital platform with an integrated 3D configurator, aimed at transforming the pre-order cycle. The solution enables customers to interactively create garment designs in a web environment, while allowing designers to optimize logistics and minimize overproduction.

The primary scientific and technical contribution of this work lies in its detailed description of the platform's target architecture and a scalable 3D model processing pipeline that ensures model optimization and correct browser-based rendering. An additional contribution is the developed methodology for preparing and optimizing 3D garment models for web visualization. Formalized as a set of technical requirements, this methodology achieves a balance between visual quality and performance.

As a result of this research, the authors have addressed the challenge of unifying 3D model formats from different designers within a multi-brand digital platform—a key distinction from existing single-brand solutions. Furthermore, the

implemented technology enables the customization of 3D clothing models with interactive real-time visualization of all design modifications on a single screen.

The technological feasibility and effectiveness of the solution are substantiated by a comparative analysis of existing alternatives, a market analysis using the PAM-TAM-SAM-SOM model, and an assessment of functional requirements.

The article also outlines a practical strategy for implementing the digital platform, making it a valuable resource for researchers and practitioners working at the intersection of e-commerce, computer graphics, and the digital transformation of business processes.

Keywords: *digital transformation, web application, digital platform, 3D configurator, 3D model, clothing customization, virtual try-on, AR fitting, technology stack, architecture, scalability, performance.*

REFERENCES

1. The Opportunity in Digital Fashion and Avatars Report // BoF Insights. November 2021.
URL: https:/www.businessoff ashion.com/reports/technology/ the-opportunity-in-digital-fashion-and-avatars-report-bof-insights/

2. *Barata J., Cardoso J., Cunha P.* Mass customization and mass personalization meet at the crossroads of Industry 4.0: A case of augmented digital engineering // Systems Engineering. 2023. No. 26. P. 715–727. https://doi.org/10.1002/sys.21682

3. *Hou S., Gao J., Wang C.* Design for mass customisation, design for manufacturing, and design for supply chain: a review of the literature // IET Collab. Intell. Manuf. 2022. No. 4(1). P. 1–16. https://doi.org/10.1049/cim2.12041

4. *Park H., Armstrong C.* Collaborative apparel consumption in the digital sharing economy: An agenda for academic inquiry // Int J Consum Stud. 2017. No. 41. P. 465–474. https://doi.org/10.1111/ijcs.12354

5. *Longo F., Padovano A., Cimmino B., Pinto P.* Towards a mass customization in the fashion industry: An evolutionary decision aid model for apparel product platform design and optimization // Computers & Industrial Engineering. 2021. Vol. 162, 107742. https://doi.org/10.1016/j.cie.2021.107742

6. . *Evdushchenko E.V., Kovaleva Yu.V.* Kontseptsiya proektirovaniya zhenskikh plat'ev na figury smezhnykh razmerov dlya uvelicheniya onlain-prodazh // Kostiumologiia. 2020. Vol. 5, No. 2.
URL: https://kostumologiya.ru/PDF/12TLKL220.pdf

7. *Qiu Y., Duan H., Xie H., Ding X., Jiao Y.* Design and development of a web-based interactive twin platform for watershed management // Transactions in GIS. 2022. No. 26. P. 1299–1317. https://doi.org/10.1111/tgis.12904

8. *Pudovkina O.E.* Formirovanie tsifrovoi ekosistemy promyshlennoi kooperatsii na baze peredovykh tsifrovykh platform v usloviiakh reindustrializatsii // Vestnik GUU. 2020. No. 9.
https://doi.org/10.26425/1816-4277-2020-9-41-48

9. Sam sebe dizainer: kastomizatsiia krossovok Louis Vuitton. Tendentsii. AR World Luxury Guide // Annarusska.
URL: https://www.annarusska.ru/fashion/trends/sam-sebe-dizayner-kastomizaciya-krossovok-louis-vuitton/

10. *Hwang C., Feng J.* Using 3D virtual fitting room stimuli to enhance older adults' spatial visualization skills // Family and Consumer Sciences Research Journal. 2023. No. 52. P. 5–18. https://doi.org/10.1111/fcsr.12486

11. *Lee H., Xu Y., Porterfield A.* Virtual Fitting Rooms for Online Apparel Shopping: An Exploration of Consumer Perceptions // Family and Consumer Sciences Research Journal. 2022. No. 50. P. 189–204.
https://doi.org/10.1111/fcsr.12428

12. *Galizia F., Bortolini M., Calabrese F.* A cross-sectorial review of industrial best practices and case histories on Industry 4.0 technologies // Systems Engineering. 2023. Vol. 2, No. 6. P. 908–924. https://doi.org/10.1002/sys.21697

13. *Davydenko E.A., Grigorian A.V.* Osobennosti vystraivaniia kommunikatsii rossiiskimi brendami odezhdy v kontekste importozameshcheniia // Marketingovye kommunikatsii. 2023. №1. S. 2–9.
https://doi.org/10.36627/2619-1407-2023-1-1-2-9

14. *Dey U., Cheruvu S.* A web-based integrated GUI for 3D modeling, kinematic study, and control of robotic manipulators // Comput Appl Eng Educ. 2020. No. 28. P. 1028–1040. https://doi.org/10.1002/cae.22282

15. *Sakharova N.A.* Tsifrovye tekhnologii v dizaine i konstruirovanii odezhdy // Innovatsii v tekstile, odezhde, obuvi (ICTAI-2022). Materialy` dokladov mezhdunarodnoj nauchno-texnicheskoj konferencii, Vitebsk, Respublika Belarus`, 23–24 noyabrya 2022 goda. Vitebskij gosudarstvenny`j texnologicheskij universitet: 2022. S. 75–79.
URL:
https://www.researchgate.net/publication/376596872_Cifrovye_tehnologii_v_dizajne_i_konstruirovanii_odezdy
(дата обращения 29.10.2025)

16. *Surai N.M., Teplyaia N.A., Baskakov V.A., Burlankov P.S., Pislegina N.V.* Marketpleisy kak draiver razvitiia elektronnoi kommertsii // Innovatsii i investitsii. 2023. No. 5. S. 154–157. URL: https://cyberleninka.ru/article/n/marketpleysy-kak-drayver-razvitiya-elektronnoy-kommertsii/viewer (дата обращения 31.10.2025).

17. Programma «Studencheskii startap»: ofitsial'nyi sait Fonda sodeistviia innovatsiiam. URL: https://www.fasie.ru/programs/programma-studstartup/

18. Svidetel`stvo o gosudarstvennoj registracii programmy` dlya E`VM № 2025685040 Rossijskaya Federaciya. 3D-konfigurator: № 2025683832: zayavl. 15.08.2025: opubl. 19.09.2025 / E.V. Evdushhenko, V.V. Kozlov; zayavitel` Obshhestvo s ogranichennoj otvetstvennost`yu «3D MODA».
URL: https://www.elibrary.ru/item.asp?id=82913375

19. *Larin S.N., Nuzhdin M.G., Danielyan A.S., Xunuzidi E.I., Nuzhdin G.A.* Strukturirovanie funkcii kachestva i model` Kano // Izvestiya Tul`skogo gosudarstvennogo universiteta. Texnicheskie nauki. 2025. No. 1. S. 263–271.
https://doi.org/10.24412/2071-6168-2025-1-263-264

СВЕДЕНИЯ ОБ АВТОРАХ

ЕВДУЩЕНКО Елена Владимировна – к. техн. н., доцент кафедры «Военная архитектура, автоматизированные системы проектирования, естественнонаучные дисциплины» ВИ (ИТ) ВА МТО, г. Санкт-Петербург. Сфера научных интересов – 3D-моделирование одежды, развитие ИТ-продуктов.

Elena Vladimirovna EVDUSHCHENKO – Candidate of of Technical Sciences, Associate Professor at the Department of Military Architecture, Automated Design Systems and Natural Sciences of Military Engineering Institute, Military Logistics Academy, St. Petersburg. Research interests include 3D modeling of clothing and the development of IT products.

email: elena.online_ktilp@mail.ru

ORCID: 0000-0003-3692-2587

ШМАТКО Марианна Владимировна – к. филос.н, доцент кафедры «Математические методы и информационные технологии в экономике» ОмГТУ. Сфера научных интересов – разработка новых цифровых продуктов, UX-аналитика и юзабилити информационных систем, гейм-дизайн. Сфера практической деятельности — цифровой маркетинг.

Marianna Vladimirovna SHMATKO – Candidate of Philosophical Sciences, Associate Professor at the Department of Mathematical Methods and Information Technologies in Economics, Omsk State Technical University. Research interests: digital product development, UX analytics, information systems usability, and game design. Practical activities: digital marketing, development of technology startup projects.

email: marin298@gmail.com

ORCID: 0000-0002-7255-8885

Материал проступил в редакцию 17 декабря 2025 года

УДК 004.8+004.912

ТИПЫ ЭМБЕДДИНГОВ И ИХ ПРИМЕНЕНИЕ В ИНТЕЛЛЕКТУАЛЬНОЙ АКАДЕМИЧЕСКОЙ ГЕНЕАЛОГИИ

А. Х. Мариносян[0000-0003-0577-2360]

Московский городской педагогический университет, г. Москва, Россия

a.marinosyan@yandex.ru

Аннотация

Рассмотрена проблема построения интерпретируемых векторных представлений научных текстов для задач интеллектуальной академической генеалогии. Предложена типология эмбеддингов, включающая три класса: статистические, выученные нейросетевые и структурированные символьные. Обоснована необходимость объединения достоинств нейросетевых (высокая семантическая точность) и символьных (интерпретируемость измерений) подходов. Для реализации такого гибридного подхода предложен алгоритм построения выученных символьных эмбеддингов путем регрессионного преобразования вектора внутреннего представления нейросетевой модели в интерпретируемый набор оценок.

Экспериментальная оценка алгоритма проведена на корпусе фрагментов авторефератов диссертаций по педагогическим наукам. Компактный трансформерный энкодер с регрессионной головой обучался воспроизводить тематические оценки, сгенерированные передовой генеративной языковой моделью. Сравнение шести режимов обучения (три типа регрессионной головы и два состояния энкодера) показало, что дообучение верхних слоев энкодера является ключевым фактором повышения качества. По результатам тестирования была выбрана наилучшая конфигурация, которая достигла коэффициента детерминации $R^2 = 0.57$ и точности определения трех наиболее релевантных концептов, равной 74%. Результаты подтверждают, что для определенного рода задач, в которых требуется формальное представление выходных данных, возможна аппроксимация поведения генеративной модели компактным энкоде-

ром с регрессионной головой при существенно меньших вычислительных затратах. В более широкой перспективе разработка алгоритмов построения выученных символьных эмбеддингов будет способствовать созданию такой модели формальной репрезентации научного знания, в которой конвергенция нейросетевых и символьных методов обеспечит как масштабируемость обработки научных текстов, так и интерпретируемость векторных представлений, кодирующих содержание.

Ключевые слова: *эмбеддинги, академическая генеалогия, трансформерный энкодер, регрессионная голова, символьные эмбеддинги, тематический профиль, обработка естественного языка, интерпретируемость, большие языковые модели, наукометрия.*

ВВЕДЕНИЕ

Академическая генеалогия – это междисциплинарная область исследований, изучающая структуру, динамику и эволюцию науки в контексте отношений научного руководства. Традиционно она опирается на формальные сведения о диссертациях (данные об авторе и руководителе, месте и годе защиты и т. д.) [1, 2]. Однако формальная связь «научный руководитель – ученик» не тождественна содержательной близости: ученик может существенно отклониться от тематики руководителя, тогда как исследователи, не связанные формально, нередко работают в едином проблемном поле. Переход от академической генеалогии формальных связей к интеллектуальной академической генеалогии, основанной на автоматизированном анализе содержания научных работ [3, 4], в своем логическом завершении предполагает построение эмбеддинг-пространства научного знания [5]. Под таким пространством понимается многомерная структура, в которой элементы знания представлены векторами, положение и свойства которых отражают их содержание, а топологические характеристики самого пространства несут информацию о структуре предметной области. К такому пространству могут быть применены математические методы для анализа динамики развития знания – от выявления концептуальной преемственности работ до отслеживания эволюции научных школ.

Существующие подходы к построению эмбеддингов можно разделить на три основных типа, различающихся по способу получения и степени интерпретируемости.

1. *Статистические эмбеддинги* основаны на алгоритмах и методах математической статистики: TF-IDF [6, 7], матрицы совместной встречаемости слов (co-occurrence), модели «мешка слов» (bag-of-words). Их достоинство состоит в простоте реализации и интерпретации: каждое измерение вектора соответствует конкретному слову, что позволяет выявлять устойчивые группы часто встречающихся терминов, формирующих лексическое ядро текста. Ограничения, однако, существенны: эти методы игнорируют семантические связи между словами (омонимия, синонимия, полисемия), не учитывают порядок слов и контекст, а потому неспособны уловить смысловую близость текстов, различающихся лексикой. Кроме того, порождаемые ими разреженные высокоразмерные матрицы плохо масштабируются при росте словаря.

2. *Выученные*[1] (*learned*) *нейросетевые эмбеддинги* получаются с помощью моделей, обученных на больших корпусах текстов. К ранним подходам относятся word2vec [8] и GloVe [9], порождающие статические векторы слов. Контекстуализированные модели: BERT [10], Sentence-BERT [11], SciBERT [12], E5 [13] – генерируют плотные векторы (размерностью от нескольких сотен до нескольких тысяч измерений), учитывающие контекст каждого слова в предложении. Достоинствами этих моделей являются высокая точность определения смысловой близости текстов, способность учитывать синонимию, полисемию, омонимию и ассоциативные связи, а также возможность работы с

[1]Поясним выбор терминологии. Нейросетевые эмбеддинги второго типа нередко называют «семантическими» (semantic embeddings). Однако в статье отдано предпочтению термину *«выученные»* (learned). Определение «семантический» содержательно перегружено и не формулируемо строгим образом. С одной стороны, в широком смысле все три типа эмбеддингов в той или иной мере кодируют семантику – статистические через лексические паттерны, нейросетевые через дистрибутивную гипотезу, символьные через экспертное определение шкал. С другой стороны, использование термина «семантический» несет в себе философско-филологическую нагрузку и ставит вопрос о том, что есть семантика. Термин же «выученный» точно указывает на ключевое различительное (и легко выявляемое) свойство – способ получения: эти представления *выучены* моделью из данных в ходе обучения на большом корпусе, а не сконструированы экспертом и не получены прямым подсчетом статистик.

нечеткими запросами и текстами на разных языках. Ключевое ограничение – это «непрозрачность» (black box): измерения векторного пространства не имеют явного символьного значения, что затрудняет интерпретацию и вызывает обоснованные сомнения в применимости таких представлений в областях, требующих прозрачности, в частности в науке.

3. ***Структурированные символьные эмбеддинги*** – это такие представления, в которых каждое измерение вектора соответствует заранее определенной экспертной шкале, а значения задаются на интерпретируемой числовой шкале (например, от 0 до 10). В контексте анализа научных работ такой вектор может описывать *тематический профиль* исследования – формальное представление тематики работы, заданное иерархическим классификатором [14]. Был предложен алгоритм формирования тематических профилей диссертаций с помощью большой языковой модели (БЯМ): для каждой работы на основании иерархического классификатора БЯМ присваивает числовые оценки релевантности по каждому элементу классификатора, формируя тем самым *тематический вектор* – символьный эмбеддинг с интерпретируемыми координатами [14]. Следует отметить, что в общем случае символьные эмбеддинги могут отражать не только тематический профиль, но и более сложное формальное представление работы: например, результаты исследования, закодированные по правилам специализированной онтологии предметной области [15].

Вместе с тем опыт генерации символьных эмбеддингов напрямую с помощью БЯМ выявил ряд проблем. Процесс содержит высокую долю произвольности: результат зависит от формулировки промпта, версии модели, температуры и других факторов; воспроизводимость ограничена, а систематический контроль качества затруднен. Поэтому необходим инструментарий, который позволил бы получать символьные эмбеддинги более контролируемо и эффективно.

Таким образом, перед нами стоит следующая проблема: каким алгоритмическим образом можно создать формальное векторное представление научной работы, которое одновременно обладало бы высокой семантической точностью (свойством выученных нейросетевых эмбеддингов) и интерпретируемостью координат (свойством символьных эмбеддингов), чтобы затем, применив различные метрики, выявить и оценить концептуальную преемствен-

ность научных работ. Цель настоящей статьи состоит не в полном решении этой проблемы, а в разработке и экспериментальной проверке подхода, являющегося компонентом ее решения: алгоритма построения ***выученных символьных эмбеддингов*** (learned symbolic embeddings), сочетающих достоинства типов представлений, описанных выше.

1. АЛГОРИТМ ПОСТРОЕНИЯ ВЫУЧЕННЫХ СИМВОЛЬНЫХ ЭМБЕДДИНГОВ

1.1. Обоснование

В основе современных подходов к обработке естественного языка лежит архитектура трансформера [16], преобразующая входную последовательность токенов в высокоразмерные векторные представления (эмбеддинги) в латентном пространстве. В классическом варианте трансформерной архитектуры, реализованном в большинстве БЯМ, механизм «выхода» из латентного пространства осуществляется через проекционный слой, который отображает латентный вектор в вектор логитов размерности словаря, после этого для получения вероятностного распределения следующего токена применяется функция Softmax [16].

Применительно к задачам извлечения формализованных знаний (например, оценки релевантности набора концептов) использование стандартной генеративной парадигмы сопряжено с рядом фундаментальных ограничений. Получение тематического вектора, описанного выше, происходит посредством генерации текста (например, в формате JSON), и процесс принимает вид двойной трансформации: сначала потенциально богатая семантика латентного вектора «сжимается» до одного дискретного токена естественного языка; затем сгенерированная последовательность токенов должна быть разделена на составные части, структурирована и преобразована обратно в числовой вектор.

Такой подход неэффективен с позиции генерации (инференса), поскольку требует авторегрессионной генерации последовательности токенов для описания структуры, которую можно выразить одним вектором. Кроме того, он неэффективен с позиции обучения: латентный вектор в принципе способен кодировать многомерные семантические конструкции, однако при авторегрессионной генерации весь обучающий сигнал на каждом шаге проходит через «узкое горлышко» (bottleneck) предсказания единственного токена, что ограничивает

объем структурированной информации, который модель может извлечь из обучающих данных за один шаг обратного распространения.

Альтернативным и потенциально более эффективным подходом для решения задачи создания формальной репрезентации текста является прямой выход из латентного слоя в целевое формальное пространство, минуя этап генерации естественного языка. Это можно достигнуть путем замены стандартного проекционного слоя на регрессионную голову (regression head) – выходной модуль, который с использованием регрессии преобразует вектор внутреннего (латентного) состояния в непрерывные числовые значения целевых признаков [17] (в нашем случае – в компоненты тематического вектора). Предлагаемый нами подход позволяет осуществлять предсказание всего вектора признаков за один проход (forward pass), существенно снижает вычислительные затраты и предоставлет более широкий инструментарий по контролю качества генерации.

Технически это может быть реализовано на базе трансформерного энкодера – архитектурного блока, использующего механизмы двунаправленного внимания (self-attention) для формирования целостного, контекстно-зависимого векторного представления входного текста [16]. Использование трансформерных энкодеров (самая известная их архитектура – BERT [10]) является стандартом в машинном обучении для задач извлечения признаков. Эффективность решения, когда регрессионная голова надстраивается над выходом энкодера, подтверждается рядом исследований. Например, BERT с регрессионной головой использовался для генерации оценок качества машинного перевода [17]. Мультиязычный энкодер mT0 с дополнительным предобучением на задачу различения близких и далеких пар предложений был дополнен регрессионной головой, преобразующей выходной семантический вектор напрямую в числовую оценку качества перевода [18]. Ансамбль энкодеров с регрессионным выходом применялся для предсказания сложности экзаменационных заданий [19].

1.2. Архитектуры регрессионной головы

Трансформерный энкодер обрабатывает входной текст потокенно: каждому токену (слову или его части) на выходе соответствует отдельный вектор. Таким образом, для текста из *n* токенов энкодер порождает *n* векторов раз-

мерности *d*. Для получения единого вектора предложения применено усреднение (mean pooling) этих *n* векторов. Поскольку тексты в наборе данных дополняются до одинаковой длины служебными «пустыми» токенами, усреднение произведено только по содержательным токенам (служебные исключаются маской). Полученный вектор $\mathbf{e} \in \mathbb{R}^d$ подается на вход регрессионной головы, которая отображает его в вектор предсказанных оценок $\hat{\mathbf{y}} \in \mathbb{R}^k$, где *k* – число компонент этого вектора (в случае академической генеалогии $\hat{\mathbf{y}}$ – это тематический вектор научной работы).

В алгоритме предусмотрены два варианта реализации регрессионной головы: посредством линейной регрессии и многослойного перцептрона (multilayer perceptron, MLP). Данные, полученные с использованием этих двух вариантов, будут сравниваться друг с другом на этапе интерпретации результатов.

Линейная регрессионная голова осуществляет аффинное преобразование

$$\hat{y} = We + b,$$

где $\mathbf{W} \in \mathbb{R}^{k \times d}$ – обучаемая матрица весов, каждая строка которой задает «направление» в пространстве эмбеддингов, соответствующее одной компоненте выходного вектора; $\mathbf{b} \in \mathbb{R}^k$ – обучаемый вектор сдвига (bias). Число обучаемых параметров составляет $k \times d + k$, что обеспечивает устойчивость к переобучению при ограниченных объемах данных. Линейные пробы над предобученными энкодерами являются стандартным инструментом анализа латентных представлений в естественной обработке языка [20].

MLP-регрессионная голова в рамках исследования была реализована как двухслойный перцептрон с нелинейной активацией:

$$\hat{y} = W_2 \cdot Dropout(ReLU(W_1 e + b_1), p) + b_2,$$

где $\mathbf{W}_1 \in \mathbb{R}^{h \times d}$ и $\mathbf{W}_2 \in \mathbb{R}^{k \times h}$ – обучаемые матрицы весов первого и второго слоев соответственно, $\mathbf{b}_1 \in \mathbb{R}^h$ и $\mathbf{b}_2 \in \mathbb{R}^k$ – обучаемые сдвиги, *h* – размерность промежуточного скрытого слоя; ReLU – нелинейная функция активации, обнуляющая отрицательные значения; Dropout – механизм регуляризации, при котором случайная доля *p* нейронов обнуляется на каждом шаге обучения, что препятствует «соадаптации» нейронов и снижает переобучение [21]. С одной сторо-

ны, перцептрон MLP способен моделировать нелинейные зависимости между вектором энкодера и целевыми оценками. С другой стороны, если MLP-голова показывает существенно лучшие значения, чем линейный регрессор, то это означает, что сам вектор энкодера плохо уловил «семантику» исходного текста и «семантическая» связь между исходным текстом и оценками достраивается самим MLP.

1.3. Процесс обучения

Тренировочные данные разбиваются на тренировочную, валидационную и тестовую подвыборки в пропорции 70% – 15% – 15%. Целевые оценки $\mathbf{y}_i \in \mathbb{R}^k$ (или, используя терминологию машинного обучения, оценки, рассматриваемые как эталонные данные (ground truth)) могут формироваться как экспертным образом, так и путем использования передовой БЯМ, которая считается достаточно квалифицированной для данной задачи (то есть разрыв в способностях между передовой (state-of-the-art) БЯМ и небольшим используемым энкодером настолько большой, что оценки, проставленные БЯМ, могут приниматься за эталон (ground truth) для данной задачи).

Для определения влияния различных архитектурных решений предусмотрены шесть режимов обучения, образованных комбинацией по двум факторам.

Первый фактор – тип обучения регрессионной головы:

- *линейная голова с нахождением матрицы весов алгоритмическим путем.* Использована регуляризация Тихонова, известная в англоязычной литературе как гребневая регрессия (ridge regression) [22][2];
- *линейная голова с нахождением матрицы весов путем машинного обучения.* Параметры выучиваются итеративно посредством градиентного спуска[3];

[2]В исследовании для реализации алгоритма гребневой регрессии использована функция Ridge python-библиотеки sklearn. URL: https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression

[3]Для построения линейного преобразования использован класс torch.nn.Linear библиотеки PyTorch. URL: https://docs.pytorch.org/docs/stable/generated/torch.nn.Linear.html

- *MLP-регрессионная голова.* Архитектура была описана выше[4].

Вторым фактором является состояние энкодера:

- *замороженный энкодер.* Все параметры предобученного трансформерного энкодера фиксируются и не обучаются;
- *частично дообученный (fine-tuned) энкодер.* Основная часть параметров энкодера остается замороженной, но последние *N* слоев трансформера размораживаются и обучаются совместно с регрессионной головой.

Комбинация двух факторов (три типа обучения головы и два состояния энкодера) образует шесть режимов, что позволяет изолировать вклад каждого фактора: сравнение при фиксированном типе головы показывает эффект дообучения энкодера, сравнение при фиксированном состоянии энкодера – эффект выбора регрессионной головы.

Во всех режимах машинного обучения в качестве функции потерь выбрана среднеквадратичная ошибка[5]. Обучение контролируется механизмом ранней остановки: если значение функции потерь на валидационной выборке не улучшается в течение заданного числа последовательных эпох, обучение прекращается и восстанавливается состояние модели с наилучшим валидационным результатом. Для ускорения обучения использован алгоритм оптимизации Adam[6].

1.4. Метрики оценки качества

Для оценки алгоритма использован набор метрик, каждая из которых характеризует определенный аспект качества регрессии. Все метрики вычислены на тестовой выборке в исходных единицах шкалы.

[4]Для реализации последовательной архитектуры слоев перцептрона использован класс Sequential python-библиотеки PyTorch. URL: https://pytorch.org/docs/stable/generated/torch.nn.Sequential.html

[5]Для реализации использован класс MSELoss python-библиотеки PyTorch. URL: https://docs.pytorch.org/docs/stable/generated/torch.nn.MSELoss.html

[6]Для реализации алгоритма использован класс Adam python-библиотеки PyTorch. URL: https://pytorch.org/docs/stable/generated/torch.optim.Adam.html

Средняя абсолютная ошибка **(*mean absolute error, MAE*)** принимает значения от 0 до +∞; значение 0 соответствует идеальному предсказанию[7]. MAE измеряет среднее абсолютное отклонение предсказаний от истинных значений в единицах исходной шкалы. Эта метрика наиболее интуитивно интерпретируема и, в отличие от среднеквадратичной ошибки, одинаково учитывает все ошибки, не усиливая влияние выбросов.

Корень из среднеквадратичной ошибки **(*root mean square error, RMSE*)** принимает значения от 0 до +∞ (0 – идеальное предсказание) и измеряется в тех же единицах, что и целевая переменная[8]. Метрика RMSE выбрана как дополнение к MAE, поскольку она непропорционально сильнее штрафует большие отклонения: возведение в квадрат делает вклад ошибки в 2 балла вчетверо большим, чем ошибки в 1 балл. Если RMSE существенно превышает MAE, это сигнализирует о наличии отдельных примеров с аномально большими ошибками.

Коэффициент детерминации (R^2) принимает значения от −∞ до 1[9]. Значение 1 означает идеальное совпадение предсказаний с истинными значениями; 0 – то, что модель предсказывает не лучше, чем тривиальная модель, всегда возвращающая среднее значение по выборке; < 0 – то, что модель предсказывает хуже среднего. Коэффициент R^2 позволяет судить о качестве модели безотносительно к масштабу данных, что отличает его от MAE и RMSE.

Корреляция Спирмена **(*ρ*)** принимает значения от −1 до +1[10]. Значение 1 означает идеальное совпадение рангов (то есть порядка) предсказанных и истинных значений; 0 – отсутствие монотонной связи; −1 – полностью обратный порядок. Метрика вычисляется для каждой из *k* компонент выходного вектора отдельно и затем усредняется. Корреляция Спирмена выбрана потому, что она

[7]Для вычисления этой метрики использована функция mean_absolute_error python-библиотеки sklearn. URL: https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-error

[8]Для реализации использована функция mean_squared_error python-библиотеки sklearn. URL: https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error

[9]При реализации использована функция r2_score python-библиотеки sklearn. URL: https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

[10]При реализации использована функция spearmanr python-библиотеки SciPy. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

оценивает качество ранжирования, а не абсолютную точность (как последнее имеет место в корреляции Пирсона). Для анализа данных гуманитарных наук это важно, поскольку правильное определение порядка величин относительно друг друга зачастую имеет большее значение, чем точное предсказание абсолютной величины

***Точность извлечения* (*Top-K*)** принимает значения от 0 до 1 (от 0% до 100% совпадения). Для каждого примера определяются *K* компонент с наибольшими значениями в предсказанном и в целевом векторах, затем вычисляется доля пересечения этих двух множеств:

$$Top - K_i = \frac{|top_K(\hat{y}_i) \cap top_K(y_i)|}{K},$$

где top_K (**y**) – множество индексов *K* наибольших компонент вектора **y**, а итоговое значение метрики есть среднее Top-K_i по всем примерам тестовой выборки. Эта метрика отвечает на практический вопрос: если из всех компонент тематического вектора необходимо отобрать *K* наиболее релевантных, то какая доля отобранных компонент (на основе предсказанных алгоритмом значений) совпадет с действительно наиболее значимыми (исходя из значений, принятых в исследовании за эталон)? Значение 1 означает, что модель безошибочно определяет все *K* ведущих компонент; 0 – ни одна из выбранных компонент не входит в число действительно наиболее значимых. В нашем исследовании были выбраны значения *K* = 3 и *K* = 5.

2. РЕЗУЛЬТАТЫ

Экспериментальная оценка была проведена на корпусе из 1105 текстов диссертаций по педагогическим наукам; каждый объект представляет собой объединенное текстовое поле, включающее название диссертации, объект и предмет исследования (ограниченность текстового поля объясняется тем, что контекст используемого трансформерного энкодера ограничен 512 токенами). Данные были разделены на три непересекающихся подвыборки: 773 текста в тренировочной выборке, 166 – в валидационной и 166 – в тестовой (соответственно 70%, 15% и 15%). Целевой вектор оценок $\mathbf{y}_i \in \mathbb{R}^{40}$ состоит из 40 компонент, из которых 15 отражают уровень образования, а оставшиеся 25 описывают предметную область исследования (простота компонентов была обусловле-

на тем, что целью оценки было проверить сам алгоритм, а не способность энкодеров различать сложную «семантику»). Оценки принимают целочисленные значения по шкале от 0 до 10, где 10 означает максимальную релевантность соответствующей компоненты, а 0 – ее отсутствие в тексте.

Эталонные оценки были сформированы с использованием одной генеративной модели, рассматриваемой как наилучшее доступное решение для данного типа задач, – Gemini 3.0 Flash [23], которой предъявлялись информация о диссертации (название, объект, предмет) и список из 40 концептов (взаимно соответствующим 40 компонентам целевого вектора). Модель получала инструкцию оценить релевантность каждого концепта по 11-балльной шкале (0–10), после этого ответы интерпретировались как компоненты вектора $\mathbf{y}_i$. Такой дизайн эксперимента был специально выбран для проверки гипотезы: может ли компактный трансформерный энкодер с относительно небольшим числом параметров и регрессионной головой воспроизвести поведение передовой генеративной модели.

В качестве предобученного трансформерного энкодера была использована модель LaBSE-en-ru – компактная версия Language-Agnostic BERT Sentence Embedding, адаптированная для русского и английского языков. Модель генерирует эмбеддинги размерности d = 768 и содержит порядка 128 млн параметров, что существенно меньше, чем у современных генеративных моделей общего назначения. Выбор LaBSE-en-ru обусловлен, во-первых, ее хорошими показателями на задачах семантического сходства для русского языка, во-вторых, прагматическими соображениями: модель достаточно небольшая и может быть дообучена с использованием ограниченных вычислительных ресурсов.

Во всех нейросетевых режимах обучение проводилось в течение максимального количества эпох - 20, однако фактическое число эпох для каждой модели контролировалось механизмом ранней остановки: если значение функции потерь на валидационной выборке не улучшалось в течение трех последовательных эпох, обучение прекращалось и восстанавливалось состояние модели с наименьшим значением функции потерь. При дообучении энкодера пересчитывались веса последних трех слоев (всего в LaBSE-en-ru 12 слоев).

Табл. 1. Результаты шести режимов на тестовой выборке.

Режим	MAE	RMSE	R^2	ρ	Top-3	Top-5
Гребневая регрессия, замороженная голова	0.92	1.74	0.380	0.475	0.641	0.628
Гребневая регрессия, дообученная	**0.66**	**1.45**	0.562	0.524	0.735	**0.680**
Линейная голова, замороженная	0.94	1.75	0.363	0.452	0.635	0.608
Линейная голова, дообученная	0.65	1.45	**0.569**	**0.540**	**0.744**	0.666
MLP-голова, замороженная	0.87	1.71	0.411	0.477	0.647	0.629
MLP-голова, дообученная	0.68	1.48	0.538	0.538	0.719	0.671

Результаты обучения представлены в табл. 1. Дообучение энкодера дало наибольший эффект: сравнение пар режимов с одинаковой регрессионной головой, но с разным состоянием энкодера показывает, что размораживание трех верхних слоев и их совместное обучение с головой приводит к существенному улучшению всех метрик. Например, для линейной нейросетевой головы коэффициент детерминации R^2 на тестовой выборке возрастает с 0.36 до 0.57, средняя абсолютная ошибка снижается примерно с 0.94 до 0.65 балла, а доля правильно определенных трех наиболее релевантных концептов увеличивается с 0.63 до 0.75. Аналогичная картина наблюдается и для MLP-головы, а также для гребневой регрессии, где R^2 растет с 0.38 до 0.56, а MAE – уменьшается с 0.92 до 0.66. Эти результаты свидетельствуют о том, что предобученные эмбеддинги, хотя и содержат богатую «семантическую» информацию, неоптимальны для конкретной задачи тематической оценки и адаптация верхних слоев энкодера к этой задаче играет положительную роль.

Эффект архитектуры регрессионной головы оказался более умеренным и значительно менее выраженным, чем эффект дообучения энкодера: при замороженном энкодере MLP-голова демонстрирует небольшое преимущество над линейными вариантами по R^2 и ранжирующим метрикам (Спирмен, Top-K), что указывает на способность нелинейного преобразования частично компенсировать ограниченность фиксированных эмбеддингов. Однако после дообучения энкодера различия между типами голов существенно сглаживаются. Это пока-

зывает, что при дообученном пространстве представлений линейного отображения в целевое пространство в значительной степени достаточно, а усложнение головы дает минимальный выигрыш при повышенной вычислительной стоимости и риске переобучения.

Результат сравнения компактной регрессионной архитектуры с эталонным поведением Gemini 3.0 Flash можно оценить, по достигнутым значениям: дообученные режимы объясняют более 56% дисперсии целевых оценок и достигают средней корреляции по Спирмену около 0.54–0.56 между предсказаниями и эталонными оценками по 40-мерному пространству концептов. Три наиболее релевантных концепта определяются правильно примерно в 74% случаев, а пять – примерно в 67–68% случаев. Это указывает на то, что компактный энкодер с регрессионной головой способен достаточно точно при существенно меньших вычислительных затратах аппроксимировать оценки, выставленные передовой генеративной моделью. На практике это означает, что для описанного рода задач массовой автоматической разметки научных текстов можно использовать энкодер с регрессионной головой, сохраняя при этом приемлемый уровень качества и интерпретируемости

ЗАКЛЮЧЕНИЕ

Предложен подход к построению формальных тематических репрезентаций научных текстов, основанный на прямом отображении латентного пространства трансформерного энкодера в целевое метрическое пространство посредством регрессионной головы. Экспериментальная оценка на корпусе фрагментов авторефератов диссертаций по педагогическим наукам подтвердила работоспособность подхода: малый по числу параметров трансформерный энкодер воспроизводит сгенерированную передовой БЯМ оценку трех наиболее релевантных признаков (в рамках задач академической генеалогии) с точностью до 74%.

Практическая значимость предложенного подхода определяется возможностью его развертывания в локальных моделях, способных обрабатывать массивы научных текстов объемом в сотни тысяч и миллионы единиц – масштаб, типичный для библиотечных фондов и наукометрических баз данных. В отличие от обращений к генеративным языковым моделям, инференс с ис-

пользованием регрессионной головы выполняется за один прямой проход и не требует значительных вычислительных ресурсов. Результатом является интерпретируемая формальная репрезентация тематического профиля работы – числовой вектор в пространстве предметных концептов, к которому в дальнейшем могут быть применены разнообразные метрики для анализа динамики научного знания: выявление тематических трендов, кластеризация исследовательских направлений, измерение семантической близости между работами и предметными областями.

Исследование проводилось в условиях малой выборки и ограниченных вычислительных ресурсов, что подтвердило способность подхода функционировать при минимальных затратах. Вместе с тем именно эти ограничения указывают на перспективные направления масштабирования. Во-первых, перспективным является переход от плоского вектора оценок к более сложным формам целевого пространства, например к структурам, согласованным с онтологиями конкретных отраслей знания [24], где выходом модели является не вектор релевантностей, а фрагмент формального графа знаний. Во-вторых, представляет интерес интеграция регрессионной головы не с компактным энкодером, а с современными большими декодерными архитектурами со «смесью экспертов» (mixture-of-experts, MoE): замена выхода одного из экспертов на регрессионную голову позволила бы модели в процессе «рассуждения» порождать формальные репрезентации и тем самым «рассуждать» не только человеческим, но и строгим формальным языком предметной области. Совершенствование логики построения формализованного представления в сочетании с масштабированием на большие модели может дать не только повышение эффективности обработки, но и качественно новые возможности для формального анализа научных текстов.

В более широкой перспективе движение к построению эмбеддинг-пространства научного знания требует конвергенции нескольких традиционно развивающихся обособленно подходов: нейросетевых методов распределенных представлений и символьных методов формализации знания, экспертной работы по построению предметных онтологий и возможностей больших языковых моделей по обработке текстов, эпистемологических концепций структуры научного знания и инструментария машинного обучения. Для этого в том

числе необходимо совершенствование понимания того, какой может быть структура выученного символьного эмбеддинг-пространства и каким образом можно описывать динамические связи в этом пространстве.

СПИСОК ЛИТЕРАТУРЫ

1. *Mulcahy C.* The Mathematics Genealogy Project comes of age at twenty-one // Notices of the AMS. 2017. Vol. 64. No. 5. P. 466–470.

2. *David S.V., Hayden B.Y.* Neurotree: A Collaborative, Graphical Database of the Academic Genealogy of Neuroscience // PLoS ONE. 2012. Vol. 7. No. 10. e46608. https://doi.org/10.1371/journal.pone.0046608

3. *Лернер И.М., Мариносян А.Х., Григорьев С.Г., Юсупов А.Р., Аникиева М.А., Гарифуллина Г.А.* Подход к формированию интеллектуальной академической генеалогии с использованием больших языковых моделей // Электромагнитные волны и электронные системы. 2024. Т. 29. № 4. С. 108–120. https://doi.org/10.18127/j5604128-202404-09

4. *Григорьев С.Г., Лернер И.М., Мариносян А.Х., Григорьева М.А.* К вопросу отбора учебно-методической информации для реализации адаптивной системы управления обучением: алгоритм априорной классификации авторов // Информатика и образование. 2025. Т. 40. № 2. С. 66–78. https://doi.org/10.32517/0234-0453-2025-40-2-66-78

5. *Мариносян А.Х., Григорьев С.Г.* Научные публикации и эмбеддинг-пространство знаний // Электронные библиотеки. 2026. Т. 29. № 2 (в печати).

6. *Salton G., Buckley C.* Term-Weighting Approaches in Automatic Text Retrieval // Information Processing & Management. 1988. Vol. 24. No. 5. P. 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

7. *Sparck Jones K.* A Statistical Interpretation of Term Specificity and Its Application in Retrieval // Journal of Documentation. 1972. Vol. 28. No. 1. P. 11–21. https://doi.org/10.1108/eb026526

8. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // arXiv preprint. 2013. arXiv:1301.3781.

9. *Pennington J., Socher R., Manning C.D.* GloVe: Global Vectors for Word Representation // Proceedings of EMNLP. 2014. P. 1532–1543.

10. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep

Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186. https://doi.org/10.18653/v1/N19-1423

11. *Reimers N., Gurevych I.* Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks // Proceedings of EMNLP. 2019. P. 3982–3992. https://doi.org/10.18653/v1/D19-1410

12. *Beltagy I., Lo K., Cohan A.* SciBERT: A Pretrained Language Model for Scientific Text // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2019. P. 3615–3620. https://doi.org/10.18653/v1/D19-1371

13. *Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F.* Multilingual E5 Text Embeddings: A Technical Report // arXiv preprint. 2024. arXiv:2402.05672.

14. *Мариносян А.Х., Григорьев С.Г., Лернер И.М., Аникиева М.А.* Автоматизированное сравнение научных исследований на базе академической генеалогии // Информатика и образование. 2025. Т. 40. № 6. С. 16–27. https://doi.org/10.32517/0234-0453-2025-40-6-16-27

15. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G.* Mathematical Knowledge Representation: Semantic Models and Formalisms // Lobachevskii Journal of Mathematics. 2014. Vol. 35. No. 4. P. 348–354. https://doi.org/10.1134/S1995080214040143

16. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I.* Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.

17. *Shimanaka H., Kajiwara T., Komachi M.* Machine Translation Evaluation with BERT Regressor // arXiv preprint. 2019. arXiv:1907.12679.

18. *Viskov V., Kokush G., Larionov D., Eger S., Panchenko A.* Semantically-Informed Regressive Encoder Score // Proceedings of the Eighth Conference on Machine Translation (WMT). 2023. P. 815–821. https://doi.org/10.18653/v1/2023.wmt-1.69

19. *Gombert S., Menzel L., Di Mitri D., Drachsler H.* Predicting Item Difficulty and Item Response Time with Scalar-Mixed Transformer Encoder Models and Rational Network Regression Heads // Proceedings of the 19th Workshop on Innova-

tive Use of NLP for Building Educational Applications (BEA 2024). 2024. P. 483–492. URL: https://aclanthology.org/2024.bea-1.40/ (дата обращения 02.02.2026).

20. *Alain G., Bengio Y.* Understanding Intermediate Layers Using Linear Classifier Probes // arXiv preprint. 2017. arXiv:1610.01644.

21. *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research. 2014. Vol. 15. No. 1. P. 1929–1958.

22. *Hoerl A.E., Kennard R.W.* Ridge Regression: Biased Estimation for Non-orthogonal Problems // Technometrics. 1970. Vol. 12. No. 1. P. 55–67. https://doi.org/10.1080/00401706.1970.10488634

23. *Pichai S., Hassabis D., Kavukcuoglu K.* A new era of intelligence with Gemini 3 // Google. The Keyword. URL: https://blog.google/products-and-platforms/products/gemini/gemini-3/#note-from-ceo (дата обращения 02.02.2026).

24. *Елизаров А.М., Кириллович А.В., Липачев Е.К., Невзорова О.А.* Цифровая экосистема OntoMath как подход к построению пространства математических знаний // Электронные библиотеки. 2023. Т. 26. № 2. С. 154–202. https://doi.org/10.26907/1562-5419-2023-26-2-154-202

TYPES OF EMBEDDINGS AND THEIR APPLICATION IN INTELLECTUAL ACADEMIC GENEALOGY

A. Kh. Marinosyan[0000-0003-0577-2360]

Moscow City University, Moscow, Russia

a.marinosyan@yandex.ru

Abstract

The paper addresses the problem of constructing interpretable vector representations of scientific texts for intellectual academic genealogy. A typology of embeddings is proposed, comprising three classes: statistical, learned neural, and structured symbolic. The study argues for combining the strengths of neural embeddings (high semantic accuracy) with those of symbolic embeddings (interpretable dimensions). To operationalize this hybrid approach, an algorithm for learned symbolic embeddings is introduced, which utilizes a regression-based mapping from a model's internal representation to an interpretable vector of scores.

The approach is evaluated on a corpus of fragments from dissertation abstracts in pedagogy. A compact transformer encoder with a regression head was trained to reproduce topic relevance scores produced by a state-of-the-art generative language model. A comparison of six training setups (three regression-head architectures and two encoder settings) shows that fine-tuning the upper encoder layers is the primary driver of quality improvements. The best configuration achieves R^2 = 0.57 and a Top-3 accuracy of 74% in identifying the most relevant concepts. These results suggest that, for tasks requiring formalized output representations, a compact encoder with a regression head can approximate a generative model's behavior at substantially lower computational cost. More broadly, the further development of algorithms for constructing learned symbolic embeddings contributes to building a model of formal knowledge representation in which the convergence of neural and symbolic methods ensures both the scalability of scientific text processing and the interpretability of vector representations that encode their content.

Keywords: *embeddings, academic genealogy, transformer encoder, regression*

head, symbolic embeddings, topic profile, natural language processing, interpretability, large language models, scientometrics.

REFERENCES

1. *Mulcahy C.* The Mathematics Genealogy Project comes of age at twenty-one // Notices of the AMS. 2017. Vol. 64. No. 5. P. 466–470.

2. *David S.V., Hayden B.Y.* Neurotree: A Collaborative, Graphical Database of the Academic Genealogy of Neuroscience // PLoS ONE. 2012. Vol. 7. No. 10. e46608. https://doi.org/10.1371/journal.pone.0046608

3. *Lerner I.M., Marinosyan A.Kh., Grigoriev S.G., Yusupov A.R., Anikieva M.A., Garifullina G.A*. An Approach to the Formation of Intellectual Academic Genealogy Using Large Language Models // Journal Electromagnetic Waves and Electronic Systems. 2024. Vol. 29. No. 4. P. 108–120. https://doi.org/10.18127/j5604128-202404-09 (In Russ.)

4. *Grigoriev S.G., Lerner I.M., Marinosyan A.Kh., Grigorieva M.A.* On the Issue of Educational and Methodological Information Selection for Implementing an Adaptive Learning Management System: Algorithm of A Priori Authors Classification // Informatics and Education / Informatika i obrazovanie. 2025. Vol. 40. No. 2. P. 66–78. https://doi.org/10.32517/0234-0453-2025-40-2-66-78 (In Russ.)

5. *Marinosyan A.Kh., Grigoriev S.G.* Scientific Publications and the Embedding Space of Knowledge // Electronic Libraries / Elektronnye biblioteki. 2026. Vol. 29. No. 2. (In press.) (In Russ.)

6. *Salton G., Buckley C.* Term-Weighting Approaches in Automatic Text Retrieval // Information Processing & Management. 1988. Vol. 24. No. 5. P. 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

7. *Sparck Jones K.* A Statistical Interpretation of Term Specificity and Its Application in Retrieval // Journal of Documentation. 1972. Vol. 28. No. 1. P. 11–21. https://doi.org/10.1108/eb026526

8. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // arXiv preprint. 2013. arXiv:1301.3781.

9. *Pennington J., Socher R., Manning C.D.* GloVe: Global Vectors for Word Representation // Proceedings of EMNLP. 2014. P. 1532–1543.

10. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep

Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186. https://doi.org/10.18653/v1/N19-1423

11. *Reimers N., Gurevych I.* Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks // Proceedings of EMNLP. 2019. P. 3982–3992. https://doi.org/10.18653/v1/D19-1410

12. *Beltagy I., Lo K., Cohan A.* SciBERT: A Pretrained Language Model for Scientific Text // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2019. P. 3615–3620. https://doi.org/10.18653/v1/D19-1371

13. *Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F.* Multilingual E5 Text Embeddings: A Technical Report // arXiv preprint. 2024. arXiv:2402.05672.

14. *Marinosyan A.Kh., Grigoriev S.G., Lerner I.M., Anikieva M.A.* Automated Comparison of Scientific Research Based on Academic Genealogy // Informatics and Education / Informatika i obrazovanie. 2025. Vol. 40. No. 6. P. 16–27. https://doi.org/10.32517/0234-0453-2025-40-6-16-27 (In Russ.)

15. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G.* Mathematical Knowledge Representation: Semantic Models and Formalisms // Lobachevskii Journal of Mathematics. 2014. Vol. 35. No. 4. P. 348–354. https://doi.org/10.1134/S1995080214040143

16. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I.* Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.

17. *Shimanaka H., Kajiwara T., Komachi M.* Machine Translation Evaluation with BERT Regressor // arXiv preprint. 2019. arXiv:1907.12679.

18. *Viskov V., Kokush G., Larionov D., Eger S., Panchenko A.* Semantically-Informed Regressive Encoder Score // Proceedings of the Eighth Conference on Machine Translation (WMT). 2023. P. 815–821. https://doi.org/10.18653/v1/2023.wmt-1.69

19. *Gombert S., Menzel L., Di Mitri D., Drachsler H.* Predicting Item Difficulty and Item Response Time with Scalar-Mixed Transformer Encoder Models and Rational Network Regression Heads // Proceedings of the 19th Workshop on Innova-

tive Use of NLP for Building Educational Applications (BEA 2024). 2024. P. 483–492. URL: https://aclanthology.org/2024.bea-1.40/ (date accessed: 02.02.2026).

20. *Alain G., Bengio Y.* Understanding Intermediate Layers Using Linear Classifier Probes // arXiv preprint. 2017. arXiv:1610.01644.

21. *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research. 2014. Vol. 15. No. 1. P. 1929–1958.

22. *Hoerl A.E., Kennard R.W.* Ridge Regression: Biased Estimation for Non-orthogonal Problems // Technometrics. 1970. Vol. 12. No. 1. P. 55–67. https://doi.org/10.1080/00401706.1970.10488634

23. *Pichai S., Hassabis D., Kavukcuoglu K.* A new era of intelligence with Gemini 3 // Google. The Keyword. URL: https://blog.google/products-and-platforms/products/gemini/gemini-3/#note-from-ceo (date accessed: 02.02.2026).

24. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A. Digital Ecosystem OntoMath as an Approach to Building the Space of Mathematical Knowledge* // Electronic Libraries / Elektronnye biblioteki. 2023. Vol. 26. No. 2. P. 154–202. https://doi.org/10.26907/1562-5419-2023-26-2-154-202 (In Russ.)

СВЕДЕНИЯ ОБ АВТОРЕ

МАРИНОСЯН Андреас Хачатурович – аспирант Института цифрового образования Московского городского педагогического университета. Область научных интересов: наукометрия, обработка естественных языков, архитектуры языковых моделей.

Andreas Khachaturovich MARINOSYAN – PhD Student at the Institute of Digital Education, Moscow City University. Research interests: scientometrics, natural language processing, language model architectures.

email: a.marinosyan@yandex.ru

ORCID: 0000-0003-0577-2360

Материал поступил в редакцию 10 декабря 2025 года

УДК 004.81

КВАНТОВАНИЕ VISION TRANSFORMER: CPU-ЦЕНТРИЧНЫЙ АНАЛИЗ КОМПРОМИССА МЕЖДУ РАЗМЕРОМ МОДЕЛИ И СКОРОСТЬЮ ИНФЕРЕНСА

А. Р. Нигматуллин[1] [0009-0001-6884-1119], Р. А. Лукманов[2] [0000-0001-9257-7410], А. Таха[3] [0009-0006-6346-4162]

[1–3]*Университет Иннополис, г. Иннополис, Россия*

[1]*Центр искусственного интеллекта Университета Иннополис, г. Иннополис, Россия*

[1]am.nigmatullin@innopolis.university, [2]r.lukmanov@innopolis.university, [3]a.taha@innopolis.university

Аннотация

Использование моделей Vision Transformer (ViT) в реальной медицинской практике, например в больницах или диагностических центрах, часто затруднено, потому что на рабочих компьютерах врачей обычно нет мощных графических процессоров (GPU), а имеющиеся вычислительные ресурсы ограничены. В настоящей работе рассмотрен полный путь практической реализации модели на этапе применения (pipeline инференса), направленный на снижение вычислительных затрат без существенной потери качества.

Предложенный подход объединяет несколько методов оптимизации. Во-первых, использована дистилляция знаний (knowledge distillation) – метод обучения, при котором компактная модель копирует поведение более крупной и точной модели-учителя. Во-вторых, применено экспоненциальное скользящее среднее (Exponential Moving Average, EMA) весов, позволяющее стабилизировать обучение и повысить обобщающую способность модели. В-третьих, исследована посттренировочная квантизация до целочисленного формата INT8 (post-training quantization, PTQ), направленная на уменьшение размера модели и ускорение инференса. Дополнительно рассмотрен упрощенный вариант квантизации совместно с обучением (QAT-lite), при котором эффекты квантизации частично учитываются во время

дообучения модели.

Эксперименты проведены на датасете ISIC, содержащем дерматоскопические изображения кожных новообразований. Оценка качества моделей включает стандартные метрики классификации: точность (accuracy), макроусредненную F1-меру и площадь под ROC-кривой (ROC-AUC). Проанализированы характеристики производительности на центральном процессоре (CPU), включая задержку инференса, пропускную способность, потребление памяти и итоговый размер модели.

Полученные результаты показали, что посттренировочная INT8-квантизация позволяет сохранить качество, близкое к модели в формате FP32, при существенном снижении требований к памяти и вычислительным ресурсам. В то же время использование QAT-lite не демонстрирует устойчивых и воспроизводимых улучшений по сравнению с PTQ.

***Ключевые слова**: Визуальный трансформер (ViT), дистилляция знаний, экспоненциальная скользящая средняя (EMA), посттренировочная квантизация, обучение с учетом квантования.*

ВВЕДЕНИЕ

Визуальные трансформеры (Vision Transformers, ViT) показывают высокие результаты в задачах компьютерного зрения, включая анализ медицинских изображений. Однако их практическое использование в клинических условиях остается ограниченным. Основная причина этого заключается в высокой вычислительной нагрузке, особенно при работе на стандартных центральных процессорах (CPU), которые широко используются в медицинских учреждениях.

В работе рассмотрен практический подход к применению ViT в медицинской визуализации при ограниченных вычислительных ресурсах. Исследован оптимизационный пайплайн, направленный на уменьшение размера модели и ускорение инференса без заметного ухудшения качества. В качестве основного метода использована дистилляция знаний (knowledge distillation), при которой компактная модель обучается на основе выходов более крупной и точной модели. Для стабилизации обучения применено экспоненциальное скользящее среднее (Exponential Moving Average, EMA) весов. Дополнительно использована посттренировочная квантизация

до целочисленного формата INT8, позволяющая снизить требования к памяти и сократить время обработки изображений.

Представлена последовательная оценка этих методов в контексте медицинских задач. Основное внимание уделено балансу между точностью классификации и вычислительной эффективностью моделей при выполнении инференса на CPU.

За последние годы ViT-модели стали широко применяться в анализе медицинских изображений. Обзорные работы показали, что по сравнению со сверточными нейронными сетями такие модели лучше учитывают глобальный контекст изображения, что важно для медицинских данных [1, 2]. Применения ViT в гистопатологии рассмотрена в [3], где также обсуждены существующие ограничения. Более общие обзоры использования трансформеров в медицинской визуализации представлены в [4, 5]. Работы, посвященные сегментации, подчеркивают роль трансформеров в точном анализе структур и границ на медицинских изображениях [6].

Основным ограничением для применения ViT в клинических задачах медицинской визуализации остается их высокая вычислительная стоимость. Одно из направлений исследований связано с разработкой облегченных архитектур [7]. Однако на практике чаще используют методы сжатия уже обученных моделей. К таким методам относится квантизация, которая позволяет уменьшить объем памяти и ускорить вычисления [8–11]. Обобщенный обзор методов сжатия представлен в [12].

Дистилляция знаний является еще одним распространенным подходом к уменьшению размера моделей при сохранении приемлемой точности. В классической работе Хинтона и соавторов [13] было показано, что компактная модель может эффективно обучаться на выходах более сложной модели. В дальнейшем этот подход был расширен и адаптирован для различных сценариев, включая медицинские задачи [14–17].

В целом проводимые исследования в области медицинской визуализации развиваются в двух направлениях: создание более компактных архитектур и применение методов сжатия для адаптации моделей к ограниченным вычислительным условиям. Настоящая работа объединяет эти подходы и оценивает совместное использование дистилляции знаний,

экспоненциального скользящего среднего весов и INT8-квантизации для ViT-моделей, предназначенных для инференса на CPU.

МЕТОДЫ

Для получения компактной, но при этом точной модели мы используем дистилляцию знаний. В рамках этого подхода меньшая нейронная сеть, называемая моделью-студентом, обучается с опорой на выходы более крупной и предварительно обученной модели-учителя. Основная идея заключается в том, что модель-учитель передает модели-студенту не только правильные ответы, но и более богатую информацию о структуре задачи, содержащуюся в распределении выходных вероятностей.

Обучение модели-студента проводится с использованием комбинированной функции потерь. С одной стороны, используется стандартная функция кросс-энтропии, которая измеряет соответствие предсказаний модели-студента истинным меткам классов. С другой стороны, добавляется функция потерь дистилляции, которая поощряет совпадение распределения выходов модели-студента с распределением выходов модели-учителя. Такое сочетание позволяет сохранить высокую точность даже при существенном уменьшении размера модели.

Комбинированная функция потерь имеет следующий вид:

$$\mathrm{LKD} = (1 - \alpha)\mathrm{CE}(z_s, y) + \alpha T^2 \mathrm{KL}(softmax(\tfrac{z_t}{T})||softmax\left(\tfrac{z_s}{T}\right)),$$

где z_t и z_s обозначают логиты модели-учителя и модели-студента соответственно, y – истинные метки классов, CE(·) — функция кросс-энтропии, KL(·‖·) — дивергенция Кульбака–Лейблера, T – температурный параметр, сглаживающий распределение вероятностей, $\alpha \in [0,1]$ — коэффициент, определяющий баланс между вкладом стандартной функции потерь и потерь дистилляции.

Экспоненциальное скользящее среднее весов

Для повышения устойчивости обучения и улучшения обобщающей способности модели применим экспоненциальное скользящее среднее (Exponential Moving Average, EMA) весов. Вместо использования мгновенных значений параметров модели, EMA поддерживает сглаженную версию весов,

которая обновляется постепенно и менее чувствительна к шуму градиентов.

Обновление EMA выполняется по следующему правилу:

$$m_t = \beta m_{t-1} + (1-\beta)\theta_t, \ \ \beta = 0.999,$$

где θ_t – текущие значения весов модели на шаге обучения t, а m_t – соответствующие EMA-веса. Использование EMA позволяет получить более стабильную модель для оценки и инференса, что особенно важно в условиях ограниченных вычислительных ресурсов.

Квантизация модели

Для дальнейшего уменьшения вычислительной нагрузки и объема памяти используют квантизацию параметров модели. В настоящей работе применено аффинное квантизирование, при котором значения с плавающей запятой преобразуются в 8-битные целые числа. Преобразование описывается следующими соотношениями:

$$x_{int} = \mathrm{round}\left(\frac{x}{s}\right) + z, \ \ \hat{x} = s(x_{\mathrm{int}} - z).$$

где s – масштаб, а x_{int} – нулевая точка. Эти параметры подбираются отдельно для каждого слоя модели, что позволяет более точно аппроксимировать исходные значения.

При квантизации модели в целые числа переводят только полносвязные слои, так как они выполняют большинство вычислений. Другие операции, которые отвечают за нормализацию данных внутри сети и преобразование выходов модели в вероятности для классов, оставляют в привычном формате с плавающей запятой (FP32). Это делается потому, что такие операции очень чувствительны к точности чисел, и если их квантизировать полностью, модель может работать нестабильно.

Мы рассматриваем два варианта квантизации. В случае посттренировочной квантизации (Post-Training Quantization, PTQ) модель квантизируется после завершения обучения без изменения весов. В варианте QAT-lite эффекты квантизации частично учитываются во время короткого этапа дообучения за счет использования так называемой «фейковой» квантизации, имитирующей целочисленные вычисления.

Архитектуры и базовые модели

Мы используем подход «учитель – студент», при котором одна модель служит источником знаний, а другая – компактной версией, предназначенной для практического применения. В качестве модели-учителя выбрана архитектура DeiT-Small@224. Это визуальный трансформер, который демонстрирует высокую точность при работе с изображениями стандартного разрешения и широко используется в исследовательских работах как надежная и сбалансированная базовая модель. Его вычислительная сложность делает его удобным эталоном качества, однако в клинических условиях такая модель часто оказывается слишком ресурсоемкой для повседневного использования.

В роли модели-студента была использована DeiT-Tiny@224 – более компактная версия той же архитектуры. По сравнению с моделью-учителя она содержит существенно меньше параметров и требует меньших вычислительных затрат, что делает ее более подходящей для развертывания в средах с ограниченными ресурсами. В частности, такая модель может использоваться для инференса на центральном процессоре без необходимости применения графических ускорителей, что соответствует типичным условиям эксплуатации в медицинских учреждениях.

Для корректной оценки качествоа и практической применимости модели-студента ее характеристики сравниваем не только с моделью-учителем, но и с рядом широко распространенных сверточных нейронных сетей. Эти модели выбраны таким образом, чтобы представить разные поколения и различные подходы в архитектурах для анализа изображений.

В качестве базовой модели была использована ResNet-18, как одна из наиболее часто используемых архитектур в задачах компьютерного зрения. Это одна из наиболее известных и хорошо изученных архитектур, которая часто применяется в медицинской визуализации и служит удобной точкой отсчета при сравнении новых методов. Модель MobileNetV3-Large включена в сравнение как пример архитектуры, специально разработанной для эффективного инференса при ограниченных вычислительных ресурсах. Такие модели широко используют в мобильных и встроенных системах, где важны низкая задержка и малое потребление памяти. Дополнительно

рассмотрена ConvNeXt-Tiny – современная сверточная архитектура, которая заимствует ряд идей из трансформеров и демонстрирует высокое качество при относительно умеренной вычислительной сложности. Эта модель использована в качестве сильного современного ориентира среди CNN.

Таким образом, выбранный набор моделей позволяет оценить положение компактного визуального трансформера относительно как более тяжелых моделей трансформеров, так и различных сверточных архитектур, применяемых на практике.

Оценка вычислительной сложности

При анализе вычислительной сложности моделей мы учитываем не только стандартные теоретические показатели, но и характеристики, важные для реального использования в клинических условиях. В частности, оцениваем общее число параметров модели и теоретическую вычислительную нагрузку, выраженную в количестве операций с плавающей запятой (FLOPs). Эти показатели дают общее представление о сложности архитектуры, однако не всегда отражают реальные затраты при развертывании. Поэтому дополнительно рассматриваем практические метрики, такие как фактический размер контрольной точки модели на диске. Этот показатель напрямую связан с требованиями к хранению данных, скорости загрузки модели и возможностям ее обновления в медицинских информационных системах. Учет как теоретических, так и практических характеристик позволяет более полно оценить пригодность моделей для использования в клинических сценариях с ограниченными вычислительными ресурсами.

ЭКСПЕРИМЕНТЫ

Эксперименты были проведены на наборе данных ISIC – общедоступном медицинском датасете, предоставленном International Skin Imaging Collaboration и содержащем дерматоскопические изображения кожных поражений. Этот набор данных широко используется для оценки алгоритмов автоматической классификации в дерматологии и является стандартным набором в исследованиях по медицинской визуализации. Мы использовали заранее определенные разбиения данных на обучающую, валидационную

и тестовую выборки.

Для обеспечения воспроизводимости экспериментов во всех запусках были применены фиксированные значения случайных инициализаций. На этапе тестирования аугментации изображений не использовались, чтобы полученные результаты отражали поведение моделей в условиях практического применения. Измерения производительности при инференсе на центральном процессоре (CPU) проводились с учетом этапа разогрева, после чего инференс запускался несколько раз подряд. Это позволило получить устойчивые оценки времени обработки одного изображения.

Обучение моделей выполнялось с использованием оптимизатора AdamW. Скорость обучения изменялась по косинусному расписанию, обеспечивающему плавное снижение шага оптимизации. В процессе дистилляции знаний было применено экспоненциальное скользящее среднее (Exponential Moving Average, EMA) весов, что позволило получить более стабильные параметры модели для последующей оценки.

Квантизация моделей была реализована с использованием backend fbgemm, оптимизированного для целочисленных вычислений на CPU. В случае посттренировочной квантизации (Post-Training Quantization, PTQ) применялась динамическая квантизация линейных слоев без дополнительного обучения модели. Для варианта QAT-lite был использован короткий этап дообучения продолжительностью пять эпох, в течение которого эффекты квантизации учитывались во время обучения перед сохранением итоговых весов модели.

Качество классификации было оценено с использованием стандартных метрик: точности (accuracy), макроусредненной F1-меры и площади под ROC-кривой (ROC-AUC), которая отражает способность модели различать классы при разных порогах принятия решения. Для оценки практической применимости моделей в клинических условиях дополнительно были проанализированы вычислительные характеристики при работе на CPU, включая задержку инференса, пропускную способность, пиковое использование оперативной памяти (RAM) и итоговый размер модели на диске. Совокупность этих показателей позволяет оценить как качество предсказаний, так и вычислительную пригодность моделей для использования в реальных медицинских сценариях.

РЕЗУЛЬТАТЫ

Далее представлены результаты сравнения моделей по качеству классификации, требованиям к памяти и производительности при инференсе на центральном процессоре (Intel i7-12700F). Основное внимание уделено влиянию посттренировочной INT8-квантизации и ее сочетания с дистилляцией знаний и экспоненциальным скользящим средним весов.

Качество классификации и использование памяти

Эксперименты на наборе данных ISIC показали, что посттренировочная INT8-квантизация практически не ухудшает качество дистилляционной модели DeiT-Tiny по сравнению с базовой версией в формате с плавающей запятой (FP32). На валидационной выборке точность изменилась всего на − 0.13 п. п., а макроусредненная F1-мера даже незначительно выросла (+0.27 п.п.). Аналогичная картина наблюдалась и на тестовом разбиении, где изменения составили −0.10 п.п. по точности и +0.08 п.п. по макро-F1. При этом выигрыш в компактности модели оказался существенным. Размер модели на диске уменьшился с 21.13 до 5.97 МБ, то есть примерно в 3.5 раза (−71.7%). Пиковое использование оперативной памяти во время инференса также снизилось примерно на 247 МБ, что соответствует уменьшению на 14%. Эти результаты особенно важны для клинических сценариев, где ограничения по памяти и хранению данных часто являются критичными.

Задержка и пропускная способность на CPU

Производительность моделей на CPU зависит от размера обрабатываемого пакета изображений (batch size) и используемого варианта модели. При обработке одного изображения за раз (batch = 1) квантизированная модель Student INT8 (PTQ) оказалась немного медленнее версии FP32: медианная задержка составила 16.49 мс против 14.53 мс, а пропускная способность снизилась примерно на 11%. Это связано с дополнительными накладными расходами на операции квантования и деквантования. Однако при использовании экспоненциального скользящего среднего весов (KD+EMA+PTQ) эта разница практически исчезла. В данном случае задержка инференса почти совпала с лучшим вариантом FP32

и оказалась заметно ниже, чем у модели KD+EMA в формате FP32 (14.46 мс против 16.77 мс).

При увеличении размера пакета до batch = 8 негативный эффект квантизации исчезает. Простая PTQ-квантизация показала задержку на уровне FP32 (58.45 мс против 58.76 мс), а сочетание KD+EMA+PTQ продемонстрировало преимущество по скорости по сравнению с KD+EMA FP32 (53.38 мс против 59.22 мс), а также более высокую пропускную способность (+8.2%). Это указывает на то, что квантизация особенно эффективна при вычислительно нагруженных сценариях.

Сравнение с базовыми CNN-архитектурами

Сравнение с распространенными сверточными архитектурами показало, что дистилляционный DeiT-Tiny остается конкурентоспособным при инференсе на CPU. При batch = 1 модель DeiT-Tiny FP32 (KD) работает быстрее, чем ResNet-18 FP32 (14.53 мс против 15.29 мс), и значительно быстрее, чем ConvNeXt-Tiny FP32 (38.28 мс). При batch = 8 архитектура MobileNetV3-Large FP32 сохраняет лидерство по пропускной способности, что ожидаемо с учетом ее ориентации на высокоэффективный инференс.

Вариант квантизации с учетом обучения (QAT-lite) не продемонстрировал устойчивых преимуществ. Он уступает PTQ по качеству классификации и не обеспечивает заметного выигрыша по задержке ни при одном из рассмотренных размеров пакета.

Практические выводы

С точки зрения практического развертывания полученные результаты показали, что посттренировочная INT8-квантизация является простым и надежным способом уменьшить размер ViT-моделей примерно в 3.5 раза при сохранении качества практически на уровне FP32. Такой подход особенно полезен в сценариях, где ограничения по памяти, хранению данных или распространению моделей играют ключевую роль.

Дополнительно эти результаты позволяют четко определить условия, при которых квантизация дает наибольший эффект. Ускорение наблюдается в вычислительно нагруженных режимах, при использовании более крупных пакетов изображений и в вариантах с EMA-стабилизацией весов. В то же время

задержка обработки одного изображения остается ограниченной накладными расходами на операции квантования и деквантования вокруг чувствительных слоев, таких как Layer Normalization и Softmax. Подробные численные результаты сравнения представлены в табл. 1, 2.

Динамика обучения модели-студента проанализирована в зависимости от номера эпохи (*epoch*). Под одной эпохой понимается один полный проход всей обучающей выборки через модель с последующим обновлением параметров. На рис. 1 показаны кривые для обучающей выборки при использовании дистилляции знаний и экспоненциального скользящего среднего весов. Видно, что применение EMA способствует более стабильному и плавному снижению значения функции потерь, а также уменьшает флуктуации точности в процессе обучения.

Аналогичный анализ на валидационной выборке представлен на рис. 2. В этом случае также наблюдается отсутствие резких скачков метрик, что указывает на лучшую обобщающую способность модели и снижение риска переобучения.

Отдельно рассмотрен короткий этап дообучения с учетом квантизации (QAT-lite). Соответствующие кривые представлены на рис. 3. Можно отметить, что несмотря на небольшое улучшение сходимости на первых итерациях, дальнейшее обучение не приводит к существенному росту качества, что согласуется с результатами количественного сравнения и подтверждает ограниченную эффективность QAT-lite по сравнению с посттренировочной квантизацией.

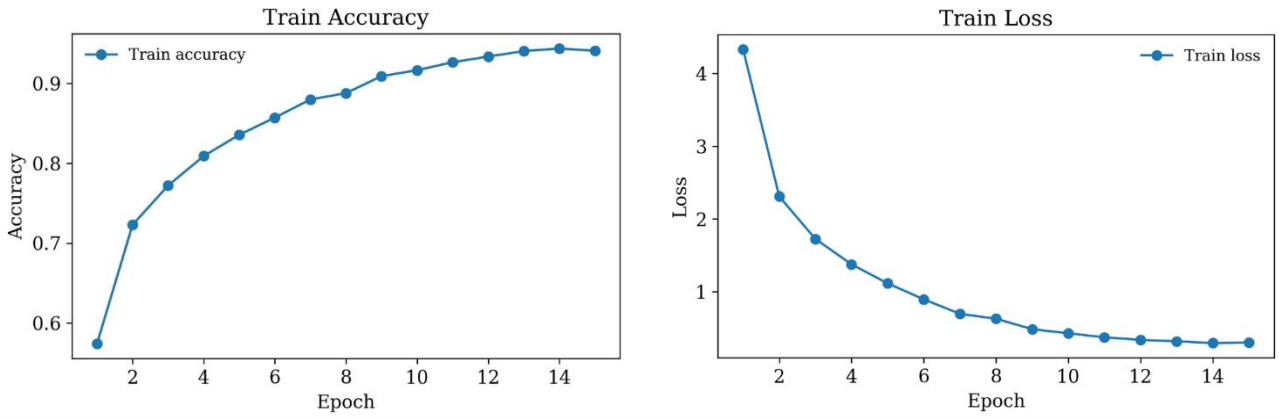


Рис. 1. Кривые изменения функции потерь (loss) и точности (accuracy) во время обучения модели-студента DeiT-Tiny с применением дистилляции знаний (KD) и экспоненциального скользящего среднего весов (EMA).

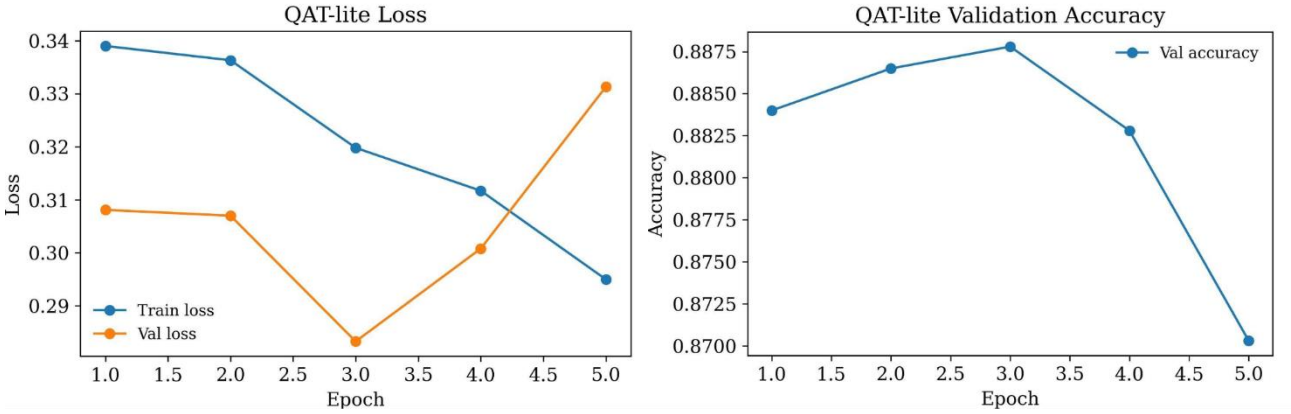


Рис. 2. Кривые изменения функции потерь и точности на валидационной выборке в процессе обучения модели-студента DeiT-Tiny.

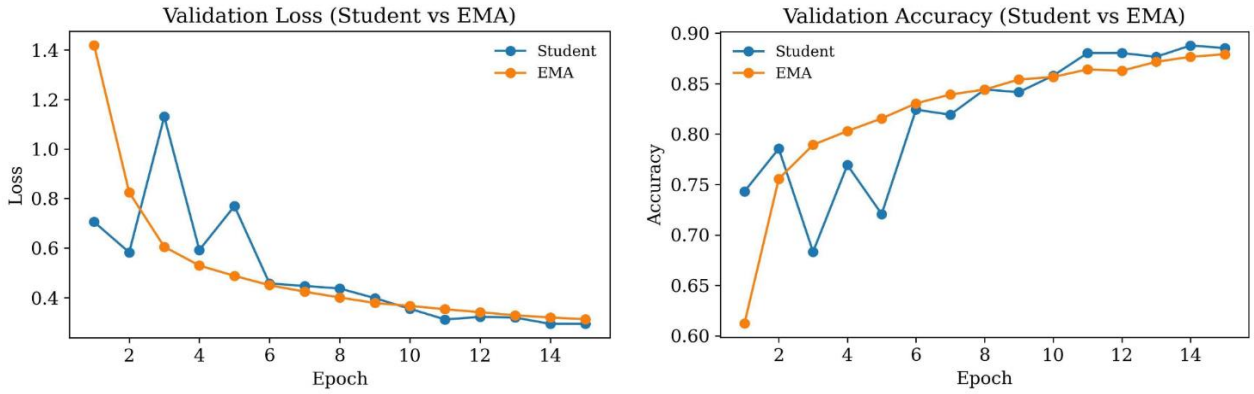


Рис. 3. Кривые изменения метрик модели во время короткого этапа дообучения с учетом квантизации (QAT-lite). Этот этап включает как дистилляцию знаний (KD), так и имитацию целочисленных вычислений во время обучения, чтобы модель могла корректно работать в формате INT8.

Для оценки влияния различных подходов к обучению и упрощению числовых представлений была выполнена экспериментальная проверка качества моделей на валидационной и тестовой выборках.

Результаты получены на наборе данных ISIC, который состоит из медицинских изображений кожи, используемых для задачи классификации кожных поражений. Для экспериментов были применены заранее зафиксированные разбиения данных на обучающую, валидационную и тестовую части. При оценке качества модели изображения использовались

без дополнительных преобразований, а все эксперименты выполнялись с фиксированными начальными условиями, что обеспечивает воспроизводимость результатов.

Табл. 1. Метрики валидации и теста.

Модель	Val Acc	Val Macro-F1	Val ROC-AUC	Test Acc	Test Macro-F1	Test ROC-AUC
Student FP32 (KD)	0.8878	0.8874	0.9903	0.8333	0.7857	0.9658
Student FP32 (KD+EMA)	0.8791	0.8787	0.9889	0.8367	0.7903	0.9657
Student INT8 (PTQ)	0.8865	0.8901	0.9901	0.8323	0.7865	0.9658
Student INT8 (KD+EMA+PTQ)	0.8741	0.8700	0.9891	0.8342	0.7908	0.9652
Student QAT-lite FP32	0.8678	0.8789	0.9908	0.8103	0.7834	0.9634
Student QAT-lite INT8	0.8666	0.8769	0.9907	0.8113	0.7897	0.9631

В табл. 1 представлены значения основных метрик качества для компактной модели DeiT-Tiny@224, обученной с использованием более крупной и точной модели DeiT-Small@224 в качестве источника знаний. Такой подход позволил уменьшить размер модели при сохранении высокой точности. В ряде экспериментов применялось экспоненциальное скользящее среднее весов (EMA) с коэффициентом $\beta = 0.999$, которое сглаживает изменения параметров модели и делает итоговые предсказания более стабильными. Обозначение INT8 (PTQ) соответствует упрощению числовых представлений параметров модели после завершения обучения, что уменьшает ее размер и требования к памяти. При этом наиболее чувствительные вычислительные операции сохраняются в исходном формате, чтобы избежать нестабильности вычислений. Вариант QAT-lite включает короткий этап дополнительного дообучения модели, во время которого она подстраивается под последующее упрощение числовых представлений.

Качество моделей было оценено на валидационной (Val) и тестовой (Test) выборках по следующим показателям:

1) Accuracy – доля правильных предсказаний;

2) Macro-F1 – усредненная мера качества по всем классам, одинаково учитывающая каждый из них;

3) ROC-AUC – показатель способности модели различать классы на основе предсказанных вероятностей.

Во всех случаях более высокие значения метрик соответствуют лучшему качеству модели.

Сравнение методов уменьшения модели

В табл. 2 представлены результаты сравнения степени уменьшения размера различных вариантов модели DeiT-Tiny, обученных и оптимизированных различными способами, на наборе медицинских изображений ISIC. Во всех случаях была использована одна и та же архитектура модели, поэтому варианты не отличаются по своей вычислительной сложности: число параметров составляет 5.53 млн, а объем вычислений – 2.15 млрд операций для изображений с разрешением 224 × 224.

Табл. 2. Сводные по степени сжатия различных вариантов модели DeiT-Tiny

Модель	Точность	Размер (МБ)	Compression factor
DeiT-T FP32 (KD)	FP32	21.13	1.00
DeiT-T FP32 (KD+EMA)	FP32	21.13	1.00
DeiT-T INT8 (PTQ)	INT8	5.97	3.54
DeiT-T INT8 (KD+EMA+PTQ)	INT8	5.97	3.54
DeiT-T QAT-lite FP32	FP32	21.13	1.00
DeiT-T QAT-lite INT8	INT8	5.97	3.54

Отличия данных для различных моделей связаны исключительно со способом представления параметров модели. Варианты с пометкой INT8 используют упрощенный числовой формат для части операций, что позволяет существенно сократить занимаемое пространство. При этом упрощение применяется только к основным линейным слоям модели, тогда как наиболее чувствительные операции, отвечающие за нормализацию и вычисление вероятностей, сохраняются в исходном формате для обеспечения стабильной работы. Указанные в таблице размеры моделей соответствуют фактическому объему файлов с сохраненными весами на диске, что напрямую отражает требования к хранению и передаче модели в реальных сценариях развертывания.

С учетом указанных различий в представлении параметров ниже приведены результаты измерения производительности моделей на CPU.

CPU benchmarks

После анализа степени сжатия и размера моделей рассмотрим их практическую производительность при инференсе на центральном процессоре (CPU).

Табл. 3. Производительность моделей на CPU при обработке одного изображения за раз (batch = 1). Более низкая задержка и меньший расход памяти, а также более высокая пропускная способность означают лучшую эффективность.

Модель	p50 (мс)	p90 (мс)	Throughput (кол-во/с)	Peak RAM (МБ)	Size (MB)
DeiT-T FP32 (KD)	14.53	15.91	67.7	1765.1	21.13
DeiT-T FP32 (KD+EMA)	16.77	17.87	59.2	1779.0	21.13
DeiT-T INT8 (PTQ)	16.49	17.17	60.3	1518.3	5.97
DeiT-T INT8 (KD+EMA+PTQ)	14.46	16.84	66.9	1532.2	5.97
DeiT-T QAT-lite FP32	17.26	18.34	58.0	1539.8	21.13
DeiT-T QAT-lite INT8	15.77	17.55	62.3	1533.2	5.97
ResNet-18 FP32	15.29	17.69	61.9	1539.6	42.72
MobileNetV3-L FP32	16.58	17.19	60.4	1616.1	16.25
ConvNeXt-T FP32	38.28	39.21	26.4	1642.2	106.21
ResNet-18 INT8 (PTQ)	17.54	19.04	56.3	1675.7	42.71
MobileNetV3-L INT8 (PTQ)	17.56	18.06	56.7	1675.7	16.25
ConvNeXt-T INT8 (PTQ)	36.16	36.82	27.8	1675.3	32.18

Измерения производительности выполнены по следующему протоколу: 50 итераций использованы для разогрева системы, после этого выполнены 100 синхронизированных измерительных итераций; весь процесс повторялся пять раз для повышения надежности результатов. Показатели, приведенные в табл. 3, интерпретируются следующим образом:

1. p50/p90 – медианное и 90-й перцентиль времени обработки одного изображения (в миллисекундах);
2. Throughput – количество изображений, обрабатываемых моделью в секунду в ходе измерений;
3. Peak RAM – максимальный объем оперативной памяти, используемый во время инференса (в мегабайтах);

4. Size – фактический размер файла с сохраненными весами модели на диске (в мегабайтах).

Все варианты модели DeiT-Tiny используют одинаковую архитектуру, следовательно, имеют одинаковую теоретическую вычислительную сложность: 5.53 млн параметров и 2.15 млрд операций для изображений с разрешением 224 × 224. В вариантах с INT8-квантизацией упрощение числовых представлений было применено только к линейным слоям модели с использованием библиотеки fbgemm, тогда как операции нормализации и вычисления вероятностей сохранялись в исходном формате для обеспечения стабильности вычислений. В табл. 4 представлены экспериментальные показатели производительности различных вариантов модели на CPU.

Табл. 4. Производительность моделей на CPU при batch = 8 (50 итераций разогрева, 100 измерительных итераций, 5 повторов)

Модель	p50 (мс)	p90 (мс)	Throughput (кол-во/с)	Peak RAM (МБ)	Size (MB)
DeiT-T FP32 (KD)	58.76	60.45	137.1	1779.0	21.13
DeiT-T FP32 (KD+EMA)	59.22	61.11	136.1	1779.0	21.13
DeiT-T INT8 (PTQ)	58.45	61.16	136.2	1528.5	5.97
DeiT-T INT8 (KD+EMA+PTQ)	53.38	59.25	147.3	1538.6	5.97
DeiT-T QAT-lite FP32	59.45	60.75	134.2	1539.8	21.13
DeiT-T QAT-lite INT8	57.79	59.75	138.2	1538.2	5.97
ResNet-18 FP32	78.82	81.17	101.5	1613.9	42.72
MobileNetV3-L FP32	48.04	48.58	166.4	1641.3	16.25
ConvNeXt-T FP32	189.68	193.84	42.0	1675.6	106.21
ResNet-18 INT8 (PTQ)	78.56	80.60	101.5	1675.7	42.71
MobileNetV3-L INT8 (PTQ)	48.25	48.84	165.4	1675.7	16.25
ConvNeXt-T INT8 (PTQ)	162.62	165.86	49.2	1714.2	32.18

Измерения производительности выполнены согласно протоколу, описанному для табл. 3.

В табл. 5 приведены результаты абляционных экспериментов для различных вариантов компактной модели DeiT-Tiny на валидационной выборке при инференсе на CPU с размером пакета batch = 1.

Все варианты модели используют ту же самую архитектуру (5.53 млн параметров, 2.15 млрд операций при разрешении 224 × 224). Варианты с INT8-квантизацией (PTQ или QAT-lite INT8) упрощают представление чисел только для линейных слоев с использованием библиотеки fbgemm, при этом операции LayerNorm и Softmax были сохранены в исходном формате для числовой стабильности.

Табл. 5. Абляционные эксперименты (валидационная выборка; CPU, batch=1)

Variant	Acc	Macro-F1	ROC-AUC	p50 (мс)	Thr (кол-во/с)	Size (МБ)
KD (Student FP32)	0.8878	0.8874	0.9903	14.53	67.7	21.13
EMA (KD+EMA FP32)	0.8791	0.8787	0.9889	16.77	59.2	21.13
PTQ (KD INT8)	0.8865	0.8901	0.9901	16.49	60.3	5.97
KD+EMA+PTQ	0.8741	0.8700	0.9891	14.46	66.9	5.97
QAT-lite FP32	0.8678	0.8789	0.9908	17.26	58.0	21.13
QAT-lite INT8	0.8666	0.8769	0.9907	15.77	62.3	5.97

ОБСУЖДЕНИЕ

Проведенные эксперименты показали, что сочетание дистилляции знаний и посттренировочной квантизации является эффективным способом использования моделей Vision Transformer (ViT) на обычных CPU, особенно при ограничениях по памяти. Применение INT8-квантизации к линейным слоям модели DeiT-Tiny позволило уменьшить размер модели почти в 3.5 раза – с 21.13 до 5.97 МБ, при этом точность осталась практически на прежнем

уровне. Это говорит о том, что значительное сжатие возможно без сложного дополнительного обучения.

Установлено, что меньшая модель на CPU не всегда работает быстрее. При обработке одного изображения за раз (batch = 1) модели с INT8-квантизацией иногда даже немного медленнее, чем исходные FP32-модели. Причиной служат накладные расходы на преобразование чисел между INT8 и FP32 в слоях LayerNorm и Softmax, которые компенсируют преимущества целочисленных вычислений. Существенное улучшение пропускной способности наблюдается только при пакетной обработке нескольких изображений (batch = 8), что подтверждает, что INT8 выгоден в вычислительно насыщенных сценариях, но не для одиночных картинок.

Таким образом, посттренировочная квантизация PTQ оказалась простым и безопасным методом уменьшения модели без потери точности, особенно когда важна экономия памяти. Однако если приоритет – минимальная задержка при обработке одного изображения, преимущества квантизации ограничены. Кроме того, необходимо учитывать особенности конкретного процессора, так как на других CPU результаты могут отличаться.

ЗАКЛЮЧЕНИЕ

Мы показали, что при использовании моделей Vision Transformer в медицинских приложениях на обычных CPU квантизация не всегда дает ускорение. Для успешного развертывания таких моделей важно учитывать не только методы сжатия, но и особенности инференса на конкретном процессоре. Наши результаты подчеркивают, что подходы, которые уменьшают размер модели, не всегда автоматически сокращают задержку при обработке отдельных изображений.

В будущем перспективными представляются методы, нацеленные именно на снижение задержки, такие как структурная обрезка отдельных блоков модели, схемы квантизации, учитывающие особенности аппаратуры, выборочная квантизация для снижения накладных расходов, а также изучение влияния сжатия на точность калибровки и неопределенность предсказаний. Все это поможет повысить надежность и доверие к медицинским системам на основе искусственного интеллекта.

Благодарности

Работа поддержана Академией наук Республики Татарстан, договор №254/2024-PD.

СПИСОК ЛИТЕРАТУРЫ

1. *Shamshad F., Khan S., Zamir S.W., et al.* Transformers in Medical Imaging: A Survey // arXiv. 2022.

2. *He K., Gan C., et al.* Transformers in Medical Image Analysis: A Review // arXiv. 2022.

3. *Atabansi C.C., Nie J., et al.* A Survey of Transformer Applications for Histopathological Image Analysis: New Developments and Future Directions // Biomedical Engineering Online. 2023. Vol. 22, No. 1. https://doi.org/10.1186/s12938-023-01069-5

4. *Azad R., Kazerouni A., Heidari M., et al.* Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review // arXiv. 2023.

5. *Shamshad F., Khan S., Zamir S.W., et al.* Transformers in Medical Imaging: A Survey // Medical Image Analysis. 2024. Vol. 88. https://doi.org/10.1016/j.media.2023.102843

6. *Liu Y., et al.* A Recent Survey of Vision Transformers for Medical Image Segmentation // *arXiv.* 2023.

7. *Wu F., et al.* Lite Transformer with Long-Short Range Attention // Proceedings of the International Conference on Learning Representations (ICLR). 2020.

8. *Jacob B., et al.* Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2704–2713. https://doi.org/10.1109/CVPR.2018.00286

9. *Nagel M., et al.* A White Paper on Neural Network Quantization // arXiv. 2021.

10. *Han S., Mao H., Dally W.J.* Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // arXiv. 2016.

11. *Yao Z., et al.* ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers // Advances in Neural Information

Processing Systems (NeurIPS). 2022. Vol. 35.

12. *Wikipedia contributors.* Model Compression // Wikipedia. 2025.

13. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // arXiv. 2015.

14. *Gou J., et al.* Knowledge Distillation: A Survey // International Journal of Computer Vision. 2021. Vol. 129, No. 6. P. 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

15. *Umirzakova S., et al.* Simplified Knowledge Distillation for Deep Neural Networks: Bridging the Performance Gap with a Novel Teacher–Student Architecture // Electronics. 2024. Vol. 13, No. 3. https://doi.org/10.3390/electronics13030512

16. *Liang P., et al.* Data-Free Knowledge Distillation with Feature Synthesis and Spatial Consistency for Image Analysis // Scientific Reports. 2024. Vol. 14, No. 1. https://doi.org/10.1038/s41598-024-53241-3

VIT QUANTIZATION: CPU-CENTRIC ANALYSIS OF THE TRADE-OFF BETWEEN SIZE AND SPEED

A. R. Nigmatullin[1] [0009-0001-6884-1119], R. A. Lukmanov[2] [0000-0001-9257-7410],

A. Taha[3] [0009-0006-6346-4162]

[1–3]*Innopolis University, Innopolis, Russia*

[1]*The Center of Artificial Intelligence of Innopolis University, Innopolis, Russia*

[1]am.nigmatullin@innopolis.university, [2]r.lukmanov@innopolis.university,
[3]a.taha@innopolis.university

Abstract

Using Vision Transformer (ViT) models in real medical practice – for example, in hospitals or diagnostic centers – is often difficult because doctors' work computers usually do not have powerful graphics processors (GPUs), and computing resources are limited. This work investigates a complete practical pipeline for model inference, aimed at reducing computational costs without significant loss of predictive performance.

The proposed approach combines several optimization techniques. First, knowledge distillation (KD) is used, where a compact student model learns to mimic the behavior of a larger, more accurate teacher model. Second, Exponential Moving Average (EMA) of the model weights is applied to stabilize training and improve generalization. Third, post-training INT8 quantization (PTQ) is explored to reduce model size and accelerate inference. Additionally, a simplified quantization-aware training variant (QAT-lite) is considered, where the effects of quantization are partially incorporated during fine-tuning.

Experiments are conducted on the ISIC dataset, which contains dermoscopic images of skin lesions. Model performance is evaluated using standard classification metrics, including accuracy, macro-averaged F1 score, and area under the ROC curve (ROC-AUC). CPU performance is also analyzed, including inference latency, throughput, memory consumption, and the final model size.

The results show that post-training INT8 quantization preserves performance close to the FP32 baseline while substantially reducing memory and computational requirements. In contrast, QAT-lite does not consistently provide reproducible improvements over PTQ.

***Keywords**: Vision Transformer, knowledge distillation, EMA, post-training quantization, quantization-aware training.*

REFERENCES

1. *Shamshad F., Khan S., Zamir S.W., et al.* Transformers in Medical Imaging: A Survey // arXiv. 2022.

2. *He K., Gan C., et al.* Transformers in Medical Image Analysis: A Review // arXiv. 2022.

3. *Atabansi C.C., Nie J., et al.* A Survey of Transformer Applications for Histopathological Image Analysis: New Developments and Future Directions // Biomedical Engineering Online. 2023. Vol. 22, No. 1.
https://doi.org/10.1186/s12938-023-01069-5

4. *Azad R., Kazerouni A., Heidari M., et al.* Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review // arXiv. 2023.

5. *Shamshad F., Khan S., Zamir S.W., et al.* Transformers in Medical Imaging: A Survey // Medical Image Analysis. 2024. Vol. 88.

https://doi.org/10.1016/j.media.2023.102843

6. *Liu Y., et al.* A Recent Survey of Vision Transformers for Medical Image Segmentation // arXiv. 2023.

7. *Wu F., et al.* Lite Transformer with Long-Short Range Attention // Proceedings of the International Conference on Learning Representations (ICLR). 2020.

8. *Jacob B., et al.* Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2704–2713. https://doi.org/10.1109/CVPR.2018.00286

9. *Nagel M., et al.* A White Paper on Neural Network Quantization // arXiv. 2021.

10. *Han S., Mao H., Dally W.J.* Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // arXiv. 2016.

11. *Yao Z., et al.* ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers // Advances in Neural Information Processing Systems (NeurIPS). 2022. Vol. 35.

12. *Wikipedia contributors.* Model Compression // *Wikipedia*. 2025.

13. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // arXiv. 2015.

14. *Gou J., et al.* Knowledge Distillation: A Survey // International Journal of Computer Vision. 2021. Vol. 129, No. 6. P. 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

15. *Umirzakova S., et al.* Simplified Knowledge Distillation for Deep Neural Networks: Bridging the Performance Gap with a Novel Teacher–Student Architecture // Electronics. 2024. Vol. 13, No. 3. https://doi.org/10.3390/electronics13030512

16. *Liang P., et al.* Data-Free Knowledge Distillation with Feature Synthesis and Spatial Consistency for Image Analysis // Scientific Reports. 2024. Vol. 14, No. 1. https://doi.org/10.1038/s41598-024-53241-3

СВЕДЕНИЯ ОБ АВТОРАХ

НИГМАТУЛЛИН Амир Рамисович – студент 4 курса Университета Иннополис по направлению «Искусственный интеллект», специализируется на оптимизации моделей трансформеров. Научные интересы включают эффективные архитектуры глубокого обучения, компьютерное зрение, объяснимый ИИ и обучение с подкреплением. Выпускная квалификационная работа посвящена анализу и тестированию методов оптимизации трансформеров с целью повышения эффективности и снижения вычислительных затрат. Победитель хакатона по генерации интерьеров компании Leroy Merlin с проектом в области ИИ для дизайна и визуализации пространств.

Amir Ramisovich NIGMATULLIN – 4th year student at Innopolis University with a degree in Artificial Intelligence, he specializes in optimizing transformer models. His research interests include effective deep learning architectures, computer vision, explicable AI, and reinforcement learning.The final thesis is devoted to the analysis and testing of transformer optimization methods in order to increase efficiency and reduce computational costs. The winner of the hackathon on interior generation by Leroy Merlin with a project in the field of AI for the design and visualization of spaces.

email: am.nigmatullin@innopolis.university
ORCID: 0009-0001-6884-1119

ЛУКМАНОВ Рустам Абубакирович (PhD, Бернский университет, 2021) – научный сотрудник, доцент, специализирующийся на машинном обучении, биоинформатике, анализе данных и объяснимом ИИ. Лауреат награды «Молодые лидеры БРИКС и ШОС» (2023). Преподает курсы по объяснимому ИИ и представлению знаний в Университете Иннополис.

Rustam Abubakirovich LUKMANOV (PhD, University of Bern, 2021) - is a Researcher and Associate Professor specializing in machine learning, bioinformatics, data analysis and explicable AI. Winner of the BRICS and SCO Young Leaders Award (2023). Teaches courses on explicable AI and knowledge representation at Innopolis University.

email: r.lukmanov@innopolis.university
ORCID: 0000-0001-9257-7410

ТАХА Ахмад – аспирант и научный сотрудник Центра искусственного интеллекта в Университете Иннополис. Специализируется на медицинском ИИ, самообучении (SSL) и компьютерном зрении. Его научные интересы также включают обработку естественного языка (NLP) и трансформеры. Является преподавателем на факультете ИИ.

Ahmad TAHA – is a PhD student and Researcher at the Center of Artificial Intelligence, Innopolis University. He specializes in Medical AI, Self-Supervised Learning (SSL), and Computer Vision. His research interests also include Natural Language Processing (NLP) and Transformers. He is an instructor in the AI department.

email: a.taha@innopolis.university

ORCID: 0009-0006-6346-4162

Материал поступил в редакцию 10 ноября 2025 года

УДК 004.81

АВТОМАТИЧЕСКОЕ ДОБАВЛЕНИЕ SEO-МЕТАДАННЫХ В НОВОСТНЫЕ СТАТЬИ С ИСПОЛЬЗОВАНИЕМ QWEN-CODER

Х. Салем[1] [0000-0002-9143-5231], **А. С. Тощев**[2] [0000-0003-4424-6822]

[1]Университет Иннополис, г. Иннополис, Россия
[2]Казанский (Приволжский) федеральный университет, г. Казань, Россия

[1]h.salem@innopolis.ru, [2]atoschev@kpfu.ru

Аннотация

Обобщен ранее разработанный конвейер обогащения новостных статей структурированными метаданными и представлена его обновленная конфигурация, в которой GPT-3 (Generative Pre-trained Transformer 3) – языковая модель от компании OpenAI – заменен на открытую модель Qwen-Coder. Новая версия, как и ранее, использует набор из 400 страниц, отобранных через Google News, и остается совместимой с Google Rich Results Test. Эксперименты показали, что качество, сопоставимое с GPT-3, достижимо при локальном запуске на типовом офисном настольном компьютере (CPU, без GPU). Установлено, что замена, указанная выше, снижает зависимость от платных облачных сервисов и обеспечивает более высокую производительность по сравнению с GPT-версией; дана оценка сходства результатов обогащения для Qwen-Coder относительно базовой реализации на GPT-3. Предложенные инструменты снижают порог внедрения семантической разметки и расширяют ее практическое применение, в том числе в цифровой журналистике.

Ключевые слова: *семантическая паутина, майнинг шаблонов, Qwen-Coder, новостные веб-страницы, читабельность, структурированные данные.*

ВВЕДЕНИЕ

Семантическая разметка веб-страниц позволяет представлять их содержание и метаданные в машиночитаемом виде, благодаря чему программные системы могут корректнее интерпретировать материалы и формировать расши-

ренные элементы поисковой выдачи [1–2]. В новостном сегменте такие механизмы особенно важны: агрегаторы и поисковые сервисы (включая Google News) используют структурированные метаданные в формате JSON-LD для понимания типа материала, даты публикации, авторства и других характеристик [3–4]. Однако значительная часть издателей продолжает публиковать страницы в виде «чистого» HTML без структурированной разметки, что ограничивает потенциал поисковой видимости и снижает качество представления материалов в выдаче.

Ранее в работе [5] был предложен пятиэтапный конвейер обогащения новостных страниц: он собирает статьи, извлекает и очищает основной текст и формирует корректные объекты JSON-LD, описывающие метаданные страницы. В исходной реализации ключевая операция очистки и нормализации текста выполнялась с использованием GPT-3 (Generative Pre-trained Transformer 3) – крупной языковой модели третьего поколения семейства GPT от компании OpenAI [6].

В настоящей статье рассмотрена обновленная конфигурация этого конвейера, в которой вместо GPT-3 применена Qwen-Coder – открытая языковая модель семейства Qwen, ориентированная на задачи программирования и обработки структурированных форматов (в том числе разметки и сериализаций данных) [7]. Такая замена делает решение менее зависимым от облачных платных интерфейсов, уменьшает риски, связанные с внешними ограничениями и нестабильными задержками, и позволяет использовать локальный запуск в типовой корпоративной инфраструктуре. Это важно для организаций с ограниченными ресурсами и для команд, которым требуется масштабно формировать структурированные метаданные при отсутствии специализированной экспертизы в машинном обучении [8–9]. Дополнительно переход на открытую модель поддерживает курс отрасли информационных технологий на расширение доступности языковых моделей вне подписных сервисов [7].

В статье представлены обзор литературы, описание методики и результаты сравнения полученных результатов с GPT-версией по показателям качества и производительности. Отдельно рассмотрены вопросы эксплуатации и масштабирования решения в условиях эксплуатации.

ОБЗОР ЛИТЕРАТУРЫ

Как известно, семантический веб – это подход, при котором сведения на веб-страницах описывают так, чтобы их могли однозначно интерпретировать не только люди, но и программы. Для этого применяют формальные модели представления знаний и стандартные форматы описания сущностей и их связей [3, 8, 10]. В этой области широко используются RDF (модель представления фактов в виде «субъект – связь – объект») и OWL (язык для описания онтологий, то есть набора понятий предметной области и отношений между ними) [3, 10]. В прикладных задачах веб-публикации чаще применяют JSON-LD (формат добавления «связанных данных» в виде JSON), а также альтернативную разметку RDFa в составе HTML [4, 11–12].

Для новостных сайтов структурированные метаданные важны тем, что поисковые системы могут использовать их для более точного понимания материала (тип публикации, автор, дата, рубрика, изображение и т. п.) и формирования расширенных элементов выдачи [4, 11, 12–13]. Отдельные работы показали, что корректная поисковая оптимизация (SEO – набор приемов, повышающих видимость страниц в поиске) влияет на результативность продвижения и может быть связана с бизнес-показателями [9, 14–16]. При этом поведение пользователей в поисковой выдаче (SERP – страница результатов поиска) изучается как самостоятельная тема; такие исследования показывают, какие элементы выдачи привлекают внимание и как изменяются сценарии просмотра [17].

Практическая проблема заключается в том, что многие сайты остаются «слабоструктурированными»: они публикуют корректный HTML, но не добавляют формализованные метаданные или делают это непоследовательно. Особенно заметно это у агрегаторах новостей, где материалы поступают из большого числа доменов с различными шаблонами страниц; на примере Google News в [18] подробно проанализированы эффекты нормализации и различий между источниками. Поэтому в реальных системах часто используют комбинированный подход: часть метаданных извлекают по структуре страницы, а часть – по текстовому содержанию.

Отдельное направление работ посвящено извлечению основного содержимого с страниц (заголовка и «тела» новости) из HTML-документов. Такие методы используют DOM (Document Object Model – древовидное представление элементов HTML-страницы), визуальные признаки и устойчивые шаблоны расположения блоков, чтобы отделять основной текст от меню, рекламы и служебных элементов [19–24]. Эти исследования важны для нашей задачи, поскольку качество структурированных метаданных напрямую зависит от качества выделения основного контента.

Кроме того, в ряде работ особо подчеркнуто, что автоматическая разметка требует регулярного сопровождения знаний предметной области: словарей, правил, классов сущностей и их атрибутов. Без такой поддержки система постепенно теряет качество из-за изменений в доменах, шаблонах страниц и требованиях поисковых платформ [10]. В прикладных сценариях это означает необходимость контроля качества и периодической актуализации правил формирования метаданных, а также проверки результата внешними средствами валидации, например Google Rich Results Test [13].

МЕТОДОЛОГИЯ

Мы используем исходный корпус из 1100 статей на английском и арабском языках, выбранных из 18 источников через Google News – новостного агрегатора, который объединяет материалы различных изданий [18]. Обновленный конвейер, как и ранее, выполняет следующую последовательность шагов: (1) загрузку веб-страниц; (2) удаление имеющейся разметки JSON-LD; (3) извлечение признаков из DOM; (4) формирование структурированных метаданных с использованием языковой модели и (5) проверку результата. В новой версии этап генерации выполняется моделью Qwen-Coder [7].

Для генерации метаданных использованы ключевые поля страницы: заголовок, основной текст, основное изображение и адрес страницы (URL). Эти поля формируют текстовый запрос к языковой модели, который передается в локально запущенное приложение. На выходе формируется блок JSON-LD, совместимый с общепринятыми схемами описания контента (schema.org). Полученные результаты сохраняются в кэш и проходят проверку с помощью Google Rich

Results Test [13]. Дополнительно проводится количественная проверка сохранения смысла: вычисляется расхождение Дженсена–Шеннона [25–28] между исходным и восстановленным представлениями содержимого, что позволяет контролировать семантическую эквивалентность.

Переход на Qwen-Coder потребовал адаптации развертывания под процессорный режим (CPU, без использования графического ускорителя). Для этого модель была упакована в облегченную среду выполнения и использовала квантованные веса (компактное представление параметров модели для экономии памяти), а также пакетную обработку запросов с учетом ограничений оперативной памяти. Мы сохранили те же шаблоны запросов, что применялись в конфигурации с GPT-3, чтобы остальные компоненты конвейера (проверка, хранение и обработка результатов) работали без изменений. Такой подход уменьшил риск ошибок при переходе на Qwen-Coder и позволил измерить вклад именно замены модели.

Для сопровождения процесса были добавлены показатели контроля работы приложения: время обработки запросов, доля обращений, обслуженных из кэша (показывает эффективность повторного использования результатов), а также значение расхождения Дженсена–Шеннона. Эти показатели были использованы для оперативного выявления проблем, связанных с входными данными или ограничениями аппаратной платформы.

Нагрузочное тестирование (проверка поведения системы при росте объема обработки) выполнялось в виде повторяющихся пакетных запусков в течение одного дня с постепенным увеличением масштаба: ежедневные серии по 20 страниц, еженедельные обновления по 200 страниц и архивные догрузки по 400 страниц. Каждый запуск сопровождался одинаковыми процедурами проверки результата, что позволило оценить устойчивость работы при длительной нагрузке и при разных сценариях поступления данных. В результате экспериментов были получены показатели производительности и вариативности времени обработки, которые далее использовались в сравнительном анализе.

ВОЗМОЖНОСТИ QWEN3-CODER ПО СРАВНЕНИЮ С GPT-3

Qwen3-Coder нельзя рассматривать как простую замену «один к одному» по отношению к GPT-3. Во-первых, эта модель поддерживает расширенный контекст, то есть может обрабатывать значительно больший объем входного текста в одном запросе (до 256 тыс. токенов) [7]. Во-вторых, по заявлению разработчиков, она ориентирована на широкий набор языков программирования и форматов разметки, а также поддерживает режим работы с внешними инструментами: модель может не только генерировать текст, но и выполнять последовательность действий, вызывая подключенные функции по мере необходимости [29].

Эти возможности важны для нашего конвейера: они позволяют обрабатывать более крупные пакеты статей без жесткого упрощения входного запроса и сохранять корректную работу со структурированными форматами, включая JSON LD. Для версии на GPT-3 требовалось сокращать входные данные, чтобы уложиться в ограничение на длину запроса (порядка 4–8 тыс. токенов). В результате сочетание локального развертывания, расширенного контекста и ориентации на структурированные форматы стало ключевой мотивацией миграции на Qwen-Coder [7].

РЕЗУЛЬТАТЫ

На подмножестве из 400 страниц применение Qwen-Coder обеспечило в среднем 93% сходства между обновленным и исходным текстами статьи. При этом итоговые страницы сохраняли корректное отображение материала, включая встроенные изображения и графические элементы. Локальный запуск увеличил среднее время обработки примерно на 1.5 с по сравнению с вариантами, основанными на GPT-3, однако такой скорости оказалось достаточно для выполнения ночных пакетных обработок.

Отказ от использования платных облачных сервисов позволил снизить прогнозируемые ежемесячные расходы на 68% с учетом затрат на оборудование, лицензирование и поддержку [8]. Кроме того, выполнение всех этапов обработки внутри локальной инфраструктуры упростило соблюдение внутренних требований организаций к защите и контролю данных.

Помимо усредненных показателей были рассмотрены результаты в разрезе источника публикации, языка и длины статьи. Во всех сегментах сохранялись высокие значения сходства. Более заметная вариативность наблюдалась у длинных материалов с большим числом встроенных медиа-элементов и ссылок. Ручная проверка показала, что отклонения чаще связаны с неоднородной структурой HTML-страниц разных сайтов, а не с ошибками модели; это указывает на потенциал дальнейшего улучшения правил извлечения основного содержания.

По результатам применения подхода было отмечено ускорение диагностики и устранения ошибок валидации, поскольку журналы событий и промежуточные файлы формируются и сохраняются внутри локальной инфраструктуры. Это сделало работу конвейера более прозрачной и снизило потребность во внешней поддержке, что повысило общий экономический эффект.

В табл. 1 представлены количественные показатели сравнения GPT-3 и Qwen-Coder для трех размеров пакетов. Для каждого сценария приведены среднее время выполнения и стандартное отклонение по пяти повторениям. Хотя среднее время для наименьшего пакета сопоставимо, Qwen-Coder демонстрирует значительно меньшую дисперсию, что делает ночные операции более предсказуемыми. На рис. 1 наглядно видна разница времени обработки между GPT-3 и Qwen-Coder. На рис. 2 показано сравнение дисперсии, где более стабильное время выполнения показывает Qwen-Coder.

Табл. 1. Сравнение времени обработки для GPT-3 и Qwen-Coder при различных размерах пакетов. Меньшие значения означают более быстрые или более стабильные прогоны.

Размер пакета (страниц)	Среднее время, мин (GPT-3)	Среднее время, мин (Qwen-Coder)	Стандартное отклонение, мин (GPT-3)	Стандартное отклонение, мин (Qwen-Coder)
20	18.4	17.9	4.2	1.6
200	162.7	149.3	21.5	6.8
400	339.5	301.8	48.2	11.4

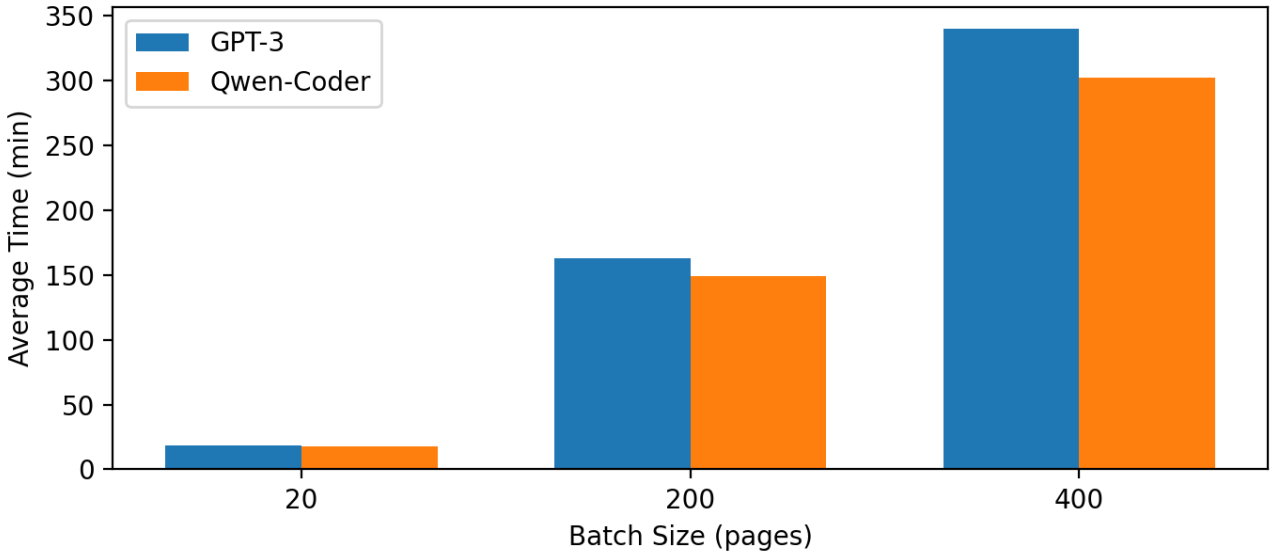


Рис. 1. Среднее время обработки для GPT-3 и Qwen-Coder при различных размерах пакетов.

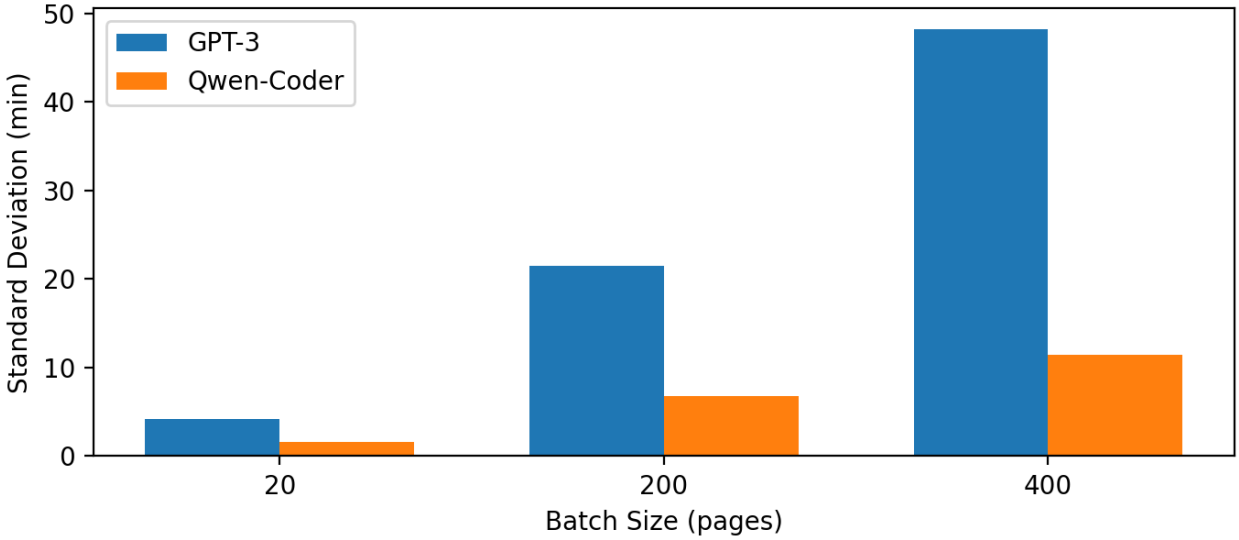


Рис. 2. Сравнение дисперсии: более стабильное время выполнения у Qwen-Coder.

Более подробный анализ данных мониторинга показал, что при локальном запуске Qwen-Coder реже возникают задержки, характерные для сетевых обращений: повторные попытки соединения и ограничения скорости со стороны удаленного сервиса, с которыми мы сталкивались в конфигурации на GPT-3. Даже при сопоставимой средней скорости обработки исчезновение редких, но

длительных задержек делает время выполнения более стабильным и упрощает планирование ночных и еженедельных пакетных запусков. Такая стабильность особенно важна для организаций, которым нужно обновлять структурированные подборки материалов до начала утреннего регионального новостного цикла.

ОГРАНИЧЕНИЯ И ПЛАНИРУЕМАЯ РАБОТА

Нами был использован корпус из 400 страниц, отобранных через Google News, который собирает публикации различных СМИ и предоставляет единый доступ к ним. Поскольку выборка ограничена материалами, попавшими в этот агрегатор, результаты могут не полностью переноситься на другие домены и типы сайтов (например, корпоративные порталы, блоги или специализированные площадки) с иной структурой страниц и правилами разметки. Мы сосредоточились на замене модели и сравнении производительности, оставив измерение долгосрочных SEO-эффектов, а также анализ потребления системой мощности для будущих исследований.

В дальнейшем планируется расширить набор данных и исследовать мультиязычные промпты. Мы также намерены экспериментировать с GPU-ускорением и измерять SEO-эффект за длительное время использования. Дополнительно включение показателей энергопотребления и доступности в систему мониторинга позволит более полно описать эксплуатационные компромиссы.

ЭКСПЛУАТАЦИОННЫЕ АСПЕКТЫ

Использование локально развернутой языковой модели предъявляет повышенные требования к эксплуатации и предполагает заранее формализованные процедуры сопровождения. Для обеспечения устойчивости при длительной работе конвейера организован регулярный контроль температуры процессора, загрузки оперативной памяти и состояния дисковой подсистемы. При приближении показателей к установленным порогам система автоматически формирует уведомления, что позволяет выполнять профилактические действия до возникновения отказов.

С учетом возможного роста вычислительной нагрузки разработан план обеспечения вычислительными ресурсами. Он предусматривает масштабирование за счет добавления вычислительных узлов, а также применение ускорения на графических процессорах в периоды пикового спроса (например, во время выборов или крупных спортивных событий). Соответствующие сценарии предварительно отрабатываются в контролируемой среде до перевода в промышленную эксплуатацию для подтверждения производительности и оценки затрат.

Отдельное внимание было уделено управлению программными зависимостями и восстановлению. Артефакты модели и конфигурационные файлы продублированы во внутреннем хранилище, что обеспечивает быстрое восстановление при отказе локального диска. Процедуры резервирования дополнительно тестировались в учебных испытаниях с имитацией сбоев инфраструктуры.

ЗАКЛЮЧЕНИЕ

Замена GPT-3 на Qwen-Coder позволила сохранить точность алгоритма добавления SEO-метаданных и повысить его производительность при работе на менее мощном оборудовании. Такая замена обеспечила существенную экономию, упростила развертывание и сохранила совместимость с алгоритмом, работающим с GPT-3. В дальнейшем планируется исследовать многоязычные шаблоны запросов и адаптивные механизмы кэширования с сохранением подхода локального запуска, который делает решение привлекательным для организаций с ограниченным бюджетом.

В более широком контексте результаты работы показали, что открытые языковые модели могут применяться в сценариях промышленной эксплуатации, которые ранее опирались на закрытые облачные сервисы. Представляя практический опыт внедрения и результаты оценки, мы рассчитываем поддержать дальнейшие исследования и разработки доступных инструментов, упрощающих применение практик семантической разметки в журналистике.

В дальнейшем планируется реализовать полуавтоматическую настройку запросов к модели, расширить языковое покрытие за пределы английского и арабского языков, а также углубить интеграцию с корпоративными системами управления контентом. Каждое изменение будет оцениваться с использованием

тех же метрик, что и в настоящей статье, чтобы повышение функциональности не снижало надежность и прозрачность результатов.

СПИСОК ЛИТЕРАТУРЫ

1. *Bashir F., Warraich N.F.* Systematic literature review of Semantic Web for distance learning // Interactive Learning Environments. 2020. Vol. 31. P. 527–543.

2. *Breit A., Waltersdorfer L., Ekaputra F.J., Sabou M., Ekelhart A., Iana A., Paulheim H., Portisch J., Revenko A., Teije A.T., et al.* Combining Machine Learning and Semantic Web: A Systematic Mapping Study // ACM Computing Surveys. 2023. Vol. 55. Art. 313.

3. *Yu L.* Introduction to the Semantic Web and Semantic Web Services. Boca Raton, FL, USA: Chapman and Hall/CRC, 2007.

4. *Sporny M., Longley D., Kellogg G., Lanthaler M., Lindström N.* JSON-LD 1.1: W3C Recommendation. 2020.

5. *Salem H., Salloum H., Orabi O., Sabbagh K., Mazzara M.* Enhancing News Articles: Automatic SEO Linked Data Injection for Semantic Web Integration // Applied Sciences. 2025. Vol. 15. Art. 1262. https://doi.org/10.3390/app15031262

6. OpenAI. GPT-3 powers the next generation of apps. 2021.
URL: https://openai.com/index/gpt-3-apps/ (дата обращения: 16.01.2026)

7. *Hui B., Yang J., Cui Z. et al.* Qwen2.5-Coder Technical Report // arXiv. 2024. arXiv:2409.12186. URL: https://arxiv.org/abs/2409.12186 (дата обращения: 10.01.2026).

8. *Shadbolt N., Berners-Lee T., Hall W.* The Semantic Web Revisited // IEEE Intelligent Systems. 2006. Vol. 21. P. 96–101.

9. *Poturak M., Keco D., Tutnic E.* Influence of search engine optimization (SEO) on business performance: Case study of private university in Sarajevo // International Journal of Research in Business and Social Science. 2022. Vol. 11. P. 59–68.

10. *Chandrasekaran B., Josephson J.R., Benjamins V.R.* What are ontologies, and why do we need them? // IEEE Intelligent Systems and Applications. 1999. Vol. 14. P. 20–26.

11. *Sporny M., Longley D., Kellogg G., Lanthaler M., Lindström N.* JSON-LD 1.0: W3C Recommendation. 2014.

12. *Adida B., Birbeck M., McCarron S., Pemberton S.* RDFa in XHTML: Syntax and processing: W3C Recommendation. 2008.

13. Rich Results Test. URL: https://search.google.com/test/rich-results (дата обращения: 08.10.2024).

14. *Iqbal M., Khalid M.N., Manzoor A.A., Malik M., Shaikh N.A.* Search Engine Optimization (SEO): A Study of important key factors in achieving a better Search Engine Result Page (SERP) Position // Sukkur IBA Journal of Computing and Mathematical Sciences. 2022. Vol. 6. P. 1–15.

15. *Alfiana F., Khofifah N., Ramadhan T., Septiani N., Wahyuningsih W., Azizah N.N., Ramadhona N.* Apply the Search Engine Optimization (SEO) Method to determine Website Ranking on Search Engines // International Journal of Cyber Services and Management. 2023. Vol. 3. P. 65–73.

16. *Mbonigaba C., Sujatha S., Kumar A.D., Vasuki M.* Leveraging Digital Channels for Customer Engagement and Sales: Evaluating SEO, Content Marketing, and Social Media for Brand Growth // International Journal of Engineering Research and Modern Education. 2024. Vol. 9. P. 32–40.

17. *Lew O.D., Kammerer Y.* Factors influencing viewing behaviour on search engine results pages: A review of eye-tracking research // Behaviour & Information Technology. 2020. Vol. 40. P. 1485–1515.

18. *Wang Q.* Normalization and Differentiation in Google News: A Multi-Method Analysis of the World's Largest News Aggregator: Thesis. Rutgers University, NJ, USA, 2020.

19. *Rahman A.F.R., Alam H., Hartono R.* Content Extraction from HTML Documents // Proceedings of the 1st International Workshop on Web Document Analysis (WDA2001). Seattle, WA, USA, 8 September 2001.

20. *Lima R., Espinasse B., Oliveira H., Pentagrossa L., Freitas F.* Information Extraction from the Web: An Ontology-Based Method Using Inductive Logic Programming // Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence. Herndon, VA, USA, 4–6 November 2013. P. 951–958.

21. *Zheng S., Song R., Wen J.-R.* Template-Independent News Extraction Based on Visual Consistency // Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver, BC, Canada, 22–26 July 2007. Washington, DC, USA: AAAI Press, 2007. P. 1507–1512.

22. *Zhu W., Dai S., Song Y., Lu Z.* Extracting news content with visual unit of web pages // Proceedings of the 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). Takamatsu, Japan, 1–3 June 2015. P. 1–5.

23. *Gupta S., Kaiser G., Neistadt D., Grimm P.* DOM-based content extraction of HTML documents // Proceedings of the 12th International Conference on World Wide Web. Budapest, Hungary, 20–24 May 2003. P. 207–214.

24. *Mirzaaghaei M., Mesbah A.* DOM-based test adequacy criteria for web applications // Proceedings of the 2014 International Symposium on Software Testing and Analysis. San Jose, CA, USA, 21–26 July 2014. P. 71–81.

25. *Lin J.* Divergence Measures Based on the Shannon Entropy // IEEE Transactions on Information Theory. 1991. Vol. 37, No. 1. P. 145–151.
https://doi.org/10.1109/18.61115

26. *Corander J., Remes U., Koski T.* On the Jensen–Shannon divergence and the variation distance for categorical probability distributions // Kybernetika. 2021. Vol. 57. P. 879–907.

27. *Nielsen F.* Jensen–Shannon divergence and diversity index: Origins and some extensions. Preprint. 2021.

28. *Menéndez M.L., Pardo J.A., Pardo L., Pardo M.C.* The Jensen–Shannon divergence // Journal of the Franklin Institute. 1997. Vol. 334. P. 307–318.

29. Qwen Team. Qwen3-Coder: GitHub repository.
URL: https://github.com/QwenLM/Qwen3-Coder (дата обращения: 11.11.2025).

AUTOMATIC ADDITION OF SEO METADATA TO NEWS ARTICLES USING QWEN-CODER

H. Salem[1] [0000-0002-9143-5231], A. S. Toshchev[2] [0000-0003-4424-6822]

[1]Innopolis University, Innopolis, Russia
[2]Kazan Federal University, Kazan, Russia

[1]h.salem@innopolis.ru, [2]atoschev@kpfu.ru

Abstract

A previously developed pipeline for enriching news articles with structured data is summarized, and an updated configuration is presented in which GPT-3–OpenAI's third-generation natural language processing model – is replaced with Qwen-Coder. As before, the updated enrichment pipeline uses a dataset of 400 pages selected from Google News, a free news aggregator by Google, remains compatible with the Google Rich Results Test (Google's tool for validating eligible structured results), and demonstrates that GPT-3-comparable output quality can be achieved on a low-power desktop PC. We describe how this substitution reduces dependence on paid GPT services and report an evaluation comparing the similarity of outputs produced by Qwen-Coder against the GPT-based baseline. The results also show higher performance of the new algorithm compared with the GPT version. The proposed tools lower the barrier to adopting semantic markup practices and thereby broaden their application in digital journalism. Overall, the findings support Qwen-Coder as a cost-effective alternative to large proprietary models for metadata enrichment tasks.

Keywords: *semantic web, pattern mining, Qwen-Coder, news web pages, readability, structured data.*

REFERENCES

1. *Bashir F., Warraich N.F.* Systematic literature review of Semantic Web for distance learning // Interactive Learning Environments. 2020. Vol. 31. P. 527–543.

2. *Breit A., Waltersdorfer L., Ekaputra F.J., Sabou M., Ekelhart A., Iana A., Paulheim H., Portisch J., Revenko A., Teije A.T., et al.* Combining Machine Learning and Semantic Web: A Systematic Mapping Study // ACM Computing Surveys. 2023. Vol. 55. Art. 313.

3. *Yu L.* Introduction to the Semantic Web and Semantic Web Services. Boca Raton, FL, USA: Chapman and Hall/CRC, 2007.

4. *Sporny M., Longley D., Kellogg G., Lanthaler M., Lindström N.* JSON-LD 1.1: W3C Recommendation. 2020.

5. *Salem H., Salloum H., Orabi O., Sabbagh K., Mazzara M.* Enhancing News Articles: Automatic SEO Linked Data Injection for Semantic Web Integration // Applied Sciences. 2025. Vol. 15. Art. 1262. https://doi.org/10.3390/app15031262

6. OpenAI. GPT-3 powers the next generation of apps. 2021.
URL: https://openai.com/index/gpt-3-apps/

7. *Hui B., Yang J., Cui Z. et al.* Qwen2.5-Coder Technical Report // arXiv. 2024. arXiv:2409.12186. URL: https://arxiv.org/abs/2409.12186

8. *Shadbolt N., Berners-Lee T., Hall W.* The Semantic Web Revisited // IEEE Intelligent Systems. 2006. Vol. 21. P. 96–101.

9. *Poturak M., Keco D., Tutnic E.* Influence of search engine optimization (SEO) on business performance: Case study of private university in Sarajevo // International Journal of Research in Business and Social Science. 2022. Vol. 11. P. 59–68.

10. *Chandrasekaran B., Josephson J.R., Benjamins V.R.* What are ontologies, and why do we need them? // IEEE Intelligent Systems and Applications. 1999. Vol. 14. P. 20–26.

11. *Sporny M., Longley D., Kellogg G., Lanthaler M., Lindström N.* JSON-LD 1.0: W3C Recommendation. 2014.

12. *Adida B., Birbeck M., McCarron S., Pemberton S.* RDFa in XHTML: Syntax and processing: W3C Recommendation. 2008.

13. Rich Results Test. URL: https://search.google.com/test/rich-results

14. *Iqbal M., Khalid M.N., Manzoor A.A., Malik M., Shaikh N.A.* Search Engine Optimization (SEO): A Study of important key factors in achieving a better Search Engine Result Page (SERP) Position // Sukkur IBA Journal of Computing and Mathematical Sciences. 2022. Vol. 6. P. 1–15.

15. *Alfiana F., Khofifah N., Ramadhan T., Septiani N., Wahyuningsih W., Azizah N.N., Ramadhona N.* Apply the Search Engine Optimization (SEO) Method to determine Website Ranking on Search Engines // International Journal of Cyber Services and Management. 2023. Vol. 3. P. 65–73.

16. *Mbonigaba C., Sujatha S., Kumar A.D., Vasuki M.* Leveraging Digital Channels for Customer Engagement and Sales: Evaluating SEO, Content Marketing, and Social Media for Brand Growth // International Journal of Engineering Research and Modern Education. 2024. Vol. 9. P. 32–40.

17. *Lew O.D., Kammerer Y.* Factors influencing viewing behaviour on search engine results pages: A review of eye-tracking research // Behaviour & Information Technology. 2020. Vol. 40. P. 1485–1515.

18. *Wang Q.* Normalization and Differentiation in Google News: A Multi-Method Analysis of the World's Largest News Aggregator: Thesis. Rutgers University, NJ, USA, 2020.

19. *Rahman A.F.R., Alam H., Hartono R.* Content Extraction from HTML Documents // Proceedings of the 1st International Workshop on Web Document Analysis (WDA2001). Seattle, WA, USA, 8 September 2001.

20. *Lima R., Espinasse B., Oliveira H., Pentagrossa L., Freitas F.* Information Extraction from the Web: An Ontology-Based Method Using Inductive Logic Programming // Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence. Herndon, VA, USA, 4–6 November 2013. P. 951–958.

21. *Zheng S., Song R., Wen J.-R.* Template-Independent News Extraction Based on Visual Consistency // Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver, BC, Canada, 22–26 July 2007. Washington, DC, USA: AAAI Press, 2007. P. 1507–1512.

22. *Zhu W., Dai S., Song Y., Lu Z.* Extracting news content with visual unit of web pages // Proceedings of the 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). Takamatsu, Japan, 1–3 June 2015. P. 1–5.

23. *Gupta S., Kaiser G., Neistadt D., Grimm P.* DOM-based content extraction of HTML documents // Proceedings of the 12th International Conference on World Wide Web. Budapest, Hungary, 20–24 May 2003. P. 207–214.

24. *Mirzaaghaei M., Mesbah A.* DOM-based test adequacy criteria for web applications // Proceedings of the 2014 International Symposium on Software Testing and Analysis. San Jose, CA, USA, 21–26 July 2014. P. 71–81.

25. *Lin J.* Divergence Measures Based on the Shannon Entropy // IEEE Transactions on Information Theory. 1991. Vol. 37, No. 1. P. 145–151.

https://doi.org/10.1109/18.61115

26. *Corander J., Remes U., Koski T.* On the Jensen–Shannon divergence and the variation distance for categorical probability distributions // Kybernetika. 2021. Vol. 57. P. 879–907.

27. *Nielsen F.* Jensen–Shannon divergence and diversity index: Origins and some extensions. Preprint. 2021.

28. *Menéndez M.L., Pardo J.A., Pardo L., Pardo M.C.* The Jensen–Shannon divergence // Journal of the Franklin Institute. 1997. Vol. 334. P. 307–318.

29. Qwen Team. Qwen3-Coder: GitHub repository.
URL: https://github.com/QwenLM/Qwen3-Coder

СВЕДЕНИЯ ОБ АВТОРАХ

САЛЕМ Хамза – аспирант, Университет Иннополис, Лаборатория программной инженерии, г. Иннополис.

Hamza SALEM – PhD student, Innopolis University, Software Engineering Lab, Innopolis.

email: h.salem@innopolis.ru

ORCID: 0000-0002-9143-5231

ТОЩЕВ Александр Сергеевич – доцент, к. н., КФУ, Институт информационных технологий и интеллектуальных систем, Кафедра программной инженерии, г. Казань.

Alexander Sergeevich TOSCHEV – Associate Professor, Ph.D., KFU, Institute of Information Technologies and Intelligent Systems, Department of Software Engineering, Kazan.

email: atoschev@kpfu.ru

ORCID: 0000-0003-4424-6822

Материал поступил в редакцию 18 ноября 2025 года

УДК 004.8:02+001.89

РОЛЬ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СОЗДАНИИ, КУРИРОВАНИИ И ИНТЕРПРЕТАЦИИ КОЛЛЕКЦИЙ ЭЛЕКТРОННЫХ БИБЛИОТЕК

Е. В. Самоходкин[1] [0000-0003-3791-0123], А. А. Эльзон[2] [0000-0003-3524-434X], Е. Г. Самоходкина[3] [0000-0002-3162-3097], Д. В. Лошадкин[4] [0000-0002-8963-2586]

[1–4] *Всероссийский институт научной и технической информации РАН, г. Москва, Россия*

[1]rodentforme@gmail.com, [2]alisaelzon@gmail.com, [3]slava-eugen@yandex.ru, [4]loshadkindv@hotmail.com

Аннотация

Исследование посвящено осмыслению роли искусственного интеллекта (ИИ) в трансформации экосистемы цифровой научной коммуникации на материале электронных библиотек и крупных агрегаторов знаний. На основе интегративного обзора новейших зарубежных и отечественных работ проанализировано, как ИИ постепенно превращается в системообразующий инфраструктурный механизм жизненного цикла электронных коллекций, структурируя процессы отбора, оцифровки, метадатирования, хранения и сервисного раскрытия ресурсов. Параллельно обоснована интерпретация интеллектуальных рекомендательных систем как эпистемического посредника, влияющего на конфигурацию научного чтения, распределение исследовательского внимания и видимость периферийных знаний в пространственно-языковой архитектуре науки. Показано, что алгоритмическая персонализация не сводится к повышению удобства поиска, а участвует в конструировании норм релевантности, языковых и региональных иерархий, новых принципов осмысления коллекций. Выявленные эффекты позволяют концептуализировать феномен алгоритмического посредничества в связке микроуровня исследовательской идентичности и макроуровня глобального распределения научного знания, а также обозначить необходимость рефлексивного управления рекомендательными контурами в целях сохранения эпистемического многообразия и повышения прозрачности цифровой инфраструктуры библиотек.

Ключевые слова*: искусственный интеллект, электронные библиотеки, рекомендательные системы, алгоритмическое посредничество, цифровая*

научная коммуникация, жизненный цикл электронных коллекций, метаданные, эпистемический медиатор, пространственно-языковая конфигурация знания, периферийные знания, исследовательская идентичность, алгоритмическая персонализация, библиометрический анализ, когнитивный менеджмент, культурное наследие.

ВВЕДЕНИЕ

Актуальность заявленной тематики определяется совмещением трех взаимосвязанных процессов: взрывным ростом объема цифровых данных, ускоренной цифровизацией библиотечных фондов и стремительным, но крайне неравномерным внедрением инструментов искусственного интеллекта(ИИ) в информационную инфраструктуру. По оценкам отраслевых исследований [1], глобальный объем данных к 2024 г. достиг порядка 149 зеттабайт, причем динамика носит экспоненциальный характер и напрямую связана с повсеместной цифровизацией научной, образовательной и культурной деятельности. На этом фоне рынок цифровых библиотек демонстрирует устойчивый рост: при совокупном объеме около 3.69 млрд долларов США в 2021 г. прогнозируется увеличение до 5.46 млрд долларов уже к 2025 г., с последующим среднегодовым темпом роста порядка 10.3% до 2033 г. [2]. Таким образом, электронные коллекции перестают быть вспомогательным ресурсом и превращаются в ключевой канал доступа к знаниям, что радикально усложняет задачи их формирования, поддержания целостности и интерпретации без опоры на интеллектуальные алгоритмы анализа.

Дополнительным индикатором трансформации служит изменение структуры пользовательского спроса. Исследования читательской активности показывают, что только в 2020 г. через публичные библиотеки было осуществлено около 428 млн заимствований электронных объектов (электронных книг, аудиокниг, цифровых журналов), причем значения фиксируются преимущественно на платформах удаленного доступа [3]. Согласно сводному технологическому обзору Public Library Association за 2023 г. [4], не менее 95% библиотек предоставляют пользователям доступ к электронным книгам и аудиокнигам, около 58% – к потоковым и загружаемым медиа, а четверть учреждений уже располагает оборудованием для цифрового медиапроизводства, при этом 40% предлагают

maker-инфраструктуру (прим. – совокупность материально-технических и организационных ресурсов, которые обеспечивают деятельность пользователей в логике maker movement, то есть в логике «сделай-сам»-культуры, ориентированной на конструирование, прототипирование, цифровое творчество и малосерийное производство). В совокупности эти тенденции демонстрируют, что цифровые коллекции перестраиваются из статичного «зеркала» печатного фонда в динамическую многомодальную экосистему, где объем, разнообразие форматов и интенсивность использования требуют перехода от традиционных библиотечных практик к моделям управления, основанным на ИИ, машинном обучении и аналитике больших данных. При этом общесистемный контекст внедрения ИИ задается процессами, выходящими далеко за пределы профессионального сообщества библиотекарей. Отчет Организации экономического сотрудничества и развития (ОЭСР – Organisation for Economic Co-operation and Development, OECD) [5] о ранних разрывах в распространении ИИ фиксирует, что доля предприятий Европейского союза, использующих ИИ, выросла с 8% в 2023 г. до 13.5% в 2024 г., причем в ряде стран (Эстония, Швеция, Греция, Норвегия) наблюдается более чем двукратное увеличение показателей за один год. Генеративные модели, снижающие порог входа и требования к технической компетентности, выводят ИИ на уровень массового инструмента, а не специализированной технологии. Нормативные документы профессиональных объединений прямо подчеркивают необходимость осмысленной позиции библиотек в данном процессе: в заявлении IFLA о библиотеках и ИИ (2020 г.) особо выделяются риски для интеллектуальной свободы, равенства доступа и приватности, а в разработанном в 2023 г. руководстве по стратегическому реагированию на ИИ отмечается необходимость формирования у библиотек роль не пассивного потребителя, а активного медиатора и критического интерпретатора алгоритмических решений [6, 7]. В результате библиотеки оказываются в ситуации, когда игнорировать ИИ уже невозможно, а стихийное внедрение без осмысленной концепции работы с коллекциями чревато усилением информационного неравенства и непрозрачности доступа к знаниям.

Мотивация к интеграции ИИ исходит и изнутри библиотечной практики. По данным все того же обследования PLA [4], 95% публичных библиотек реализуют программы цифровой грамотности, при этом среди учреждений, ответивших

на вопрос о новых тематических направлениях обучения, 61% указали ИИ как приоритетную сферу развития образовательных инициатив для населения. Инфраструктурная база при этом остается неоднородной: около 74.8% библиотек подключены к оптоволоконным сетям и 99.4% предоставляют Wi-Fi-доступ, однако 28.4% организаций подписаны на интернет-каналы, не достигающие новых федеральных стандартов широкополосной связи (100/20 Мбит/с), а в сельских и малых городах аналогичный показатель достигает только 35.4%. В академическом сегменте показательно масштабное исследование внедрения ИИ в университетских библиотеках Китая, охватившее 154 респондента из государственных и частных учреждений [8]: было зафиксировано, что 48.1% библиотек уже используют технологии преобразования текста в речь, наоборот, и речи в текст, 41.6 % внедрили голосовой поиск, 35.7% эксплуатируют RFID-системы, а расширение тексто-звуковых сервисов планируется более чем в половине библиотек (51.9%). При этом 86.9% опрошенных считают, что ИИ повышает точность поиска, 88.3% отмечают эффект автоматизации задач, 59.1% поддерживают использование чат-ботов, но 85.7% фиксируют недостаточную поддержку со стороны университетской администрации, 69.5% указывают на высокую стоимость технологий, 57.8% – на нехватку технологических ресурсов, а 66.9% – на существенные этико-правовые риски. Таким образом, вырисовывается противоречивая картина: даже при очевидной утилитарной выгоде алгоритмических решений библиотеки сталкиваются с институциональными и инфраструктурными барьерами; следовательно, требуются не просто точечное внедрение отдельных приложений, а продуманная модель включения ИИ в полный жизненный цикл электронных коллекций – от отбора и оцифровки до интерпретации и предоставления.

Тенденция к институционализации ИИ в библиотечной сфере подтверждается данными библиометрических исследований. В обзоре AI-инноваций и сервисов в академических библиотеках на основе выборки Scopus было проанализировано 922 публикации, из которых после многоступенчатого отбора сформирован корпус из 189 релевантных статей за период 2014–2024 г.; регрессионный анализ показал, что динамика роста числа работ лучше описывается экспоненциальной моделью ($R^2 = 0.964$), что интерпретируется как стадия ускоренного развития темы [9]. В другом исследовании, ориентированном на применение ИИ именно в академических библиотеках, с использованием запросов по ключевым словам

“Artificial Intelligence”, “Machine Learning”, “Chatbots” и др. из базы Scopus за 2010–2023 г. было извлечено 484 записи, среди которых 191 публикация классифицирована как журнальная статья, 158 – как материалы конференций, плюс ряд обзоров и глав в книгах [10]. Отмечены не только рост суммарного числа работ, но и тематическое расширение – от автоматизированного индексирования и рекомендательных систем до поддержки исследовательской деятельности и цифрового сохранения. При этом тематическое картирование, выполняемое в данных обзорах, демонстрирует концентрацию публикаций вокруг сервисных функций (чатботы, интеллектуальный поиск, пользовательская поддержка), тогда как вопросы, связанные с самим статусом библиотечных коллекций как данных для обучения моделей и объектов алгоритмической интерпретации, остаются существенно менее разработанными.

Сдвиг исследовательского фокуса в сторону культурного наследия и цифровых гуманитарных наук создает дополнительное измерение актуальности. Библиометрический анализ литературы по цифровым гуманитарным исследованиям и культурному наследию, выполненный на данных Scopus за 2015–2022 г., выявил 228 исходных источников, из которых после очистки было сохранено 194 статьи; годовая динамика демонстрирует рост с 24 работ в 2015 г. до пика в 41 публикацию в 2020 г., при общем числе цитирований порядка 650 и средней цитируемости около 22 ссылок на документ [11]. В обзорной работе М. Фиоруччи и соавторов [12] подчеркнуто, что сложность, разнородность и объем данных культурного наследия требуют систематического применения методов машинного обучения и компьютерного зрения – от анализа изображений и трехмерных моделей до обработки древних текстов. Концептуально важным шагом становится работа Б. С. Г. Ли [13], предложившего “Collections as ML Data” чек-лист для проектов машинного обучения на материале коллекций GLAM-институтов, где коллекция осмысливается как набор данных с явными характеристиками качества, смещения, правовых ограничений и возможностей повторного использования. Параллельно разрабатываются методики публикации коллекций «как данные» в музеях, архивах и библиотеках [14]. В результате электронные фонды начинают восприниматься не только как объект человеческого чтения, но и как тренировочная среда для алгоритмов, а решения о структуре, аннотировании и политике доступа приобретают двоякий характер: гуманитарный и машинно-ориентированный.

Технологический потенциал ИИ в управлении электронными коллекциями иллюстрируется и прикладными исследованиями в области рекомендательных систем цифровых библиотек. В работе [15], посвященной персонализации чтения на основе совмещения поведенческих характеристик пользователей и их социальных связей, предложена модель, использующая механизмы внимании и капсульные представления. На эмпирическом материале цифровой библиотеки было показано, что модель достигает точности рекомендаций 97.24% в сценариях подбора научно-популярной литературы, демонстрирует значение Recall 0.198 и Precision 0.062 для задач рекомендаций при длительных периодах заимствования, а также обеспечивает минимальные значения среднеквадратичной и средней абсолютной ошибок (RMSE = 0.731; MAE = 0.721) по сравнению с альтернативными подходами. Тем самым подтверждается возможность радикального повышения релевантности и персонализации доступа к документам за счет ИИ-моделей, однако используемые метрики по-прежнему ориентируются преимущественно на количественные показатели и не затрагивают вопросы объяснимости, сохранения тематического разнообразия, предотвращения алгоритмической фильтрации «неудобного» знания. Дополняют эту картину результаты исследований по ИИ-грамотности в библиотеках: интервью с участниками образовательных кругов показывают, что специалисты склонны рассматривать ИИ как инструмент повышения доступности и управляемости коллекций, одновременно выражая обеспокоенность влиянием алгоритмов на медийный ландшафт и испытывая трудности при включении ИИ-тематики в обучение информационной грамотности пользователей. Данные наблюдения свидетельствуют о том, что на уровне практики уже формируется запрос на модели, которые позволят интегрировать ИИ в работу с коллекциями без утраты критической рефлексии и профессиональных ценностей.

В целом приведенные количественные и качественные показатели формируют сложный контур актуальности рассматриваемой проблематики. С одной стороны, наблюдается стремительное наращивание объемов цифрового контента, почти повсеместное внедрение электронных ресурсов в библиотечные фонды и экспоненциальный рост научного интереса к ИИ в библиотечной и гуманитарной сфере. С другой стороны, отчеты ОЭСР указывают, что средний уровень использования ИИ в бизнесе остается относительно невысоким (13.5% предприятий в ЕС в

2024 г.), при этом ускоренный рост сопровождается формированием «разрывов по умолчанию» между регионами, секторами и типами организаций. Аналогичные разрывы фиксируются и в библиотеках: значительная доля учреждений не обладает доступом к современным сетевым ресурсам, сталкивается с дефицитом финансирования, кадров и компетенций, а также с этическими и правовыми ограничениями, связанными с обработкой пользовательских данных и алгоритмической фильтрацией контента. На этом фоне роль ИИ в создании, курировании и интерпретации коллекций электронных библиотек перестает быть сугубо технологическим вопросом и приобретает статус ключевой исследовательской и практической задачи, от решения которой зависит, будут ли цифровые фонды функционировать как прозрачная, подотчетная и инклюзивная инфраструктура знаний или превратятся в непрозрачные «черные ящики» алгоритмического посредничества. Поэтому в рамках предлагаемого исследования представляется необходимым комплексный анализ моделей применения ИИ на всех стадиях жизненного цикла электронных коллекций – от отбора и структурирования до алгоритмической интерпретации и пользовательского интерфейса – с опорой на эмпирические данные, нормативные рамки и гуманитарную экспертизу.

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Методологические рамки настоящей работы основаны на сочетании анализа и синтеза, что позволяет рассматривать экосистему цифровой научной коммуникации и электронных библиотек как динамичный объект, трансформируемый внедрением технологий ИИ на всех этапах жизненного цикла. В аналитическом контуре произведено разложение моделей применения ИИ в процессах создания, курирования и интерпретации коллекций электронных библиотек с опорой на интегративный обзор публикаций из наукометрических баз данных Scopus и Web of Science, а также на разбор репрезентативных кейсов внедрения (включая систему Lib2Life). На синтетическом уровне осуществлена сборка выявленных эффектов в целостное представление о роли ИИ как системообразующего инфраструктурного механизма, который не только автоматизирует рутинные операции, но и задает новые форматы существования фондов в логике «коллекций как данных» и влияет на эпистемическое многообразие.

Методологические ограничения определены вторичным характером исследования, базирующегося на анализе литературы и кейсов без проведения собственных экспериментов или разработки программного обеспечения; ИИ-системы рассмотрены как функциональный класс алгоритмических «черных ящиков», без технического аудита проприетарных моделей, что фокусирует внимание на интерпретации наблюдаемых эффектов, а не на реконструкции внутренней архитектуры. Дополнительно рамки исследования ограничены преобладанием англоязычных источников и отчетов профильных ассоциаций (IFLA, ALA, OECD), что может оставить вне поля зрения локальные практики и неформализованные режимы использования ИИ в электронных библиотеках, тем самым придав полученным выводам характер аналитико-концептуальной, а не исчерпывающе эмпирической картины.

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК СИСТЕМООБРАЗУЮЩИЙ ИНФРАСТРУКТУРНЫЙ МЕХАНИЗМ ЖИЗНЕННОГО ЦИКЛА ЭЛЕКТРОННЫХ КОЛЛЕКЦИЙ

Искусственный интеллект в контексте электронных библиотек все отчетливее позиционируется не как совокупность локальных сервисов, а как системообразующий инфраструктурный механизм, организующий жизненный цикл электронных коллекций – от отбора и цифрового захвата ресурсов до их описания, раскрытия и аналитического использования. На подобную трансформацию указывают данные новейших библиометрических и текстовых исследований. В текстовом анализе [16] проведено картирование 285 документов из базы Scopus за 2015–2024 г.; показано, что корпус работ по ИИ в библиотечной сфере сконцентрирован вокруг двух устойчивых полюсов: сервисного (128 публикаций) и инфраструктурно-технологического (149 публикаций), при этом резкий «всплеск» терминов, связанных с ChatGPT и приложениями ИИ, зафиксирован в 2023–2024 г., что трактуется как переход к фазе интенсивной интеграции ИИ в библиотечную экосистему. Параллельно scientometric-анализ [17], выполненный на материале 1462 публикаций 2012–2021 г., продемонстрировал экспоненциальный характер роста исследований об использовании ИИ в библиотеках: установлено около 5400 авторов при среднем показателе порядка четырех соавторов на документ, а также доминирование открытого доступа, что интерпретируется как формирование

междисциплинарного, кооперативного поля, в котором библиотечные системы рассматриваются как естественные полигоны для внедрения алгоритмических решений.

Дополнительное подтверждение инфраструктурного статуса ИИ дает систематический обзор [18], в рамках которого были отобраны и проанализированы 88 статей об ИИ-поддержанных библиотечных сервисах, опубликованных с 2002 по апрель 2024 г.: подчеркнуто, что значительная часть корпуса посвящена не отдельным «витринным» проектам, а автоматизации ключевых операций обслуживания, что фактически переносит акцент на перестройку всей операционной архитектуры библиотек. Сопоставимые выводы сделаны в систематическом обзоре исследований по применению ИИ и машинного обучения в библиотеках [19]: из более широкого массива записей в Web of Science, Scopus, LISA и LISTA были отобраны 32 статьи и показано, что даже при доминировании теоретических и концептуальных работ они в основном посвящены переосмыслению фундаментальных библиотечно-информационных процессов, а не точечным экспериментам, что также указывает на смещение тематик в сторону инфраструктурной логики внедрения. На макроуровне такая тенденция подтверждена библиографическим картированием [20], где на основе 4233 статей Web of Science за 2011 – июнь 2024 г. установлено, что в последнем периоде (2021–2024 гг.) вокруг понятий “artificial intelligence” и “machine learning” формируется ключевой кластер, устойчиво сопряженный с узлами “academic library”, “information literacy” и “librarian”, что фактически фиксирует встраивание ИИ в саму проблематику библиотечной информатики и управленческих решений. В совокупности данные наблюдения позволяют заключить, что в научном дискурсе последних лет ИИ закрепляется не как периферийный инструмент улучшения отдельных сервисов, а как ядро инфраструктуры, перераспределяющее функции по всей цепочке обращения цифровых ресурсов и задающее новую конфигурацию жизненного цикла электронных коллекций.

На начальной стадии жизненного цикла электронных коллекций – при отборе, оцифровке и первичной структуризации ресурсов – ИИ выступает ядром технологического конвейера. Показателен пример платформы Lib2Life, разработанной для консорциума университетских библиотек Румынии [21]. Архитектура системы выстроена вокруг непрерывного потока данных: физические книги и

периодика XVIII–XXI вв. сканируются, затем к полученным образам применяется оптическое распознавание текста; далее задействуется специализированный конвейер обработки, включающий автоматическое определение границ абзацев, слияние переносов, исправление орфографических ошибок, выделение изображений и таблиц, а также согласование структурных элементов документа. После этого текстовые блоки индексируются в Elasticsearch, разрезаются на абзацы и преобразуются в векторные представления; поверх данного уровня строятся семантический поиск и система поиска похожих документов, опирающиеся на модели обработки естественного языка и онтологию Lib2Life, которая используется для автоматической предметной классификации и наполнения графа знаний. Тем самым ИИ-модели оказываются встроенными в сами операции превращения физического документа в машиночитаемый объект, задавая формат, гранулярность сегментации и глубину семантического описания коллекции уже на этапе ее «рождения» в цифровой среде.

Следующий слой инфраструктуры составляет ИИ-центричная экосистема метаданных, обеспечивающая устойчивость и управляемость электронных коллекций. В крупном обзоре по ИИ-поддерживаемому метадатированию библиотечных и архивных ресурсов [22] выделен целый спектр тематических блоков: влияние ИИ на метаданные и рабочие процессы, платформы для автоматизированной генерации описаний, оценка качества и точности алгоритмически созданных записей, использование генеративных моделей для проектирования метаданных, подготовка данных к машинной обработке, а также этико-правовые аспекты. Показано, что системы, подобные разработанной платформе для архивных метаданных, позволяют существенно сокращать объем ручной работы при одновременном повышении согласованности полей, а проекты масштабного обогащения метаданных в секторе культурного наследия демонстрируют способность ИИ увеличивать охват и глубину описания без пропорционального роста затрат. В то же время в проведенных исследованиях отмечены опасения специалистов по поводу смещения, унаследованного из исторических каталогизационных практик, а также подчеркнута необходимость гибридной модели, при которой автоматизированное описание сопровождается человеко-ориентированным контролем, пересмотром политик и внедрением процедур прозрачной валидации.

Системный характер роли ИИ особенно отчетливо проявляется в архитектурах управления метаданными, где алгоритмические компоненты охватывают все стадии жизненного цикла описания. В работе [23], посвященной современным системам управления метаданными, описана модульная модель, включающая блоки извлечения и генерации метаданных, поиска и раскрытия, контроля качества, хранения и индексирования, а также управления политиками и безопасностью. Модуль извлечения и генерации опирается на методы машинного и глубинного обучения для выделения характеристик ресурсов из структурированных, полуструктурированных и неструктурированных источников; модуль поиска и раскрытия использует обработку естественного языка, семантический поиск и интеллектуальную фильтрацию; контур контроля качества включает автоматизированные механизмы проверки на соответствие стандартам, обнаружение несогласованностей и обогащение записей; модуль хранения обеспечивает масштабируемое индексирование и связность данных. В результате ИИ-компоненты перестают быть локальными вспомогательными сервисами и превращаются в связующее «техническое основание», на котором строятся процессы каталогизации, навигации и аналитики электронных коллекций.

Таким образом, в комплексе рассмотренные исследования и практические решения демонстрируют, что ИИ уже функционирует в качестве системообразующего инфраструктурного механизма жизненного цикла электронных коллекций: на этапе первичного отбора и цифрового захвата ресурсов он задает формат их машиночитаемого существования; в метаданных и системах управления описанием выступает связующим техническим основанием, обеспечивающим согласованность, масштабируемость и операционную управляемость фондов; на уровне архитектуры сервисов и управленческих решений он перераспределяет ключевые функции между человеком и машиной. Описанные тенденции позволяют заключить, что в современных электронных библиотеках ИИ перестает быть периферийным инструментом автоматизации и превращается в структурный элемент инфраструктуры, определяющий конфигурацию процессов создания, организации и аналитического использования цифровых ресурсов.

ВЛИЯНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ПРОСТРАНСТВЕННО-ЯЗЫКОВУЮ КОНФИГУРАЦИЮ НАУЧНОЙ КОММУНИКАЦИИ И ВИДИМОСТЬ ПЕРИФЕРИЙНЫХ ЗНАНИЙ

Искусственный интеллект в контексте электронных библиотек все отчетливее позиционируется не только как технологический инструмент обработки данных, но как полноценный эпистемический медиатор, через который структурируется доступ к знанию, конструируются интерпретации и закрепляются иерархии смыслов. В международных докладах (см., например, [24]) подчеркнуто, что интеграция алгоритмических систем в культурные экосистемы трансформирует не только способы сохранения и распространения культурного наследия, но и сами когнитивные режимы работы с коллекциями: раскрывается возможность масштабного анализа, реконструкции и переосмысления источников, однако одновременно усиливаются риски стандартизации, смещения акцентов в пользу доминирующих культурных нарративов и эрозии эпистемического многообразия. В докладе группы CULTAI при ЮНЕСКО особо отмечается, что ИИ, будучи встроенным в процессы описания, поиска и курирования, способен как расширять доступ к культурному разнообразию, так и воспроизводить предвзятости, усиливать культурную гомогенизацию и подрывать культурные права, если алгоритмическая инфраструктура выстроена вокруг узкого набора источников и интересов.

В аналитических работах по культурному наследию ИИ все чаще описывается через категорию «эпистемического цикла», в котором решения о том, что оцифровывается, как описывается и каким образом агрегируется, формируют скрытую, но устойчивую систему допущений. В исследовании [25] продемонстрировано, что смещения в музейных и архивных коллекциях воспроизводятся в обучающих выборках, затем в платформах и интерфейсах, где ИИ «масштабирует» уже имеющиеся перекосы, превращая их в норму по умолчанию для миллионов пользователей. Введено понятие «цифрового культурного колониализма», когда национальные и институциональные агрегаторы закрепляют эпистемологию доминирующих сообществ, а алгоритмическая персонализация закрепляет бинарные и центрированные оптики как естественные. Авторами указанной работы подчеркнута необходимость явного документирования и интерпретации данных

эпистемических выборов на каждом этапе ИИ-конвейера: от формирования датасета до развертывания моделей – как условия ответственного использования ИИ в сфере культурного наследия.

Параллельно формируется другая, самостоятельная линия исследований, в которой предлагается рассматривать ИИ как носитель имплицитных культурных теорий. В работе [26] выдвинут аргумент о необходимости эпистемического анализа тех культурных моделей, которые «зашиваются» в архитектуру алгоритмов, выбор метрик качества и дизайн интерфейсов, поскольку без такого анализа ИИ оказывается ориентирован на редукцию сложных культурных феноменов к упрощенным, статистически удобным категориям. В концепции «культурного ИИ» и «медленных библиотек», предложенной в [27], библиотеки и другие GLAM-институты рассматриваются как пространства, где должен сознательно замедляться темп технологического внедрения, чтобы обеспечить согласование алгоритмических решений с локальными ценностями, дискуссиями о реституции и исторической ответственности. Авторами показано, что именно библиотеки, опираясь на традицию рефлексивного каталогизирования и публичного обсуждения стандартов описания, могут стать площадкой, где алгоритмы рассматриваются как эпистемические акторы, подлежащие регулированию наравне с человеческими практиками интерпретации.

В отечественной библиотековедческой повестке также наблюдается смещение акцентов от чисто операционного понимания ИИ к трактовке его как когнитивного посредника между фондом и пользователем. В исследовании [28], посвященном когнитивному менеджменту и ИИ в библиотеках, ИИ-системы рассмотрены в логике системно-функционального подхода как инструменты управления библиотечным фондом, пользовательским обслуживанием и исследовательской деятельностью, способные реконструировать паттерны обращения к коллекциям, формировать динамические модели информационных потребностей и поддерживать персонализированные траектории чтения. Показано, что автоматизированная каталогизация, рекомендательные сервисы и чат-боты создают дополнительные уровни интерпретации фонда, где документ помещается не только в иерархию классификационных индексов, но и в сеть ассоциативных и поведенческих связей, отражающих реальные практики использования. Одновременно подчеркнута необходимость критического отношения к точности алгоритмов,

рискам предвзятости и проблемам конфиденциальности, поскольку именно через эти сервисы библиотека незаметно для пользователя конструирует образ релевантного знания и задает «норму» чтения.

Работа [29], в которой проанализировано использование нейросетевых технологий в библиографическом обслуживании, дополняет данную картину за счет эмпирического обзора российских и зарубежных практик. На основе мониторинга сайтов библиотек и анализа профессиональной литературы продемонстрировано, что нейросетевые модели все активнее используются для индексирования, тематического и фактографического поиска, справочного обслуживания, персонализированного информирования, а также для реализации диалоговых сервисов на базе чат-ботов. Показано, что перечисленные инструменты фактически переводят библиографическое обслуживание из плоскости «ручного» подбора документов в режим многоуровневой семантической фильтрации, где ИИ не только ускоряет поиск, но и интерпретирует запрос, дополняя его скрытыми ассоциациями и предугадывая интересы пользователя. В результате формируется новая конфигурация посредничества: библиограф становится куратором и критиком алгоритмических предложений, а сама библиографическая запись – гибридным продуктом совместной работы человека и машины, закрепляющим в своей структуре определенную трактовку содержания документа, его связей и значимости.

Итак, на основе рассмотренных исследований можно заключить, что ИИ в библиотечно-информационной среде уже функционирует как эпистемический медиатор, радикально влияющий на конфигурацию доступа к знаниям, распределение внимания и закрепление смысловых иерархий. Через алгоритмическое описание, персонализированный поиск, рекомендательные механизмы и нейросетевое библиографическое обслуживание формируется новая, гибридная структура посредничества, в рамках которой библиограф и библиотекарь выступают не только операторами сервисов, но и кураторами, критически переосмысляющими алгоритмические решения и отвечающими за баланс между расширением доступности и сохранением эпистемического многообразия. В результате этот технологический слой начинает определять видимость периферийных знаний и пространственно-языковую архитектуру научной коммуникации, что требует осознанного, рефлексивного управления алгоритмическими контурами интерпретации коллекций.

ЗАКЛЮЧЕНИЕ

Проведенный анализ позволяет заключить, что ИИ-инструменты в сфере электронных библиотек приобретают двойственный, но внутренне связанный статус: с одной стороны, формируется системообразующий инфраструктурный каркас жизненного цикла электронных коллекций, охватывающий стадии отбора, цифрового захвата, метадатирования, хранения и сервисного раскрытия ресурсов; с другой – укрепляется роль алгоритмического посредничества как эпистемического механизма, перераспределяющего исследовательское внимание и влияющего на видимость периферийных знаний в глобальной экосистеме научной коммуникации, что созвучно выявленной в библиометрических исследованиях зависимости между динамикой научного интереса к ИИ и публикационной активностью [30]. Показано, что рекомендательные системы и связанные с ними аналитические контуры не ограничиваются повышением оперативности обслуживания, а фактически участвуют в конструировании норм релевантности, языково-пространственных иерархий и новых режимов интерпретации коллекций.

Сделанный акцент на инфраструктурном и эпистемическом измерениях алгоритмической персонализации позволяет по-новому обозначить ответственность библиотек и крупных агрегаторов знаний: управляя ИИ как элементом технологической инфраструктуры, эти институции одновременно управляют распределением доступа к знаниям, условиями присутствия маргинализированных тематик и малых языков, а также профилем исследовательской идентичности пользователей, в том числе в контексте обсуждаемых в современной литературе социальных угроз ИИ и необходимости их институционализированного контроля [31]. Выявленный в работе разрыв между высокой степенью технологической интеграции и недостаточной разработанностью механизмов прозрачности, подотчетности и защиты эпистемического многообразия свидетельствует о необходимости перехода от интуитивных, ситуативных решений к рефлексивным стратегиям проектирования и регулирования алгоритмического посредничества в научном чтении.

Перспективным направлением дальнейших исследований представляется углубленное эмпирическое изучение влияния рекомендательных систем на практики чтения, цитирования и формирование доверия к научным и библиотечным

институтам, включая сопоставление алгоритмических траекторий с коммуникационными инструментами формирования лояльности к научным организациям [32]. Не менее значимым видится развитие методик аудита и объяснимости алгоритмов [33], ориентированных на выявление языковых, региональных и тематических смещений, а также концептуализация принципов инклюзивного рекомендательного дизайна, поддерживающего видимость периферийных знаний. Дополнительного внимания требует сравнительный анализ влияния ИИ на национальные и локальные экосистемы научной коммуникации, включая «серые» репозитории и малые архивы, что позволит преодолеть обозначенные методологические ограничения и перейти от преимущественно обзорной перспективы к многомерной, эмпирически насыщенной картине алгоритмического посредничества в глобальном и региональном масштабе.

СПИСОК ЛИТЕРАТУРЫ

1. *Bartley K.* Big data statistics: How much data is there in the world? // Rivery Blog. 2025. URL: https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world (дата обращения: 03.09.2025).

2. *Dharmadhikari S.* Digital Library Market Report 2025 (Global Edition). Cognitive Market Research, 2025. URL: https://www.cognitivemarketresearch.com/digital-library-market-report (дата об-ращения: 10.09.2025).

3. *Fox-Sowell S.* Public libraries are alive and well, thanks to Gen Z, millennials and the shift to digital collections // StateScoop. 12 March 2024.
URL: https://statescoop.com/public-libraries-alive-well-gen-z-millennials-digital-collections (дата обращения: 22.09.2025).

4. *Public Library Association.* 2023 Public Library Technology Survey: Sum-mary Report. Chicago: Public Library Association, 2024. 57 p.
URL: https://www.ala.org/sites/default/files/2024-07/PLA_Tech_Survey_Report_2024.pdf (дата обращения: 05.10.2025).

5. *Kergroach S., Héritier J.* Emerging divides in the transition to artificial intelligence. OECD Regional Development Papers. Paris: OECD Publishing, 2025. No. 147. https://doi.org/10.1787/7376c776-en.

URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/emerging-divides-in-the-transition-to-artificial-intelligence_eeb5e120/7376c776-en.pdf (дата обращения: 19.10.2025).

6. *IFLA FAIFE.* IFLA Statement on Libraries and Artificial Intelligence. The Hague: International Federation of Library Associations and Institutions, 2020.
URL: https://repository.ifla.org/items/8c05d706-498b-42c2-a93a-3d47f69f7646 (дата обращения: 28.10.2025).

7. *Cox A.* Developing a library strategic response to artificial intelligence. Working document. The Hague: International Federation of Library Associations and Institutions, 2023.
URL: https://www.ifla.org/developing-a-library-strategic-response-to-artificial-intelligence (дата обращения: 04.11.2025).

8. *Li D.* Adoption of Artificial Intelligence in Public and Private Libraries of China: Determinants, Challenges, and Perceived Benefits // Profesional de la infor-mación. 2024. Vol. 33, No. 4. e330416. https://doi.org/ 10.3145/epi.2024.ene.0416

9. *Kulkanjanapiban P., Silwattananusarn T., Lambovska M.* Research on AI-driven innovations and services in academic libraries: A bibliometric and systematic literature review // Journal of Data and Information Science. 2025. Vol. 10, no. 3. P. 1–51. https://doi.org/10.2478/jdis-2025-0036

10. *Islam M.N., Ahmad S., Aqil M., Hu G., Ashiq M., Abusharhah M.M., Saky S.A.T.M.* Application of artificial intelligence in academic libraries: a bibliometric analysis and knowledge mapping // Discover Artificial Intelligence. 2025. Vol. 5. Art. 59. https://doi.org/10.1007/s44163-025-00295-9.

11. *Othman R.* Charting Digital Humanities: A Bibliometric View of Cultural Heritage // European Proceedings of Social and Behavioural Sciences (EpSBS). 2023. P. 42–58. https://doi.org/10.15405/epsbs.2023.11.4

12. *Fiorucci M., Khoroshiltseva M., Pontil M., Traviglia A., Del Bue A., James S.* Machine learning for cultural heritage: a survey // Pattern Recognition Letters. 2020. Vol. 133. P. 102–108. https://doi.org/10.1016/j.patrec.2020.02.017

13. *Lee B.C.G.* The “Collections as ML Data” checklist for machine learning and cultural heritage // Journal of the Association for Information Science and Technology. 2025. Vol. 76, No. 2. P. 375–396. https://doi.org/10.1002/asi.24765

14. *Candela G., Gabriëls N., Chambers S. et al.* A checklist to publish collections as data in GLAM institutions // Global Knowledge, Memory and Communication. 2023. Vol. 74, No. 56. P. 1323–1355.

15. *Andersdotter K.* Artificial intelligence literacy in libraries: Experiences and critical impressions from a learning circle // Journal of Information Literacy. 2023. Vol. 17, No. 2. P. 108–130. https://doi.org/10.11645/17.2.14

16. *Santosa F.A.* Artificial Intelligence in Library Studies: A Textual Analysis // JLIS.It. 2025. Vol. 16, No. 1. P. 61–71. https://doi.org/10.36253/jlis.it-626

17. *Borgohain D.J., Bhardwaj R.K., Verma M.K.* Mapping the literature on the application of artificial intelligence in libraries (AAIL): a scientometric analysis // Library Hi Tech. 2024. Vol. 42, No. 1. P. 149–179. https://doi.org/10.1108/LHT-07-2022-0331

18. *Gunasekera D., Senevirathne W.A.R.* Application of Artificial Intelligence for Library Services: A Systematic Literature Review // Journal of the University Librarians Association of Sri Lanka. 2024. Vol. 27, No. 2. P. 257–284. https://doi.org/10.4038/jula.v27i2.8089

19. *Das R.K., Islam M.S.U.* Application of Artificial Intelligence and Machine Learning in Libraries: A Systematic Review // arXiv preprint. 2021. arXiv:2112.04573. https://doi.org/10.48550/arXiv.2112.04573

20. *Park Y., Kim S.* Research Trends on Information Technology and Artificial Intelligence for Libraries Using Bibliographic Mapping // Journal of the Korean Biblia Society for Library and Information Science. 2024. Vol. 35, No. 4. P. 45–65. https://doi.org/10.14699/kbiblia.2024.35.4.045

21. *Nitu M., Dascalu M., Dascalu M.D., Neagu L.-M., Dascalu M.-I.* Lib2Life – Digital Library Services Empowered with Advanced Natural Language Processing Techniques // Interaction Design and Architecture(s) Journal – IxD&A. 2024. No. 60. P. 147–167. https://doi.org/10.55612/s-5002-060-006

22. *Kanaujia Sukula S.* AI-Assisted Metadata and Intelligent Catalogues: Redefining Knowledge Organization in Libraries // International Journal of Library Information Network and Knowledge. 2025. Vol. 10, No. 2. P. 86–101.

23. *Yang W., Fu R., Amin M.B., Kang B.* The Impact of Modern AI in Metadata Management // Human-Centric Intelligent Systems. 2025. Vol. 5, No. 3. P. 323–350. https://doi.org/10.1007/s44230-025-00106-5

24. *United Nations Educational, Scientific and Cultural Organization (UNESCO).* Report of the Independent Expert Group on Artificial Intelligence and Culture. Paris: UNESCO, 2025. URL: https://vk.cc/cRBDkt (дата обращения: 11.11.2025).

25. *Foka A., Griffin G., Ortiz Pablo D. et al.* Tracing the bias loop: AI, cultural heritage and bias-mitigating in practice // AI & Society. 2025. Vol. 40, No. 8. P. 5835–5847. https://doi.org/10.1007/s00146-025-02349-z

26. *Mansouri M.* A call for epistemic analysis of cultural theories for AI methods // AI & Society. 2023. Vol. 38, No. 2. P. 969–971. https://doi.org/10.1007/s00146-022-01465-4

27. *Breemen K., Breemen V.* 'Slow libraries' and 'Cultural AI': Reassessing technology regulation in the context of digitised cultural heritage data // Technology and Regulation. 2025. P. 175–193. https://doi.org/10.71265/fxkhy005

28. *Каптерев А.И.* Когнитивный менеджмент и искусственный интеллект в библиотеках: возможности и особенности // Научные и технические библиотеки. 2023. № 6. С. 113–137. https://doi.org/10.33186/1027-3689-2023-6-113-137

29. *Нещерет М.Ю.* Нейросети в библиотеке: новое в библиографическом обслуживании // Научные и технические библиотеки. 2024. № 1. С. 105–128. https://doi.org/10.33186/1027-3689-2024-1-105-128

30. *Самоходкин Е. В., Эльзон А. А.* Анализ взаимосвязи научного интереса и динамики публикаций по искусственному интеллекту в Российской Федерации (2020–2024 гг.) // Научно-техническая информация. Сер. 1: Организация и методика информационной работы. 2025. № 6. С. 10–17.https://doi.org/10.36535/0548-0019-2025-06-2

31. *Самоходкин Е.В., Эльзон А.А.* Социальные угрозы искусственного интеллекта: классификация и оценка опасности // Научно-техническая информация. Сер. 1: Организация и методика информационной работы. 2025. № 9. С. 12–21. https://doi.org/10.36535/0548-0019-2025-09-2

32. *Тимохович А.Н., Самоходкин Е.В., Эльзон А.А.* Исследование коммуникационных инструментов для формирования лояльности к научной организации // Социология науки и технологий. 2025. Т. 16, № 3. С. 201–221. https://doi.org/10.24412/2079-0910-2025-3-201-221

33. *Самоходкин Е.В., Эльзон А.А.* Подходы к формализации нормативного поведения автономных интеллектуальных агентов // Искусственный интеллект и принятие решений. 2025. № 4. С. 35–46. https://doi.org/10.14357/20718594250403

THE ROLE OF ARTIFICIAL INTELLIGENCE IN CREATION, CURATION AND INTERPRETATION OF DIGITAL LIBRARY COLLECTIONS

E. V. Samokhodkin[1] [0000-0003-3791-0123], A. A. Elzon[2] [0000-0003-3524-434X], E. G. Samokhodkina[3] [0000-0002-3162-3097], D. V. Loshadkin[4] [0000-0002-8963-2586]

[1–4] *All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences*

[1]rodentforme@gmail.com, [2]alisaelzon@gmail.com, [3]slava-eugen@yandex.ru, [4]loshadkindv@hotmail.com

Abstract

This study examines the role of artificial intelligence (AI) in reshaping the ecosystem of digital scholarly communication, drawing on evidence from electronic libraries and large-scale knowledge aggregators. On the basis of an integrative review of recent international and Russian-language scholarship, AI is shown to be gradually evolving into a system-forming infrastructural mechanism across the life cycle of electronic collections, structuring processes of selection, digitisation, metadata creation, storage, and service-oriented resource discovery and access. In parallel, intelligent recommender systems are substantiated as an epistemic mediator influencing the configuration of scholarly reading, the distribution of research attention, and the visibility of peripheral forms of knowledge within the spatial–linguistic architecture of science. It is demonstrated that algorithmic personalisation cannot be reduced to improved search convenience; rather, it participates in constructing relevance norms, linguistic and regional hierarchies, and new regimes for interpreting collections. The effects identified make it possible to conceptualise algorithmic mediation at the intersection of the micro level of research identity and the macro level of the global distribution of scholarly

knowledge, while also underscoring the need for reflexive governance of recommender loops in order to preserve epistemic diversity and enhance the transparency of libraries' digital infrastructures.

Keywords: *artificial intelligence, electronic libraries, recommender systems, algorithmic mediation, digital scholarly communication, life cycle of electronic collections, metadata, epistemic mediator, spatial–linguistic architecture of science, peripheral knowledge, research identity, algorithmic personalisation, bibliometric analysis, cognitive management, cultural heritage.*

REFERENCES

1. *Bartley K.* Big data statistics: How much data is there in the world? // Rivery Blog. 2025. URL: https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world (accessed September 3, 2025).

2. *Dharmadhikari S.* Digital Library Market Report 2025 (Global Edition). Cognitive Market Research, 2025. URL: https://www.cognitivemarketresearch.com/digital-library-market-report (accessed September 10, 2025).

3. *Fox-Sowell S.* Public libraries are alive and well, thanks to Gen Z, millennials and the shift to digital collections // StateScoop. 12 March 2024.
URL: https://statescoop.com/public-libraries-alive-well-gen-z-millennials-digital-collections (accessed September 22, 2025).

4. *Public Library Association.* 2023 Public Library Technology Survey: Sum-mary Report. Chicago: Public Library Association, 2024. 57 p.
URL: https://www.ala.org/sites/default/files/2024-07/PLA_Tech_Survey_Report_2024.pdf (accessed October 3, 2025).

5. *Kergroach S., Héritier J.* Emerging divides in the transition to artificial intelligence. OECD Regional Development Papers. Paris: OECD Publishing, 2025. No. 147. https://doi.org/10.1787/7376c776-en.
URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/emerging-divides-in-the-transition-to-artificial-intelligence_eeb5e120/7376c776-en.pdf (accessed October 19, 2025).

6. *IFLA FAIFE.* IFLA Statement on Libraries and Artificial Intelligence. The Hague: International Federation of Library Associations and Institutions, 2020.

URL: https://repository.ifla.org/items/8c05d706-498b-42c2-a93a-3d47f69f7646 (accessed October 19, 2025).

7. *Cox A.* Developing a library strategic response to artificial intelligence. Working document. The Hague: International Federation of Library Associations and Institutions, 2023.
URL: https://www.ifla.org/developing-a-library-strategic-response-to-artificial-intelligence (accessed November 04, 2025).

8. *Li D.* Adoption of Artificial Intelligence in Public and Private Libraries of China: Determinants, Challenges, and Perceived Benefits // Profesional de la infor-mación. 2024. Vol. 33, No. 4. e330416. https://doi.org/ 10.3145/epi.2024.ene.0416

9. *Kulkanjanapiban P., Silwattananusarn T., Lambovska M.* Research on AI-driven innovations and services in academic libraries: A bibliometric and systematic literature review // Journal of Data and Information Science. 2025. Vol. 10, no. 3. P. 1–51. https://doi.org/10.2478/jdis-2025-0036

10. *Islam M.N., Ahmad S., Aqil M., Hu G., Ashiq M., Abusharhah M.M., Saky S.A.T.M.* Application of artificial intelligence in academic libraries: a bibliometric analysis and knowledge mapping // Discover Artificial Intelligence. 2025. Vol. 5. Art. 59. https://doi.org/10.1007/s44163-025-00295-9.

11. *Othman R.* Charting Digital Humanities: A Bibliometric View of Cultural Heritage // European Proceedings of Social and Behavioural Sciences (EpSBS). 2023. P. 42–58. https://doi.org/10.15405/epsbs.2023.11.4

12. *Fiorucci M., Khoroshiltseva M., Pontil M., Traviglia A., Del Bue A., James S.* Machine learning for cultural heritage: a survey // Pattern Recognition Letters. 2020. Vol. 133. P. 102–108. https://doi.org/10.1016/j.patrec.2020.02.017

13. *Lee B.C.G.* The “Collections as ML Data” checklist for machine learning and cultural heritage // Journal of the Association for Information Science and Technology. 2025. Vol. 76, No. 2. P. 375–396. https://doi.org/10.1002/asi.24765

14. *Candela G., Gabriëls N., Chambers S. et al.* A checklist to publish collections as data in GLAM institutions // Global Knowledge, Memory and Communication. 2023. Vol. 74, No. 56. P. 1323–1355.

15. *Andersdotter K.* Artificial intelligence literacy in libraries: Experiences and critical impressions from a learning circle // Journal of Information Literacy. 2023. Vol. 17, No. 2. P. 108–130. https://doi.org/10.11645/17.2.14

16. *Santosa F.A.* Artificial Intelligence in Library Studies: A Textual Analysis // JLIS.It. 2025. Vol. 16, No. 1. P. 61–71. https://doi.org/10.36253/jlis.it-626

17. *Borgohain D.J., Bhardwaj R.K., Verma M.K.* Mapping the literature on the application of artificial intelligence in libraries (AAIL): a scientometric analysis // Library Hi Tech. 2024. Vol. 42, No. 1. P. 149–179. https://doi.org/10.1108/LHT-07-2022-0331

18. *Gunasekera D., Senevirathne W.A.R.* Application of Artificial Intelligence for Library Services: A Systematic Literature Review // Journal of the University Librarians Association of Sri Lanka. 2024. Vol. 27, No. 2. P. 257–284. https://doi.org/10.4038/jula.v27i2.8089

19. *Das R.K., Islam M.S.U.* Application of Artificial Intelligence and Machine Learning in Libraries: A Systematic Review // arXiv preprint. 2021. arXiv:2112.04573. https://doi.org/10.48550/arXiv.2112.04573

20. *Park Y., Kim S.* Research Trends on Information Technology and Artificial Intelligence for Libraries Using Bibliographic Mapping // Journal of the Korean Biblia Society for Library and Information Science. 2024. Vol. 35, No. 4. P. 45–65. https://doi.org/10.14699/kbiblia.2024.35.4.045

21. *Nitu M., Dascalu M., Dascalu M.D., Neagu L.-M., Dascalu M.-I.* Lib2Life – Digital Library Services Empowered with Advanced Natural Language Processing Techniques // Interaction Design and Architecture(s) Journal – IxD&A. 2024. No. 60. P. 147–167. https://doi.org/10.55612/s-5002-060-006

22. *Kanaujia Sukula S.* AI-Assisted Metadata and Intelligent Catalogues: Redefining Knowledge Organization in Libraries // International Journal of Library Information Network and Knowledge. 2025. Vol. 10, No. 2. P. 86–101.

23. *Yang W., Fu R., Amin M.B., Kang B.* The Impact of Modern AI in Metadata Management // Human-Centric Intelligent Systems. 2025. Vol. 5, No. 3. P. 323–350. https://doi.org/10.1007/s44230-025-00106-5

24. *United Nations Educational, Scientific and Cultural Organization (UNESCO).* Report of the Independent Expert Group on Artificial Intelligence and Culture. Paris: UNESCO, 2025. URL: https://vk.cc/cRBDkt (accessed November 11, 2025).

25. *Foka A., Griffin G., Ortiz Pablo D. et al.* Tracing the bias loop: AI, cultural heritage and bias-mitigating in practice // AI & Society. 2025. Vol. 40, No. 8. P. 5835–5847. https://doi.org/10.1007/s00146-025-02349-z

26. *Mansouri M.* A call for epistemic analysis of cultural theories for AI methods // AI & Society. 2023. Vol. 38, No. 2. P. 969–971. https://doi.org/10.1007/s00146-022-01465-4

27. *Breemen K., Breemen V.* 'Slow libraries' and 'Cultural AI': Reassessing technology regulation in the context of digitised cultural heritage data // Technology and Regulation. 2025. P. 175–193. https://doi.org/10.71265/fxkhy005

28. *Kapterev A.I.* Kognitivnyj menedzhment i iskusstvennyj intellekt v bibliotekah: vozmozhnosti i osobennosti // Nauchnye i tekhnicheskie biblioteki. 2023. No. 6. P. 113–137. https://doi.org/10.33186/1027-3689-2023-6-113-137

29. *Neshcheret M.Yu.* Nejroseti v biblioteke: novoe v bibliograficheskom obsluzhivanii // Nauchnye i tekhnicheskie biblioteki. 2024. No. 1. P. 105–128. https://doi.org/10.33186/1027-3689-2024-1-105-128

30. *Samokhodkin E.V., Elzon A.A.* Analysis of the Relationship Between Scientific Interest and Publication Dynamics on Artificial Intelligence in the Russian Federation (2020–2024) // Scientific and Technical Information Processing. 2025. Vol. 52, No. 2. P. 152–160. https://doi.org/10.3103/S0147688225700182

31. *Samokhodkin E.V., Elzon A.A.* Social Threats of Artificial Intelligence: Classification and Risk Assessment // Scientific and Technical Information Processing. 2025. Vol. 52. P. 263–271. https://doi.org/10.3103/S0147688225700674

32. *Timokhovich A.N., Samokhodkin E.V., Elzon A.A.* Issledovanie kommunikatsionnyh instrumentov dlya formirovaniya loyal'nosti k nauchnoj organizatsii // Sotsiologiya nauki i tekhnologij. 2025. Vol. 16, No. 3. P. 201–221. https://doi.org/10.24412/2079-0910-2025-3-201-221

33. *Samokhodkin E.V., Elzon A.A.* Podkhody k formalizatsii normativnogo povedeniya avtonomnykh intellektual'nykh agentov // Iskusstvennyi intellekt i prinyatie reshenii. 2025. No. 4. P. 35–46. https://doi.org/10.14357/20718594250403

СВЕДЕНИЯ ОБ АВТОРАХ

САМОХОДКИН Евгений Вячеславович – Аспирант. Ведущий специалист Центра маркетинговых исследований и перспективного планирования Федерального государственного бюджетного учреждения науки «Всероссийский институт научной и технической информации Российской академии наук». Области исследований: искусственный интеллект, публикационная активность, маркетинг.

SAMOKHODKIN Evgeniy Vyacheslavovich – PhD student, Lead Specialist at the Center for Marketing Research and Strategic Planning, Federal State Budgetary Scientific Institution «All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences». Research interests: artificial intelligence, publication activity, marketing.

rodentforme@gmail.com

ORCID 0000-0003-3791-0123

ЭЛЬЗОН Алиса Андреевна – Аспирант. Ведущий специалист Центра маркетинговых исследований и перспективного планирования Федерального государственного бюджетного учреждения науки «Всероссийский институт научной и технической информации Российской академии наук». Области исследований: искусственный интеллект, публикационная активность, маркетинг.

ELZON Alisa Andreevna – PhD student, Lead Specialist at the Center for Marketing Research and Strategic Planning, Federal State Budgetary Scientific Institution «All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences». Research interests: artificial intelligence, publication activity, marketing.

alisaelzon@gmail.com

ORCID 0000-0003-3524-434X

САМОХОДКИНА Елена Геннадьевна – Главный специалист Группы баз данных Отдела информационных ресурсов Федерального государственного бюджетного учреждения науки Всероссийский институт научной и технической информации Российской академии наук. Области исследований: информетрия, библиометрия, наукометрия.

SAMOKHODKINA Elena Gennadievna – Chief Specialist, Database Group, Department of Information Resources, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences (Federal State Budgetary Institution of Science). Research areas: informetrics, bibliometrics, scientometrics.

slava-eugen@yandex.ru

ORCID 0000-0002-3162-3097

ЛОШАДКИН Дмитрий Владимирович – Кандидат химических наук, старший научный сотрудник, заместитель генерального директора по развитию ООО «Эверс Груп Рус». Области исследований: разработка и внедрение новых конструкционных материалов, патентоведение, фрактальный анализ, наукометрия, библиометрия.

LOSHADKIN Dmitrii Vladimirovich – PhD in Chemistry, Senior Research Fellow, Deputy General Director for Development at Evers Group Rus LLC. Research areas: development and implementation of advanced structural materials, patent studies, fractal analysis, scientometrics, and bibliometrics.

loshadkindv@hotmail.com

ORCID 0000-0002-8963-2586

Материал поступил в редакцию 15 декабря 2025 года

УДК 004.942:624.04

РАСЧЕТ СТЕРЖНЕВЫХ ЭЛЕМЕНТОВ С ТРЕЩИНАМИ НА ОСНОВЕ СОЧЕТАНИЯ ТЕОРИИ СТЕРЖНЕЙ И ТЕОРИИ УПРУГОСТИ

М. Н. Серазутдинов[0000-0003-1675-3819]

Казанский национальный исследовательский технологический университет, г. Казань, Россия

serazmn@mail.ru

Аннотация

Представлены математические модели для расчета напряженно-деформированного состояния стержней с трещинами при деформациях растяжения-сжатия и изгибе. В работе использовано сочетание соотношений теории упругости и теории стержней. Основные положения предложенного метода моделирования основаны на разделении стержня на фрагменты и нахождении для каждого из выделенных фрагментов деформаций и напряжений на базе теории стержней или теории упругости. Описаны алгоритмы расчетов, которые сравнительно просты в реализации. Для подтверждения достоверности и точности расчетов, основанных на предложенных моделях, представлены результаты решения задач.

Ключевые слова*: стержень, балка, трещина, напряженно-деформированное состояние, формула Мора, теория упругости, теория стержней.*

ВВЕДЕНИЕ

В эксплуатируемых стержневых конструкциях от воздействия различных техногенных и климатических факторов могут возникать повреждения и дефекты. В частности, часто образуются локальные изменения размеров поперечных сечений стержней, которые оказывают существенное влияние на надежность сооружения. К таким повреждениям относятся и трещины.

Для минимизации последствий образования трещины нужно ее обнаружить (идентифицировать) и оценить последствия ее влияния на напряженно-деформированное состояние (НДС).

В настоящее время идентификация трещин часто проводится на основе анализа частотных характеристик стержней. Обзор работ по этой теме представлен в статье [1]. В работах [2–10] описаны математические модели, разработанные для обнаружения трещин и определения их геометрии.

НДС, возникающее в сравнительно простых элементах конструкций с трещинами, рассчитывается с использованием математических моделей, основанных на соотношениях теории упругости. Известны работы, в которых предложены модели расчета деформированного состояния балок с трещинами, в которых используются теория стержней и моментная теория упругости [11–13].

В работах [14–17] представлены математические модели для определения НДС балок с трещинами, основанные на гипотезах теории стержней. Трещина в этих моделях заменяется деформируемым элементом, механические характеристики которого назначаются по результатам дополнительных исследований.

В [18–20] для определения деформаций в стержнях с различными дефектами в геометрии поперечных сечений предложено моделировать участки с дефектами стержнем постоянного сечения и для нахождения его геометрический характеристик применять уравнения теории упругости.

В настоящей статье представлена математическая модель для расчета НДС стержней с трещинами. Модель базируется на синтезе соотношений теории упругости и теории стержней. Учтены следующие особенности:

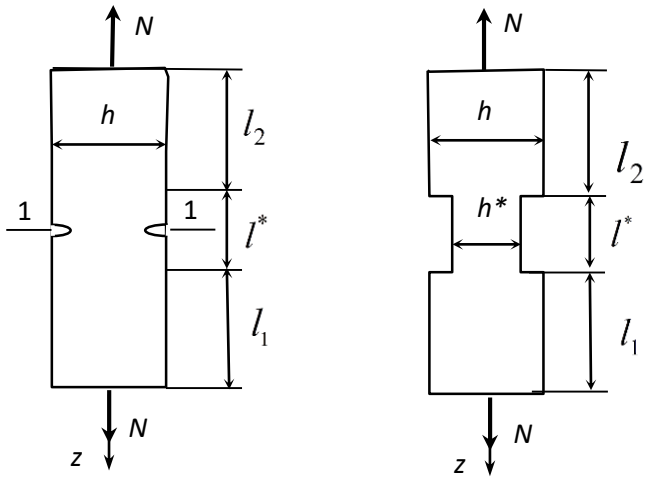
– в окрестности трещины для определения НДС нельзя применять приближенные теории, нужно использовать соотношения теории упругости;

– длина элементов стержневой системы значительно больше размеров поперечных сечений, поэтому в областях, удаленных от трещины, для определения НДС можно использовать приближенные методы теории сопротивления материалов.

Кроме того, описаны результаты расчета стержней с трещиной при деформации растяжения-сжатия и изгибе.

СТЕРЖЕНЬ С ТРЕЩИНОЙ ПРИ РАСТЯЖЕНИИ И СЖАТИИ

Выделим в стержне некоторый участок, содержащий в сечении 1–1 трещины (рис. 1а). Полагаем, что на этом участке продольная сила *N* постоянна. Определим изменение длины выделенного участка. Так как трещины ослабляют сечение, для нахождения продольной деформации полагаем, что в окрестности сечения 11, на участке длиной l^*, ширина поперечного сечения стержня h^* постоянна и ее размер меньше исходной величины h (рис. 1).



а б

Рис. 1. Фрагмент стержня с трещиной (а); схема стержня при моделировании трещины (б).

Полагаем, что величина изменения длины Δl стержня, показанного на рис. 1б, равна изменению длины выделенного фрагмента стержня с трещинами (рис. 1а). Для определения неизвестных параметров l^*, h^* используем условие

$$\Delta l_a^{\text{ту}} = \Delta l_б^{\text{см}}. \qquad (1)$$

Здесь $\Delta l_a^{\text{ту}}$ – изменение длины фрагмента стержня, показанного на рис. 1а (индекс «ту»); $\Delta l_б^{\text{см}}$ – изменения длины стержней, показанных соответственно на рис. 1б (индекс «см»). Величина $\Delta l_a^{\text{ту}}$ определяется на основе соотношений теории упругости, $\Delta l_б^{\text{см}}$ вычисляется по формулам сопротивления материалов.

Записав эти величины в виде суммы деформаций по участкам, длины которых равны l_1 , l^*, l_2 (рис. 1), получим

$$\Delta l_a^{\text{ту}} = \Delta l_1^{\text{ту}} + \Delta l_*^{\text{ту}} + \Delta l_3^{\text{ту}} \ ; \quad \Delta l_б^{\text{см}} = \Delta l_1^{\text{см}} + \Delta l_*^{\text{см}} + \Delta l_3^{\text{см}} . \qquad (2)$$

Подставив выражения (2) в равенство (1), найдем

$$\Delta l_1^{\text{ту}} + \Delta l_*^{\text{ту}} + \Delta l_3^{\text{ту}} = \Delta l_1^{\text{см}} + \Delta l_*^{\text{см}} + \Delta l_3^{\text{см}} . \qquad (3)$$

Если выбрать участки с длинами l_1 и l_2 достаточно удаленными от трещины, то при постоянных размерах поперечных сечений продольные деформации, вычисленные по теории стержней и теории упругости, будут различаться незначительно. Поэтому можно считать, что $\Delta l_1^{\text{ту}} = \Delta l_1^{\text{см}}$, $\Delta l_2^{\text{ту}} = \Delta l_2^{\text{см}}$. Следовательно, равенство (3) упрощается и записывается в виде

$$\Delta l_*^{\text{ту}} = \Delta l_*^{\text{см}} . \qquad (4)$$

Получилось, что продольная деформация $\Delta l_*^{\text{см}}$ участка стержня с размерами l^* и h^*, вычисленная по формуле теории стержней, должна быть равной удлинению $\Delta l_*^{\text{ту}}$ участка стержня с трещиной, вычисленному по соотношениям теории упругости.

С учетом $\Delta l_*^{\text{см}} = \dfrac{N l^*}{E A^*}$ из уравнения (4) получим

$$\Delta l_*^{\text{ту}} = \frac{N l^*}{E A^*} . \qquad (5)$$

Здесь *E* – модуль упругости; A^* – площадь поперечного сечения. Для сечения в виде прямоугольника шириной *b* получим $A^* = h^* \cdot b$.

Так как участки с длинами l_1 и l_2 достаточно удалены от трещины, в (5) можно положить, что $N = \sigma_0 A$, где σ_0 – нормальные напряжения, равномерно распределенные по сечению с площадью $A = h \cdot b$. Учитывая, что $N = \sigma_0 A$, а $A^* = h^* \cdot b = \dfrac{h^* \cdot h \cdot b}{h} = \dfrac{h^* \cdot A}{h}$, формулу (5) можно представить в виде

$$\Delta l_*^{\text{ту}} = \frac{\sigma_0 h \, l^*}{E h^*} . \qquad (6)$$

На рис. 2 представлен фрагмент стержня с трещинами глубиной *a*, для которого нужно определить $\Delta l_*^{\text{ту}}$.

Как известно, перемещения, возникающие при деформации тела, прямо пропорциональны действующим внешним напряжениям и обратно пропорциональны модулю упругости материала. Поэтому можно считать

$$\Delta l_{*}^{\text{ту}} = \frac{\sigma_0}{E} \Delta l_{1*}^{\text{ту}}, \tag{7}$$

где $\Delta l_{1*}^{\text{ту}}$ – продольная деформация участка стержня, показанного на рис. 2, при $\sigma_0 = E$. Отметим, что $\Delta l_{1*}^{\text{ту}}$ вычисляется с использованием соотношений теории упругости.

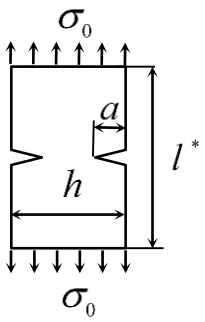


Рис. 2. Фрагмент стержня для расчета по теории упругости

С учетом формулы (7) равенство (6) примет вид

$$\frac{l^*}{h^*} = \frac{\Delta l_{1*}^{\text{ту}}}{h}. \tag{8}$$

Используя условие (8), можно определить отношение параметров l^* и h^*, т. е. найти размер участка l^* (рис. 1b), введенного для моделирования трещины.

Поскольку значение $\Delta l_{1*}^{\text{ту}}$ находится при $\sigma_0 = E$, значения параметров σ_0 и E не влияют на результаты расчетов, т. е. величина $\Delta l_{1*}^{\text{ту}}$ не зависит от модуля упругости материала и напряжений, возникающих в окрестности трещины. Для определения изменения длины полосы с трещинами при растяжении и сжатии отношение параметров l^* / h^* можно найти без учета значений напряжений, возникающих на некотором удалении от трещины. Следовательно, если по соотношениям теории упругости вычислить удлинение только фрагмента стержня, то по формулам сопротивления материалов можно определить деформацию стержня в целом.

Представленная модель может быть использована и при расчете статически неопределимой системы. В такой системе при возникновении трещины происходит перераспределение усилий и перемещений. В этом случае следует проводить расчеты в два этапа. На первом этапе, с использованием предлагаемой модели, нужно определить перераспределение усилий и перемещений стержневой системы с трещиной. По теории упругости находится величина $\Delta l_{1^*}^{\text{ту}}$ для небольшого фрагмента стержня (рис. 2) и определяются параметры l^* и h^*. Затем на базе теории стержней вычисляются перемещения в статически неопределимой системе. На втором этапе расчетов, с учетом полученных расчетных данных, по соотношениям теории упругости можно определить НДС фрагмента с трещиной. При этом в качестве граничных условий для расчета указанного фрагмента можно использовать значения перемещений, вычисленные на первом этапе.

БАЛКА С ТРЕЩИНОЙ ПРИ ИЗГИБЕ

Рассмотрим плоский изгиб балки. Представим математическую модель определения НДС балки при наличии в ней трещины. Основная особенность этого случая изгиба стержня заключается в том, что для адекватного описания деформированного состояния в некоторой окрестности трещины теория стержней неприменима, и поэтому нужно использовать соотношения теории упругости.

Выделим фрагмент *AD* балки, содержащий трещину на участке *BC* (рис. 3).

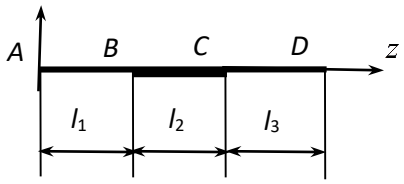


Рис. 3. Фрагмент балки

Положим, что трещина достаточно удалена от сечений, проходящих через точки *B* и C, поэтому на *AB* и *CD* НДС балки можно описать, с помощью классической теории стержней. Из-за наличия трещины на *BC* для расчетов необходимо использовать соотношения теории упругости.

Обозначим через $v_1^{\text{см}}(z_1)$, $v_2^{\text{ту}}(z_2, y)$, $v_3^{\text{см}}(z_3)$ прогибы балки на участках *AB*, *BC*, *CD* соответственно. Здесь $z_1 = z$, $z_2 = z - l_1$, $z_2 = z - (l_1 + l_2)$ – компоненты локальных систем координат, начала которых находятся соответственно в точках *A*, *B* и *C*. Верхний индекс «см» показывает, что функция получена с использованием классической теории стержней (сопротивления материалов), индекс «ту» обозначает функцию, определенную по соотношениям теории упругости. Функция $v_2^{\text{ту}}(z_2, y)$, в отличие от $v_1^{\text{см}}(z_1)$ и $v_3^{\text{см}}(z_3)$, зависит от двух переменных z_2 и y, следовательно, будет изменяться в направлении продольной оси и по высоте поперечного сечения балки. Это связано с тем, что рассматривается плоский изгиб балки, и из-за наличия трещины на участке *BC* перемещения следует определять по теории упругости.

Согласно классической теории стержней уравнение изгиба балки на участках *AB* и *CD* можно представить в следующем виде:

$$\frac{d^2 v_i^{\text{см}}(z_i)}{dz_i^2} = -\frac{M_{xi}(z_i)}{EI_{xi}}, \quad i = 1, 3. \tag{9}$$

Здесь EI_{xi} – изгибная жесткость, $M_{xi}(z_i)$ – изгибающий момент, $v_i^{см}(z)$ – прогиб. При действии на балку сосредоточенных моментов и сил, а также равномерно распределенных нагрузок изгибающий момент можно записать в виде $M_{xi}(z_i) = M_{0i} + (M_{1i}/l_i)\, z_i + (M_{2i}/l_i^2)\, z_i^2$, где M_{0i}, M_{1i}/l_i, M_{2i}/l_i^2 – слагаемые, которые находятся с учетом внешних сил и реакций опор, действующих на балку.

Проинтегрировав уравнение (9), получим

$$v_i^{\text{см}}(z_i) = f_i^{\text{см}}(z_i) + c_{1i}^{\text{см}}\, z_i + c_{2i}^{\text{см}}, \quad i = 1, 3. \tag{10}$$

Здесь $f_i^{\text{см}}(z_i) = -\left[M_{0i}\, z_i^2 + (M_{1i}/3l_i)\, z_i^3 + (M_{2i}/6l_i^2)\, z_i^4 \right] / (2\, EI_{xi})$, $c_{1i}^{\text{см}}$, $c_{2i}^{\text{см}}$ – постоянные интегрирования.

Если $v_2^{\text{ту}}(z_2, y)$ на участке *BC* находить на основе классической схемы теории упругости, то необходимо в расчетную схему внести условия связи для перемещений на границах участков *AB, BC* и *CD*. Это усложняет алгоритм решения и существенно увеличивает размерность решаемой системы уравнений. С учетом этой особенности далее будем использовать специально разработанный метод

определения $v_2^{\text{ту}}(z_2, y)$ на *BC* по теории упругости. При этом условия связи перемещений на границах участков стержней не будут заранее (при решении уравнений теории упругости) вводиться.

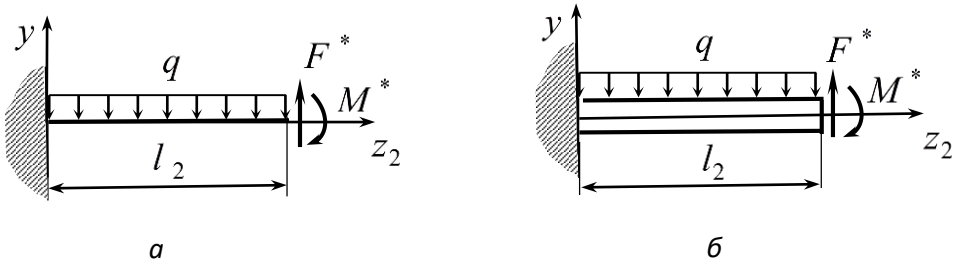
Далее описан метод определения $v_2^{\text{ту}}(z_2, y)$ на *BC* по теории упругости, в котором заранее не требуется учет условия связи для перемещений, но необходимо в расчетную схему внести параметры, отражающие действие внешних сил и наличие реакций опор. Если этого не сделать, например, для балки, не нагруженной на участке *BC*, то уравнения равновесия теории упругости будут однородными, а решение этих уравнений будет равным нулю.

Для применения на участке *BC* уравнений теории упругости с учетом внешних сил (реакций опор), действующих на других частях балки, воспользуемся результатом, вытекающим из решения дифференциального уравнения изгиба балки (9) в виде (10).

Если положить в уравнениях (10) равными нулю постоянные интегрирования, то

$$v_i^{\text{см}}(z_i) = f_i^{\text{см}}(z_i) = -\left[M_{0i}\, z_i^2 + (M_{1i}/3l_i)\, z_i^3 + (M_{2i}/6l_i^2)\, z_i^4 \right] / (2\, EI_{xi}). \quad (11)$$

Очевидно, что для этой функции $v_i^{\text{см}}(0) = 0,\ \dfrac{d\, v_i^{\text{см}}(0)}{d\, z_i} = 0$. Получили, что $v_i^{\text{см}}(z_i)$ в виде (11) – это решение дифференциального уравнения изгиба консольной балки, заделанной на левом конце при z_i =0 (рис. 4а).



а *б*

Рис. 4. Схема консоли для расчетов по теории стержней (а); полоса для расчетов по теории упругости (b).

Функция $f_i^{\text{см}}(z_i)$ получена на базе классической теории изгиба балок, которая из-за наличия трещины неприменима для участка *BC*. Предположим, что для построения, согласно теории упругости, функции $f_j^{\text{ту}}(z_j, y)$, аналогичной $f_i^{\text{см}}(z_i)$, нужно получить решение двумерной задачи плоской теории упругости,

использовав элемент в виде полосы, аналогичный тому, для которого определена функция (11). Поэтому считаем, что для определения $v_2^{\text{ту}}(z_2, y)$ нужно решить задачу теории упругости для полосы, левая грань которой (при z_2 =0) закреплена (рис. 2b), а в поперечных сечениях действуют изгибающие моменты

$$M_{x2}(z_2) = M_{02} + (M_{12}/l_2)\, z_2 + (M_{22}/l_2^2)\, z_2^2 . \qquad (12)$$

Здесь M_{02} и M_{12}/l_2 – значения изгибающего момента и поперечной силы в начале участка BC; $M_{22}/l_2^2 = q/2$.

Отметим, что при расчетах по теории упругости форму и размеры участка *BC*, показанного на рис. 4b, нужно устанавливать с учетом наличия в ней трещины.

Задав силу и момент, приложенные на правом краю полосы (рис. 4b), в виде

$$F^* = \frac{M_{12}}{l_2} - q l_2 , \quad M^* = M_{02} - M_{12} + \frac{q l_2^2}{2} ,$$

получим, что изгибающий момент в ее сечениях будет определяться по формуле (12).

При решении задачи теории упругости на поверхностях полосы (рис. 4b) должны действовать напряжения, обусловленные приложенными к телу внешними силами. Поэтому на гранях полосы, показанной на рис. 4b, вместо *q*, F^* и M^* нужно задать соответственно напряжения $\sigma_y^*(z_2)$, $\tau^*(y)$ и $\sigma^*(y)$. В дальнейшем при расчетах положим

$$\sigma_y^*(z_2) = \frac{q}{h} , \quad \tau^* = \frac{F^*}{A} , \quad \sigma^*(y) = \frac{M^*}{I_x}\, y .$$

Здесь h, *A* и I_x – ширина, площадь и осевой момент инерции сечения полосы соответственно; *y* – координаты точек сечения. При таком задании $\tau^*(y)$ и $\sigma^*(y)$ выполняются условия

$$F^* = \iint_A \tau^*(y)\, dA , \; M^* = \iint_A \sigma^*(y) \cdot y \; dA , \; \iint_A \sigma^*(y)\, dA = 0 .$$

Добавив к функции $f_2^{\text{ту}}(z_2, y)$, определенной как решение по теории упругости для участка стержня *BC* в виде полосы (рис. 4b), слагаемое $c_{12}^{\text{ту}} z_2 + c_{22}^{\text{ту}}$, получим

$$v_2^{\text{ту}}(z, y) = f_2^{\text{ту}}(z_2, y) + c_{12}^{\text{ту}} z_2 + c_{22}^{\text{ту}} . \qquad (13)$$

Здесь $c_{12}^{\text{ту}}$, $c_{22}^{\text{ту}}$ – постоянные, которые определяют перемещения участка *BC* как жесткого целого.

С учетом (10) и (13) получим

$$v_1^{\text{см}}(z) = f_1^{\text{см}}(z_1) + c_{11}^{\text{см}} z_1 + c_{21}^{\text{см}}, \quad v_2^{\text{ту}}(z_2, y) = f_2^{\text{ту}}(z_2, y) + c_{12}^{\text{ту}} z_2 + c_{22}^{\text{ту}},$$
$$v_3^{\text{см}}(z) = f_3^{\text{см}}(z_3) + c_{13}^{\text{см}} z_3 + c_{23}^{\text{см}}.$$

Таким образом, для каждого из участков *AB, BC* и *CD* (рис. 3) получены решения, которые должны быть состыкованы между собой и с частями балки, оставшимися за пределами фрагмента *AD*. Необходимо, чтобы на границах выделенных участков выполнялись условия стыковки поперечных сил Q_y и изгибающих моментов M_x, а прогибы и первые производные прогибов были непрерывны.

На границе участков с номерами J и I ($I > J$) условия стыковки сил Q_y и моментов M_x имеют вид

$$Q_{yI}(0) = Q_{yJ}(l_J) + Q_J^*, \qquad M_{yI}(0) = M_{yJ}(l_J) + M_J^*, \tag{14}$$

где Q_J^* и M_J^* – сосредоточенная поперечная сила и изгибающий момент, приложенные в конце *J*-го участка.

Нетрудно убедится, что равенства (14) будут выполняться при соответствующем задании изгибающего момента $M_{yI}(z_I)$. Например, если на границе участков с номерами 1 и 2 действуют сосредоточенная поперечная сила Q_1^* и изгибающий момент M_1^*, то, положив в (12)

$$M_{0\,2} = M_1^* + M_{01} + (M_{11}/l_1)\, l_1 + (M_{21}/l_1^2)\, l_1^2,$$
$$M_{1\,2}/l_2 = Q_1^* + (M_{11}/l_1) + 2\,(M_{21}/l_1^2)\, l_1,$$

получим выполнение условий (14).

Постоянные $c_{11}^{\text{см}}$, $c_{21}^{\text{см}}$, $c_{13}^{\text{см}}$, $c_{23}^{\text{см}}$, $c_{12}^{\text{ту}}$, $c_{22}^{\text{ту}}$ (*i*=1, 2, 3) определяются из граничных условий, а также из условий стыковки прогибов, углов поворота на границах участка *BC* и границах контакта выделенного фрагмента (рис. 3) с оставшейся частью балки. В качестве примера представим условия стыковки на границах участка *BC*

$$v_1^{\text{см}}(l_1) = v_2^{\text{ту}}(0, y), \qquad \frac{d\, v_1^{\text{см}}(l_1)}{dz_1} = \frac{\partial v_2^{\text{ту}}(0, y)}{\partial z_2},$$

$$v_2^{\text{ту}}(l_2, y) = v_3^{\text{см}}(0), \qquad \frac{\partial v_2^{\text{ту}}(l_1, y)}{\partial z_2} = \frac{d\, v_3^{\text{см}}(0)}{d\, z_3}. \qquad (15)$$

Обратим внимание на особенности, связанные с тем, что функции $v_1^{см}(z_1)$, $v_2^{см}(z)$ являются одномерными, а $v_2^{\text{ту}}(z_2, y)$ – двумерной. Вследствие этого в условиях вида (15) константы $v_1^{\text{см}}(l_1)$, $v_3^{\text{см}}(0)$, $\frac{d\, v_1^{\text{см}}(l_1)}{dz_1}$, $\frac{d\, v_3^{\text{см}}(0)}{d\, z_3}$ должны быть равны функциям от переменной y. Чтобы избавиться от этого противоречия, предлагаем в условиях стыковки прогибов и первых производных от прогибов доопределять функцию $v_2^{\text{ту}}(z_2, y\ )$. Например, в уравнениях вида (14) вместо $v_2^{\text{ту}}(z_2, y\ )$ можно использовать средние значения этой функции на интервале изменения толщины балки $-h/2 \le y\ \le h/2$, найденные из решения теории упругости. Можно также в условиях стыковки прогибов и первых производных от прогибов задавать некоторое значение координаты y. Такое дополнительное определение $v_2^{\text{ту}}(z_2, y\ )$ на границах участка *BC* не оказывает существенного влияния на результаты расчетов. Это обусловлено тем, что равенства (15) являются условиями стыковки решений, полученных по теории изгиба балки и теории упругости, а в соответствии с классической теорией изгиба прогибы и углы поворота балки не изменяются по высоте поперечного сечения.

Отметим, что приведенные далее результаты расчетов получены в случае, когда в равенствах вида (15) полагалось $y = -h/2$, то есть вместо переменных по высоте балки прогибов и углов поворота функции $v_2^{\text{ту}}(z_2, y\ )$ использовались их значения на нижней грани балки.

В соответствии с теорией упругости нормальные напряжения вычисляются по формуле

$$\sigma_z(z_2, y) = E\, \varepsilon_z(z_2, y)\ ,$$

где $\varepsilon_z(z_2, y) = \partial u(z_2, y) / \partial z_2$ – продольная деформация; $u(z_2, y)$ – продольное перемещение в полосе.

При расчетах по теории стержней имеем

$$\sigma_z(z_i, y) = \frac{M_{xi}(z_i)}{I_{xi}} y .$$

Отметим, что в качестве функции $f_2^{\text{ту}}(z_2, y)$ можно использовать аналитическое или численное решение задачи теории упругости, полученное с помощью различных методов на основе уравнений равновесия или вариационных соотношений.

ФОРМУЛА МОРА НА ОСНОВЕ СООТНОШЕНИЙ ТЕОРИИ УПРУГОСТИ

Для выделенного фрагмента *AD* (рис. 3) по классической формуле Мора (теория стержней) перемещение балки $v(z^1)$ при $z = z^1$ определяется по формуле

$$v(z^1) = \sum_{i=1}^{3} v_i^{\text{см}}(z^1) = \sum_{i=1}^{3} \frac{1}{EI_{xi}} \int_{l_i} M_{xi}^F M_{xi}^1 \, dl , \qquad (16)$$

где M_{xi}^F и M_{xi}^1 – моменты, обусловленные действием соответственно внешних нагрузок и единичной силы, приложенной при $z = z^1$.

Теория стержней неприменима к участку с трещиной. Следовательно, чтобы воспользоваться формулой (16), нужно функцию $v_2^{\text{см}}(z^1)$ заменить на аналогичную функцию $v_2^{\text{ту}}(z^1)$, которая учитывает наличие трещины на участке.

Представим вывод формула для вычисления $v_2^{\text{ту}}(z^1)$ по соотношениям плоской теории упругости.

При изгибе балки ее поперечная деформация значительно больше продольной. Поэтому при использовании соотношений плоской теории упругости положим, что потенциальная энергия деформации

$$U_2 = \frac{1}{2} \iiint_V E \, \varepsilon_z^2(z_2, y) \, dV ,$$

где V – объем стержня на участке *BC*, ε_z – продольная деформация, которая определяется по формуле

$$\varepsilon_z = \varepsilon_0^1(z) M_{02} + \varepsilon_1^1(z) M_{12} + \varepsilon_2^1(z) M_{22} .$$

Здесь $\varepsilon_0^1(y, z)$, $\varepsilon_1^1(y, z)$, $\varepsilon_2^1(y, z)$ – деформации от компонентов изгибающего момента при $M_{02} = 1$, $M_{12} = 1$ и $M_{22} = 1$ соответственно.

Приложим в точке с координатой $z = z^1$ силу *Ф*. По теореме Кастильяно, обобщенное перемещение в этой точке имеет вид

$$v_{2\Phi}^{\text{ту}}(\Phi, z^1) = \frac{dU_2}{d\Phi} = \iiint_V E(\varepsilon_z + \Phi\varepsilon_z^1)\ \varepsilon_z^1\ dV\ ,. \qquad (17)$$

где $\varepsilon_z^1 = \varepsilon_0^1(z)M_{02}^1 + \varepsilon_1^1(z)M_{12}^1$; ε_z^1 – продольная деформация, вызванная действием единичной силы (*Ф*=1); M_{02}^1, M_{12}^1 – компоненты изгибающих моментов от единичной силы.

Так как в действительности силы *Ф* в точке $z = z^1$ нет, то, положив *Ф* = 0 в выражении (17), получим

$$v_2^{\text{ту}}(z^1) = v_{2\Phi}^{\text{ту}}(0, z^1) = \iiint_V E\,\varepsilon_z(z_2, y)\ \varepsilon_z^1(z_2, y\ dV\ . \qquad (18)$$

Формула (17) является аналогом формулы Мора, используемой в теории сопротивления материалов. При расчетах по описанной модели в (16) вместо функции $v_2^{\text{см}}(z^1)$ следует включить $v_2^{\text{ту}}(z^1)$ в виде (18).

РЕЗУЛЬТАТЫ РАСЧЕТОВ

Численные результаты, представленные ниже, получены с применением для вычислений на участке, содержащем трещину, вариационного метода плоской задачи теории упругости [21].

Данные расчетов стержня с двумя симметрично расположенными трещинами (рис. 1) представлены в табл. 1, где табл. l^*, h^* – длина и ширина участка стержня, выделенного в окрестности трещины (рис. 1б); $\Delta l_a^{\text{ту}}$ и $\Delta l_б^{\text{см}}$ – изменения длин стержней, показанных на рис. 1а и 1б;

$$\Delta = (\left|\Delta l_a^{\text{ту}} - \Delta l_б^{\text{см}}\right| / \Delta l_a^{\text{ту}}) \cdot 100\%\,.$$

Табл. 1. Удлинения стержня с трещинами

l^*, м	h^*, м	$\Delta l_a^{\text{ту}} \cdot 10^3$, м	$\Delta l_б^{\text{см}} \cdot 10^3$, м	Δ, %
0.2	0.161	1.089	1.097	0.7
0.4	0.133		1.101	1.1
0.8	0.11		1.082	0.6

При проведении расчетов полагалось, что $N = 2 \cdot 10^6$ н; $l = l_1 + l^* + l_2 = 2$ м; $h = 0.2$ м; $b = 0.2$ м; $E = 2 \cdot 10^5$ МПа; коэффициент Пуассона $\nu = 0$; глубина трещины $a = 0.35\, h$.

Как видно из табл. 1, значения $\Delta l_a^{\text{ту}}$ и $\Delta l_б^{\text{см}}$ отличаются незначительно. Следовательно, описанная математическая модель может быть использована для определения продольных деформаций стержней по формулам теории сопротивления материалов. При этом по теории упругости рассчитывается только часть стержня, содержащая трещину.

В табл. 2 представлены данные расчетов консольной балки с прямоугольным поперечным сечением (рис. 5). Трещина расположена в верхней части сечения, проходящего через точку *C*. Ее глубина $a = h/2$. Полагалось $E = 2 \cdot 10^5$ МПа, $F = 10$ кН. Длина балки $l_1 + l_2 = 3$ м, ширина $в = 0{,}1$м. Высота сечений без трещины $h = 0.12$ м. Стыковка решений, полученных по теории стержней и теории упругости, производилась в сечении на расстоянии l_2 от трещины.

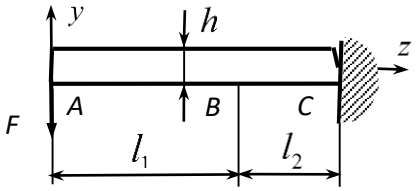


Рис. 5. Консольная балка с трещиной

В табл. 2 через $v_{мах}$ обозначен максимальный прогиб, σ_o – максимальное значение нормального напряжения при z = 2.8 м. Численные результаты, полученные с помощью метода сопряжения решений по теории сопротивления материалов и теории упругости, изложенного выше, приведены в столбце «СМ, ТУ». Результаты расчетов по соотношениям плоской задачи теории упругости для всей консоли представлены в столбце «ТУ», данные, вычисленные с учетом соотношения (18) по формуле Мора, – «ФМ».

Табл. 2. Прогибы и напряжения в консольной балке с трещиной (h = 0.12 м)

l_2, м	$v_{мах} \cdot 10^2$, м			σ_o, Мпа	
	СМ, ТУ	ФМ	ТУ	СМ, ТУ	ТУ
0.5	3.54	3.47	3.35	120	121
1	3.4	3.42		112	
2	3.23	3.33		115	

Как видно из табл. 2, величины прогибов $v_{мах}$ и напряжений σ_o, полученные с помощью предложенной модели, незначительно отличаются от значений этих величин, найденных по теории упругости.

ЗАКЛЮЧЕНИЕ

В работе представлена математическая модель для определения НДС стержня с трещинами при растяжении и сжатии. Для определения продольной деформации стержня предложено фрагмент стержня с трещиной заменить участком с постоянной площадью сечения и рассчитывать параметры этого участка по соотношениям теории упругости. Представлены формулы для расчетов. Изложены основные положения модели НДС стержней при изгибе, основанной на разделении балки на части и нахождении для каждой из частей деформаций и напряжений по теории стержней или теории упругости. Дан вывод аналога формулы Мора, основанный на соотношениях теории упругости. Отметим, чтори решении задач с использованием предложенных математических моделей могут применяться как численные, так и аналитические методы расчетов. Алгоритмы реализации описанных моделей сравнительно просты, так как решения с помощью различным теорий находятся для разных фрагментов стержней. Кроме того, достоверность и точность расчетов, основанных на описанных моделях, подтверждены численными результатами решения задач.

СПИСОК ЛИТЕРАТУРЫ

1. *Ахтямов А.М., Ильгамов М.А*. Обзор исследований по идентификации локальных дефектов стержней // Проблемы машиностроения и надежности машин. 2020. № 2. С. 3–15. https://doi.org/10.31857/S0235711920020042

2. *Ватульян А.О., Солуянов Н.О.* Идентификация полости в упругом стержне при анализе поперечных колебаний // ПМТФ. 2008. Т. 49, Вып. 6. С. 152–158

3. *Shifrin E.I.* Inverse spectral problem for a rod with multiple cracks // Mechanical Systems and Signal Processing. 2015. Vol. 56–57. P. 181–196. https://doi.org/10.1016/j.ymssp.2014.11.004

4. *Shifrin E.I.* Inverse spectral problem for a non-uniform rod with multiple cracks // Mechanical Systems and Signal Processing. 2017. Vol. 96. P. 348–365. https://doi.org/10.1016/j.ymssp.2017.04.029

5. *Лебедев И.М., Шифрин Е.И.* Решение обратной спектральной задачи для стержня, ослабленного поперечными трещинами, с помощью оптимизационного алгоритма Левенберга–Марквардта // Известия Российской академии наук. Механика твердого тела. 2019. Т. 4. С. 8–26. https://doi.org/10.1134/ 80572329919040056

6. *Лебедев И.М., Шифрин Е.И.* Идентификация поперечных трещин в стержне по собственным частотам поперечных колебаний // Известия Российской академии наук. Механика твердого тела. 2020. Т. 4. С. 50–70. https://doi.org/10.31857/s057232992004008x

7. *Ахтямов А.М., Ильгамов М.А.* Модель изгиба балки с надрезом: прямая и обратная задачи // Прикладная механика и техническая физика. 2013. Т. 54. № 1. С. 152–162. https://doi.org/10.1134/S0021894413010161

8. *Ильгамов М.А.* Продольные колебания стержня с зарождающимися поперечными трещинами // МТТ. 2017. № 1. С. 23–31. URL: https://rucont.ru/efd/592439

9. *Khiem N., Tran T., Ninh V.* A closed-form solution to the problem of crack identification for a multistep beam on Rayleigh quotient // Int. J. Solid Struct. 2018. Vol. 150. P. 154–165. https://doi.org/10.1016/j.ijsolstr.2018.06.010

10. *Акуленко Л.Д., Гавриков А.А., Нестеров С.В.* Идентификация дефектов поперечного сечения стержня по собственным частотам и особенностям формы продольных колебаний // Известия Российской академии наук. Механика твердого тела. 2019. Вып. 6. С. 98–107. https://doi.org/10 . 1134/ S0572329919060023

11. *Loya J., Lopez-Puente J., Zaera R., Fernandez-Saez J.* Free transverse vibrations of cracked nanobeams using a nonlocal elasticity model // J. Appl. Phys. 2009. Vol. 105. 044309. https://doi.org/10.1063/1.3068370. Corpus ID: 121034133

12. *Akbarzadeh Khorshidi M., Shariati M.* Buckling and postbuckling of size-dependent cracked microbeams based on a modified couple stress theory // J. Appl. Mech. Tech. Phys. 2017. Vol. 58. № 4. P. 717–724.
https://doi.org/10.1134/S0021894417040174

13. *Фу Ч., Ян С.* Анализ изгиба балки Тимошенко с трещиной с использованием нелокальной градиентной теории упругости // Прикладная механика и техническая физика. 2019. Т. 60. № 3. С. 196–206.
https://doi.org/10.15372/PMTF20190320

14. *Xiao Y., Jin H., Yu O.* Bending of Timoshenko beam with effect of crack gap based on equivalent spring model // Appl. Math. Mech. 2016. Vol. 37. P. 513–528. https://doi.org/10.1007/S10483-016-2042-9. Corpus ID: 124769412

15. *Batihan A.Ç., Kadioğlu F.S.* Vibration Analysis of a Cracked Beam on an Elastic Foundation // International Journal of Structural Stability and Dynamics. 2016. Vol. 16. № 5. 15500066. https://doi.org/10.1142/S0219455415500066

16. *Mazaheri H., Rahami H., Kheyroddin A.* Static and Dynamic Analysis of Cracked Concrete Beams Using Experimental Study and Finite Element Analysis // Periodica Polytechnica Civil Engineering. 2018. Vol. 62. No. 2. P. 337–345.
https://doi.org/10.3311/PPci.11450

17. *Fu C., Wang Y. u Tong D. Chunyu F., Yuyang W., Dawei T.* Stiffness Estimation of Cracked Beams Based on Nonlinear Stress Distributions Near the Crack // Mathematical Problems in Engineering. 2018. 5987973.
https://doi.org/10.1155/2018/5987973

18. *Серазутдинов М.Н.* Моделирование трещин и изменений в поперечных размерах стержня при продольной деформации // Вестник Технологического университета. 2024. Т. 27. № 7. С. 144–150

19. *Серазутдинов М.Н.* Определение перемещений балки с трещиной с использованием теории стержней // Известия КГАСУ. 2024. № 2(68). С. 114–123, https://doi.org/10.48612/NewsKSUAE/68.10, EDN: JPGQMS

20. *Серазутдинов М.Н., Убайдуллоев М.Н.* Метод расчета возможных параметров визуально недоступных повреждений стержневых конструкций // Вестник технологического университета. 2025. Т. 28. № 9. С. 112–115.

21. *Серазутдинов М.Н.* Метод построение финитной функций класса C^0 высокой степени аппроксимации // Вестник Технологического университета. 2016. Т. 19. № 11. С. 160–162.

CALCULATION OF ROD ELEMENTS WITH CRACKS BASED ON A COMBINATION OF ROD THEORY AND ELASTICITY THEORY

M. N. Serazutdinov[0000-0003-1675-3819]

Kazan National Research Technological University

serazmn@mail.ru

Abstract

Mathematical models for calculating the stress-strain state of rods with cracks under tension-compression and bending deformations are presented. A combination of the relations of the theory of elasticity and the theory of rods is used. The main provisions of the proposed modeling method are based on dividing the rod into fragments and finding deformations and stresses for each of the selected fragments according to the theory of rods or the theory of elasticity. Calculation algorithms are described, which are relatively simple to implement. Numerical data for solving problems are provided to illustrate the reliability and accuracy of calculations based on the models described in the article.

Keywords: *rod, beam, crack, stress-strain state, Mohr's formula, theory of elasticity, theory of rods*

REFERENCES

1. Akhtyamov A.M., Ilgamov M.A. Review of studies on the identification of local defects of rods // Problems of mechanical engineering and machine reliability. 2020. No. 2. P. 3–15. https://doi.org/10.31857/S0235711920020042

2. *Vatul`yan A.O., Soluyanov N.O.* Identifikaciya polosti v uprugom sterzhne pri analize poperechny`x kolebanij // PMTF. 2008. T. 49, Vy`p. 6. S. 152–158

3. *Shifrin E.I.* Inverse spectral problem for a rod with multiple cracks // Mechanical Systems and Signal Processing. 2015. Vol. 56–57. P. 181–196. https://doi.org/10.1016/j.ymssp.2014.11.004

4. *Shifrin E.I.* Inverse spectral problem for a non-uniform rod with multiple cracks // Mechanical Systems and Signal Processing. 2017. Vol. 96. P. 348–365. https://doi.org/10.1016/j.ymssp.2017.04.029

5. *Lebedev I.M., Shifrin E.I.* Reshenie obratnoj spektral`noj zadachi dlya sterzhnya, oslablennogo poperechny`mi treshhinami, s pomoshh`yu optimizacionnogo algoritma Levenberga–Markvardta // Izvestiya Rossijskoj akademii nauk. Mexanika tverdogo tela. 2019. T. 4. S. 8–26. https://doi.org/10.1134/ 80572329919040056

6. *Lebedev I.M., Shifrin E.I.* Identifikaciya poperechny`x treshhin v sterzhne po sobstvenny`m chastotam poperechny`x kolebanij // Izvestiya Rossijskoj akademii nauk. Mexanika tverdogo tela. 2020. T. 4. S. 50–70. https://doi.org/10.31857/s057232992004008x

7. *Akhtyamov A.M., Ilgamov M.A.* The model of bending a beam with an incision: direct and inverse problems // Applied mechanics and technical physics. 2013. Vol. 54. No. 1. P. 152–162. https://doi.org/10.1134/S0021894413010161

8. *Ilgamov M.A.* Longitudinal vibrations of a rod with incipient transverse cracks // MTT. 2017. No. 1. P. 23–31. URL: https://rucont.ru/efd/592439

9. *Khiem N., Tran T., Ninh V.* A closed-form solution to the problem of crack identification for a multistep beam on Rayleigh quotient // Int. J. Solid Struct. 2018. Vol. 150. P. 154–165. https://doi.org/10.1016/j.ijsolstr.2018.06.010

10. *Akulenko L.D., Gavrikov A.A., Nesterov S.V.* Identifikaciya defektov poperechnogo secheniya sterzhnya po sobstvenny`m chastotam i osobennostyam formy` prodol`ny`x kolebanij // Izvestiya Rossijskoj akademii nauk. Mexanika tverdogo tela. 2019. Vy`p. 6. S. 98–107. https://doi.org/10 . 1134/ S0572329919060023

11. *Loya J., Lopez-Puente J., Zaera R., Fernandez-Saez J.* Free transverse vibrations of cracked nanobeams using a nonlocal elasticity model // J. Appl. Phys. 2009. Vol. 105. 044309. https://doi.org/10.1063/1.3068370. Corpus ID: 121034133

12. *Akbarzadeh Khorshidi M., Shariati M.* Buckling and postbuckling of size-dependent cracked microbeams based on a modified couple stress theory // J. Appl. Mech. Tech. Phys. 2017. Vol. 58. No. 4. P. 717–724. https://doi.org/10.1134/S0021894417040174

13. Fu Ch., Yan S. Analysis of the Timoshenko beam bending with a crack using a non-local gradient theory of elasticity // Applied mechanics and technical physics. 2019. Vol. 60. No. 3 P. 196–206. https://doi.org/10.15372/PMTF20190320

14. *Xiao Y., Jin H., Yu O.* Bending of Timoshenko beam with effect of crack gap

based on equivalent spring model // Appl. Math. Mech. 2016. Vol. 37. P. 513–528. https://doi.org/10.1007/S10483-016-2042-9. Corpus ID: 124769412

15. *Batihan A.Ç., Kadioğlu F.S.* Vibration Analysis of a Cracked Beam on an Elastic Foundation // International Journal of Structural Stability and Dynamics. 2016. Vol. 16. № 5. 15500066. https://doi.org/10.1142/S0219455415500066

16. *Mazaheri H., Rahami H., Kheyroddin A.* Static and Dynamic Analysis of Cracked Concrete Beams Using Experimental Study and Finite Element Analysis // Periodica Polytechnica Civil Engineering. 2018. Vol. 62. No. 2. P. 337–345. https://doi.org/10.3311/PPci.11450

17. *Fu C., Wang Y. and Tong D. Chunyu F., Yuyang W., Duway T.* Stiffness Estimation of Cracked Beams Based on Nonlinear Stress Distributions Near the Crack // Mathematical Problems in Engineering. 2018. 5987973. https://doi.org/10.1155/2018/5987973

18. *Serazutdinov M.N.* Modelirovanie treshhin i izmenenij v poperechny`x razmerax sterzhnya pri prodol`noj deformacii // Vestnik Texnologicheskogo universiteta. 2024. T.27. № 7. S. 144–150

19. *Serazutdinov M.N.* Opredelenie peremeshhenij balki s treshhinoj s ispol`zovaniem teorii sterzhnej // Izvestiya KGASU. 2024. № 2(68). S. 114–123. https://doi.org/10.48612/NewsKSUAE/68.10, EDN: JPGQMS

20. *Serazutdinov M.N., Ubajdulloev M.N.* Metod rascheta vozmozhny`x parametrov vizual`no nedostupny`x povrezhdenij sterzhnevy`x konstrukcij // Vestnik texnologicheskogo universiteta. 2025. T. 28. № 9. S. 112–115.

21. *Serazutdinov M.N*. Metod postroenie finitnoj funkcij klassa vy`sokoj stepeni approksimacii // Vestnik Texnologicheskogo universiteta. 2016. T. 19. № 11. S. 160–162

СВЕДЕНИЯ ОБ АВТОРЕ

СЕРАЗУТДИНОВ Мурат Нуриевич – докт. физ.-мат. наук, профессор Казанского национального исследовательского технологического университета.

Murat Nurievich SERAZUTDINOV – Doctor of Physical and Mathematical Sciences, Professor of Kazan National Research Technological University.

email: serazmn@mail.ru

ORCID: 0000-0003-1675-3819

Материал поступил в редакцию 12 декабря 2025 года

УДК 004.81

МУЛЬТИ-ТАЙМФРЕЙМОВЫЕ DRUMMOND-ПАТЧИ И JEPA-ПРЕДОБУЧЕНИЕ ДЛЯ КРАТКОСРОЧНОГО ПРОГНОЗА РОЗНИЧНЫХ OHLC-РЯДОВ

А. С. Сизов[1] [0000-0001-8110-9929], **Ю. А. Халин**[2] [0000-0002-7020-8515], **А. А. Белых**[3] [0009-0005-7408-0052]

[1–3]*Юго-западный государственный университет, г. Курск, Россия*

[1]kafedra-ipm@mail.ru, [2]yur-khalin@yandex.ru, [3]belykhartem.a@mail.ru

Аннотация

Предложен метод построения инвариантных к масштабу представлений временных рядов розничной выручки на базе трехбарной (по трем соседним периодам) геометрии Драммонда (DG), расширенной мульти-таймфреймовым контекстом (день, частичная календарная неделя и скользящая 7-дневка). На этих «патчах» выполнено self-supervised предобучение по схеме Joint-Embedding Predictive Architecture (JEPA) со спатио-темпоральным маскированием, после чего модель дообучена с выходными слоями, оценивающими неопределенность, для прогноза на следующий день и следующую неделю. Проанализированы свойства аффинной инвариантности признаков и идентифицируемости недельной фазы; эмпирически продемонстрировано улучшение по сравнению с сильными базовыми моделями на реальных данных.

***Ключевые слова*:** геометрия Драммонда, Joint-Embedding Predictive Architecture (JEPA), временные ряды, Open-High-Low-Close (OHLC), розничная торговля, краткосрочный прогноз, самообучение.

ВВЕДЕНИЕ

Современные розничные временные ряды характеризуются высокой вариативностью, выраженной недельной сезонностью и наличием шума, что создает существенные сложности для краткосрочного прогнозирования ключевых метрик: выручки, количества кассовых чеков и среднего чека [1, 2]. Традиционные методы, такие как ARIMA и экспоненциальное сглаживание, могут демонстрировать недостаточную точность на нелинейных зависимостях и при резких изменениях режимов [3]. Нейросетевые подходы, включая N-BEATS и Temporal

Fusion Transformer (TFT), показывают лучшие результаты, но требуют больших объемов данных и могут быть неустойчивы к изменениям масштаба и сдвигам уровня ряда [4, 5].

В настоящей работе рассмотрена комбинация классических технических индикаторов и современных методов самообучающихся представлений. В качестве основы для признаков использована геометрия Драммонда (DG) [6] – набор интерпретируемых уровней, отражающих локальную геометрию ценового движения, обобщенную на случай произвольных временных рядов. Для обучения представлений применена архитектура JEPA (Joint-Embedding Predictive Architecture), показавшая свою эффективность в задачах компьютерного зрения и обработки сигналов [7, 8]. Ключевая идея JEPA – предсказание представлений одних частей данных по контексту других, что позволяет модели извлекать устойчивые скрытые (латентные) зависимости без реконструкции входных данных.

Перечислим основные полученные результаты.

1. Предложена единая многошкальная по времени (мультитаймфреймовая) постановка задачи для трех каналов розничных данных (выручка, чеки, средний чек) с использованием недельно-базисных приращений и OHLC-агрегирования, где Open – цена открытия, High – максимальная цена, Low – минимальная цена и Close – цена закрытия.

2. Разработан метод построения аффинно-инвариантных фрагментов данных (Drummond-патчей), объединяющих информацию с дневного и двухнедельных горизонтов.

3. Адаптирована и доработана схема JEPA-предобучения для временных рядов с пространственно-временным маскированием и позиционным кодированием, учитывающим недельную фазу.

4. На реальных данных выполнено экспериментальное сравнение предложенного подхода с рядом сильных базовых моделей, показавшее статистически значимое улучшение качества на краткосрочных горизонтах: следующий день (D+1) и следующая неделя (W+1). Эффективность самообучения для розничных продаж, таким образом, получила дополнительное подтверждение [9].

ДАННЫЕ И ОБОЗНАЧЕНИЯ

Использованы три канала, отражающие изменение динамики базовых розничных метрик:

- GainVal – изменение выручки;
- CheckCount – изменение количества покупок (кассовых чеков);
- ARVal – изменение среднего чека (Average Receipt Value).

Пусть V_t – выручка в день t, N_t – число кассовых чеков и $A_t = V_t/N_t$ – средний чек. Для подавления сезонности по дням недели используем недельно-базисные приращения (WeekBasis) в виде лог-отношений:

$$\Delta^W X_t = \log(X_t + \varepsilon) - \log\left(X_{(t-7)} + \varepsilon\right), \text{ где } X \in \{V, N, A\},\ \varepsilon > 0. \quad (1)$$

По умолчанию в тексте под именами каналов понимаем именно эти WeekBasis-приросты:

$$\text{GainVal}_t \equiv \Delta^W V_t,\ \ \text{CheckCount}_t \equiv \Delta^W N_t,\ \ \text{ARVal}_t \equiv \Delta^W A_t.$$

Для задач Day(t) может использоваться дневной аналог недельного приращения (1):

$$\Delta^D X_t = \log(X_t + \varepsilon) - \log(X_{t-1} + \varepsilon).$$

Дополнительно используем OHLC-агрегирование в заданном окне $[a, b]$:

$$\text{OHLC}_{[a,b]}(y) = (O, H, L, C) = \left(y_a,\ \max_{t\in[a,b]} y_t,\ \min_{t\in[a,b]} y_t,\ y_b\right).$$

Рассмотрим два недельных варианта окон для OHLC над приращениями $\Delta^W X$:

- Week_cal: календарная неделя (понедельник – воскресенье), содержащая t;
- Week_roll: скользящее окно $[t-6,\ t]$ из 7 дней.

На рис. 1 проиллюстрировано построение OHLC для канала GainVal; аналогично могут быть построены OHLC для каналов CheckCount и ARVal.

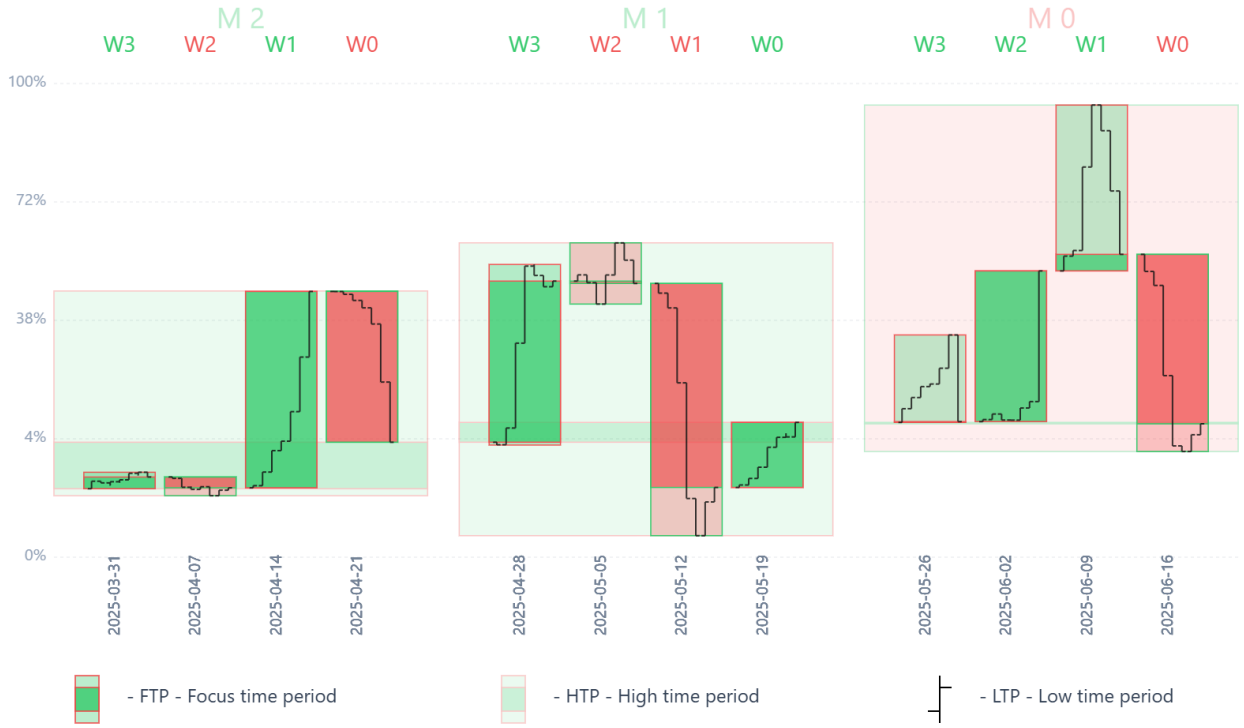


Рис. 1. Схема построения OHLC по WeekBasis-приростам выручки (GainVal).

1. МЕТОД: DRUMMOND-ПАТЧ И ДЕРА-ПРЕДОБУЧЕНИЕ

1.1. Построение мульти-таймфреймового Drummond-патча

Геометрия Драммонда (DG) определяет набор уровней на основе цен OHLC трех последовательных временных интервалов (баров) [6]. Для окна из трех баров (H_i, L_i, C_i), $i \in \{0,1,2\}$ (0 – текущий) определим базовые уровни:

$$\text{pivot}_i = (H_i + L_i + C_i)/3,$$

$$h_3 = (H_0 + H_1 + H_2)/3,$$

$$l_3 = (L_0 + L_1 + L_2)/3,$$

$$\text{pld} = (\text{pivot}_0 + \text{pivot}_1 + \text{pivot}_2)/3,$$

$$\text{rbird} = (\text{pivot}_0 + \text{pivot}_1 + C_0)/3.$$

На их основе вычислим производные уровни, такие как $\text{et}_1 = 2\,\text{pld} - l_3$, $\text{eb}_1 = 2\,\text{pld} - h_3$ и др. [6].

Патч P_t на конец дня t включает DG-уровни и z-координаты (нормализованные значения) для трех временных масштабов (тайм фреймов) τ:

- Day: последние три дня $[t-2, t]$;
- Week_cal: две прошлые полные календарные недели + текущая частичная (с понедельника по день t) (Mon…t);
- Week_roll: три скользящих окна $[t-20 \ldots t-14]$, $[t-13 \ldots t-7]$, $[t-6 \ldots t]$.

Для обеспечения аффинной инвариантности используем нормализацию внутри каждого окна $[a, b]$ по High/Low для каждой компоненты j:

$$z_{[a,b]}(x_t)^{(j)} = \frac{x_t^{(j)} - L_j}{\max\{H_j - L_j, \varepsilon\}}, \quad (2)$$

где $H_j = \max_{t \in [a,b]} x_t^{(j)}$, $L_j = \min_{t \in [a,b]} x_t^{(j)}$.

Лемма 1 (Аффинная инвариантность нормализации). *Для любого аффинного преобразования* $x \mapsto ax + b$ *с* $a > 0$ *и любого окна* [a,b] *выполняется* $z_{[a,b]}(ax + b) = z_{[a,b]}(x)$.

Доказательство следует непосредственно из определения (2).

Патч также обогащается межмасштабными признаками (позиция дневного бара относительно недельных уровней) и календарными мета-признаками (день недели, признак незавершенной (частичной) недели). Все расчеты произведены строго на исторических данных без утечки из будущего.

1.2. JEPA-обучение для розничных OHLC-представлений

Архитектура обучения, представленная на рис. 2, следует принципам JEPA [7, 8] в контексте задач временных рядов [10].

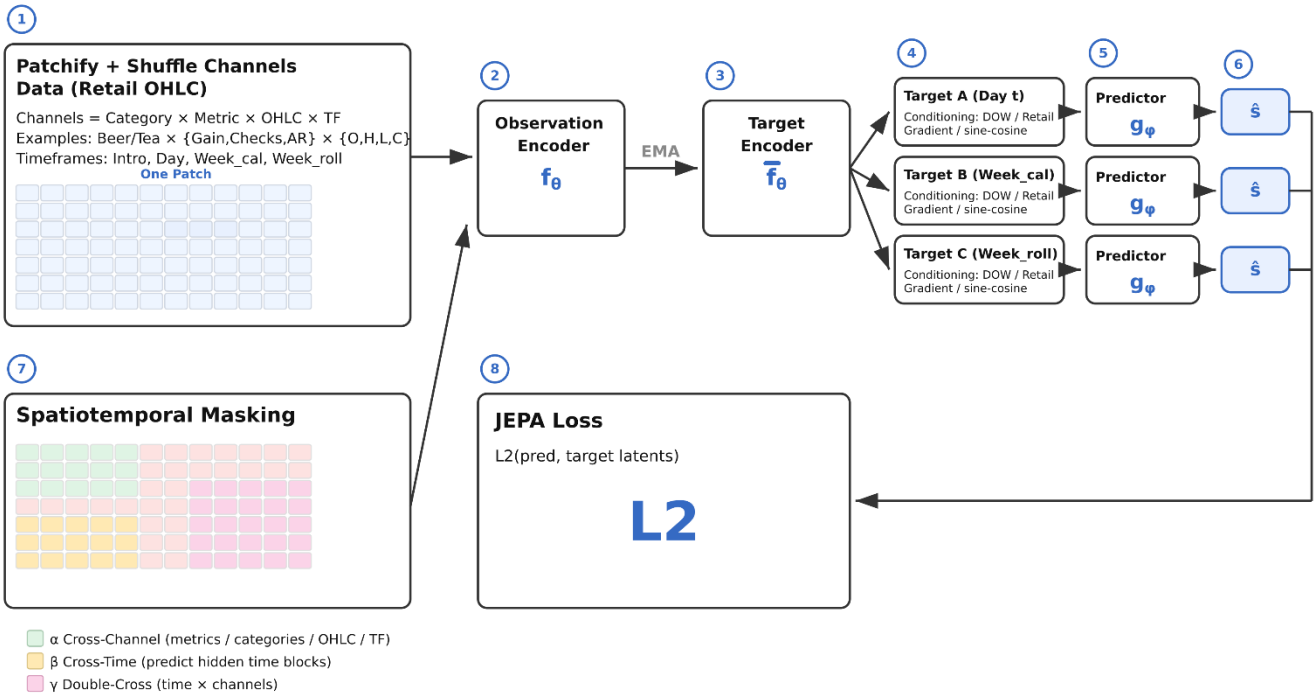


Рис. 2. Retail-JEPA: схема обучения и предсказания для мультишкальных OHLC-данных.

Пояснения к рис. 2.

Разбиение на фрагменты и перемешивание каналов (Patchify+Shuffle Channels). Вход – многоканальные розничные временные ряды формата Категория × Метрика × OHLC × Тайм фрейм.

Метрики: {GainVal, CheckCount, ARVal}; OHLC: {O,H,L,C}; тайм фреймы: {Intro, Day, Week_cal, Week_roll}. Поток разрезан на компактные DG-патчи (трехбарные блоки); каналы могут перемешиваться для регуляризации.

Кодировщик наблюдаемой части (Observation Encoder) f_θ. Незамаскированная (наблюдаемая) часть патча кодируется кодировщиком (энкодером) наблюдения f_θ в скрытое представление (латент) s_x. Во вход уже заложена DG-нормализация $(X - PLdot)/\Delta$, что делает признаки аффинно-инвариантными и снижает влияние локальных амплитуд.

EMA (экспоненциальное скользящее среднее) → Target Encoder $\bar{f}_\theta$. Целевой кодировщик (таргет-энкодер) $\bar{f}_\theta$ есть экспоненциально сглаженная (EMA) копия параметров f_θ. Он кодирует скрытые блоки (замаскированные регионы пат-

ча) в целевые скрытые представления (таргет-латенты) s_y^τ с остановкой градиента. Это стабилизирует целевые представления и предотвращает вырождение представлений (коллапс).

Позиционное обусловливание (Positional Conditioning). К латенту s_x добавляются позиционные признаки: (а) Temporal – день недели (DOW), синус/косинус-время, флаг partial-week (неполная неделя обрезана на текущем дне t); (б) Retail Gradient Positioning – векторные представления (эмбеддинги) структурных осей (категория товара, тип метрики, OHLC-канал, тип таймфрейма, Week_cal vs Week_roll). Эти признаки сообщают модели фазу недели (например, пятница/суббота пик спроса) и контекст тайм фрейма.

Предсказывающие головы/модули (Predictors) g_ϕ. Небольшие прогнозирующие (предикторные) головы g_ϕ (ViT/MLP-блоки) по объединенному представлению строят оценки латентов скрытых целей: Target A: дневной блок Day(t); Target B: Week_cal (календарная неделя Mon…t, обрезанная на текущем дне); Target C: Week_roll (скользящее окно t–6…t). Выходы обозначены как $\hat{s}_y^\tau$.

Предсказанные скрытые представления целей (Predicted Target Latents). $\hat{s}_y^\tau$ – предсказанные латентные представления целевых (скрытых) блоков в пространстве $\bar{f}_\theta$. Мы восстанавливаем напрямую не сами ряды, а их латенты, что подчеркивает структуру зависимостей «день ↔ неделя» и улучшает переносимость признаков к головам прогноза D+1/W+1.

Пространственно-временное скрытие (Spatiotemporal Masking). Применяем комбинированную маску без доступа к будущему:

α Cross-Channel – скрываем набор каналов (например, все каналы Week_cal или часть OHLC/метрик);

β Cross-Time – скрываем целиком временной блок (например, весь день t);

γ Double-Cross – одновременное скрытие по времени и каналам (например, Day(t)+Week_cal).

Для Week_cal берем только данные Mon…t; для Week_roll – строго окно t–6…t, т. е. информация из будущих дней не используется.

Функция потерь JEPA (JEPA Loss). Обучение идет по L2-расхождению между предсказанными и таргет-латентами: $L_{JEPA} = \sum_\tau \| \hat{s}_y^\tau - s_y^\tau \|_2^2$. Параметры $\bar{f}_\theta$

обновляются только через EMA, а не напрямую градиентом, что стабилизирует целевое пространство.

Пусть $\mathrm{P_t} = \{\mathrm{p_t^{(k)}}\}_{\mathrm{k=1}}^{\mathrm{K}}$ – набор патчей вокруг момента t. Применим двоичную маску $\mathrm{M} \subseteq \{1, \dots, \mathrm{K}\}$, разделяющую P_t на контекст (наблюдаемую часть) $\mathrm{P}_t^{\setminus \mathrm{M}}$ и таргет (замаскированную часть) $\mathrm{P}_t^{\mathrm{M}}$.

Контекстный энкодер $\mathrm{f_\theta}$ преобразует наблюдаемую часть в латентное представление $\mathrm{s_x} = \mathrm{f_\theta}\left(\mathrm{P}_t^{\setminus \mathrm{M}}\right)$. Целевой энкодер $\bar{\mathrm{f}}_\theta$, являющийся экспоненциально сглаженной (EMA) копией $\mathrm{f_\theta}$, обрабатывает исходный, незамаскированный патч P_t и извлекает эталонные представления $\mathrm{s_y^\tau} = \bar{\mathrm{f}}_\theta\ (\mathrm{P_t^M})$ для тех его частей, которые в данном примере обучения соответствуют маске M и тайм фрейму τ. Градиент через $\bar{\mathrm{f}}_\theta$ не пропускается (stop-gradient).

Предиктор $\mathrm{g_\phi}$, получая на вход $\mathrm{s_x}$ и позиционные признаки (день недели, тип тайм фрейма), предсказывает латенты целевых патчей: $\hat{\mathrm{s}}_\mathrm{y}^\tau = \mathrm{g_\phi}(\mathrm{s_x}, \tau)$.

Целевая функция обучения – это минимизация L2-расстояния между предсказанными и эталонными скрытыми представлениями (латентными векторами):

$$\mathrm{L_{JEPA}}(\theta, \phi) \ = \ \mathbb{E}\left[\ \| \ \mathrm{g_\phi}\left(\mathrm{f_\theta}(\mathrm{P_t^{\setminus M}})\right) - \mathrm{sg}\left(\bar{\mathrm{f}}_\theta(\mathrm{P_t^M})\right) \ \|_2^2\ \right],$$

где $\mathrm{P_t^M}$ обозначает часть исходного фрагмента данных (патча), соответствующую схеме маскирования M (какие части входа скрываются), математическое ожидание берем по t, маскам M и весам $\mathrm{w_\tau}$.

В соответствии с теорией оптимального прогнозирования, это оптимальный предиктор в таком сценарии стремится к условному математическому ожиданию целевых представлений при данном контексте.

2. ЭКСПЕРИМЕНТЫ

2.1. Данные и настройка эксперимента

Эксперименты проводились на реальных данных розничной сети за период с 2020 по 2025 г. Были использованы данные по двум товарным категориям: пиво (Beer) и чай/кофе/какао (TeaCoffeeCocoa). Прогноз строился для трех каналов: GainVal, CheckCount, ARVal. Выборка была разделена на обучающую (2020–

2024), проверочную (I квартал 2025 г.) и тестовую (II квартал 2025 г.) части с соблюдением временного порядка.

2.2. Детали реализации

Энкодер f_θ и предиктор g_ϕ были реализованы на основе трансформерной архитектуры с 4 слоями, 8 «головами» механизма внимания (параллельными каналами attention) и размерностью скрытого состояния 256. Размерность латентного представления s = 128. Вероятность маскирования патча – 30%. Коэффициент EMA ($\bar{\theta}$) для целевого энкодера – 0.99. Обучение проводилось оптимизатором AdamW со скоростью обучения (learning rate) 10^{-4} и размером минипакета данных (батча) 128 в течение 100 эпох. Предобучение JEPA заняло приблизительно два дня на графическом ускорителе GPU NVIDIA V100. После предобучения к латентным представлениям добавлялись простые выходные слои, оценивающие распределение будущих значений (два полносвязных слоя), и модель дообучалась на задаче прогнозирования.

2.3. Модели и метрики

Предложенный метод (JEPA+Heads) сравнивался со следующими бэйзлайнами:

- SeasonalNaive: наивный прогноз с недельной сезонностью;
- Ridge/LightGBM: линейная модель и градиентное усиление (бустинг) на табличных признаках, лежащие в основе современных ансамблевых решений [11];
- N-BEATS/N-HiTS: современные нейросетевые модели для прогнозирования временных рядов [4];
- TFT: Temporal Fusion Transformer [5].

С целью оценки вкладов компонентов (ablation study) были также протестированы:

- LightGBM (DG-признаки): LightGBM, обученный на сконструированных Drummond-патчах;
- TFT (DG-признаки): модель TFT, обученная на тех же патчах с DG-признаками сквозным образом (end-to-end);

- JEPA+Heads (Raw Features): наш метод, но на сырых недельных приращениях без DG-обработки.

Целями прогнозирования были:

- D+1: прогноз дневного приращения $\Delta^{D}X_{(t+1)}$;
- W+1: прогноз OHLC по $\Delta^{W}X$ на следующей неделе.

Использовались следующие метрики: симметричная средняя абсолютная процентная ошибка (sMAPE), средняя абсолютная масштабированная ошибка (MASE), для оценки статистической значимости различий применялся тест Диболда – Мариано [12].

2.4. Результаты и их обсуждение

Основные результаты, усредненные по всем каналам и категориям, представлены в табл. 1.

Табл. 1. Основные результаты на GainVal/CheckCount/ARVal: усредненные ошибки (меньше – лучше).

Модель	sMAPE (D+1)	sMAPE (W+1)	MASE (D+1)	MASE (W+1)
SeasonalNaive	21.8	24.9	1.00	1.00
Ridge / LightGBM	19.6	22.3	0.92	0.95
N-BEATS / N-HiTS	18.0	20.5	0.86	0.90
TFT	17.2	19.9	0.84	0.88
JEPA+Heads (предл.)	**14.9**	**17.6**	**0.76**	**0.82**

Предложенный метод JEPA+Heads показал наилучшие результаты по всем метрикам на обоих горизонтах прогноза. Улучшение по сравнению с лучшим из бэйзлайнов (TFT) составило около 13% по sMAPE на горизонте D+1 и 11%

на горизонте W+1. Результаты теста Диболда – Мариано подтвердили статистическую значимость улучшений (значение p-value против модели TFT составило 0.003 для D+1 и 0.007 для W+1). Анализ разбивки результатов по отдельным каналам и категориям показал согласованные улучшения, с максимальным выигрышем на канале GainVal (15.2% на D+1).

Результаты ablation study (табл. 2) показывают вклад каждого компонента метода. LightGBM на DG-признаках уже демонстрирует улучшение над LightGBM на сырых данных, что подтверждает полезность самих Drummond-патчей. TFT, обученная на DG-признаках, показывает результат, близкий к оригинальной TFT, что говорит о сложности прямого использования этих признаков без специального предобучения. Наш метод без DG-признаков (JEPA+Heads (Raw Features)) уступает полной модели, но все же превосходит TFT, что доказывает эффективность JEPA-предобучения. Наилучший результат достигается только при совместном использовании DG-признаков и JEPA-предобучения.

Табл. 2. Исследование методом абляции: sMAPE на горизонте D+1 (усреднено).

Модель	sMAPE (D+1)
LightGBM (Raw Features)	19.6
LightGBM (DG-признаки)	18.1
TFT (Raw Features)	17.2
TFT (DG-признаки)	17.4
JEPA+Heads (Raw Features)	16.0
JEPA+Heads (DG-признаки, полный метод)	**14.9**

Эффективность предложенного метода объясняется сочетанием аффинно-инвариантных DG-признаков, которые устойчивы к изменениям масштаба ряда, и JEPA-предобучения, которое позволяет извлекать информативные представления, согласованные между различными временными горизонтами.

ЗАКЛЮЧЕНИЕ

Представлен метод прогнозирования розничных временных рядов, сочетающий построение мульти-таймфреймовых Drummond-патчей и self-supervised предобучение по схеме JEPA, когда целевые сигналы формируются автоматически из исходных данных. Ключевыми особенностями метода являются аффинно-инвариантная нормализация признаков, пространственно-временное маскирование патчей и использование EMA-таргет энкодера для стабилизации обучения.

Эксперименты на реальных данных показали, что предложенный подход статистически значимо превосходит сильные бэйзлайны на горизонтах прогноза D+1 и W+1. Исследование методом абляции подтвердило важность каждого компонента метода. Полученные результаты свидетельствуют о перспективности метода для практического применения в задачах операционного планирования в розничной торговле.

Основными ограничениями работы являются локальность трехбарного анализа DG, зависимость от схемы маскирования и отсутствие учета внешних факторов (праздники, акции). Перспективными направлениями дальнейших исследований являются интеграция внешних признаков, разработка более сложных стратегий маскирования и масштабирование метода на большее число товарных категорий.

СПИСОК ЛИТЕРАТУРЫ

1. *Fildes R., Ma S., Kolassa S.* Retail forecasting: Research and practice // International Journal of Forecasting. 2022. Vol. 38, No. 4. P. 1283–1318. https://doi.org/10.1016/j.ijforecast.2019.06.004

2. *Lim B., Arik S. O., Loeff N., Pfister T.* Temporal Fusion Transformers for interpretable multi-horizon time series forecasting // International Journal of Forecasting. 2021. Vol. 37, No. 4. P. 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

3. *Hyndman R. J., Athanasopoulos G.* Forecasting: Principles and Practice. 2nd ed. Melbourne: OTexts, 2018. 380 p. Цит. по с.: 183–220, 221–274, 347–368.

4. *Oreshkin B.N., Carpov D., Chapados N., Bengio Y.* N-BEATS: Neural basis expansion analysis for interpretable time series forecasting // arXiv preprint

arXiv:1905.10437. 2019. https://doi.org/10.48550/arXiv.1905.10437; Challu C. et al. NHITS: Neural hierarchical interpolation for time series forecasting // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. Vol. 37, No. 6. P. 6989–6997. https://doi.org/10.1609/aaai.v37i6.25854

5. *Yue Zh. et al.* TS2Vec: Towards Universal Representation of Time Series // Proceedings of the AAAI Conference on Artificial Intelligence. 2022. Vol. 36, No. 8. P. 8980–8987. https://doi.org/10.1609/aaai.v36i8.20881

6. *Hearne T.* Drummond Geometry: Picking Yearly Highs and Lows in Interbank Forex Trading // Breakthroughs in Technical Analysis: New Thinking from the World's Top Minds / ed. by D. Keller. Princeton: Bloomberg Press, 2007. P. 1–19. https://doi.org/10.1002/9781119204749.ch1.

7. *Dawid A., LeCun Y.* Introduction to Latent Variable Energy-Based Models: A Path Towards Autonomous Machine Intelligence // Journal of Statistical Mechanics: Theory and Experiment. 2024. No. 10. Art. 104011. https://doi.org/10.1088/1742-5468/ad292b. (arXiv:2306.02572)

8. *Assran M. et al.* Self-supervised learning from images with a joint-embedding predictive architecture // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 15619–15629. https://doi.org/10.1109/CVPR52729.2023.01499

9. *Park Y.J. et al.* A scalable and transferable time series prediction framework for demand forecasting. 2024. arXiv preprint arXiv:2402.19402. https://doi.org/10.48550/arXiv.2402.19402

10. *Ragab M., Liu Q., Jia W., Chen M., Yun U.* Self-Supervised Learning for Time Series Analysis: Taxonomy, Progress, and Prospects // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024. Vol. 46, No. 10. P. 6775–6794. https://doi.org/10.1109/TPAMI.2024.3387317

11. *Волошин Т.А., Зайцев К.С., Дунаев М.Е.* Применение адаптивных ансамблей методов машинного обучения к задаче прогнозирования временных рядов // International Journal of Open Information Technologies. 2023. Т. 11, №. 8. С. 57–63.

12. *Diebold F.X., Mariano R.S.* Comparing Predictive Accuracy // Journal of Business & Economic Statistics. 1995. Vol. 13, No. 3. P. 253–263. https://doi.org/10.1080/07350015.1995.10524599

MULTI-TIMEFRAME DRUMMOND PATCHES AND JEPA PRE-TRAINING FOR SHORT-TERM RETAIL OHLC SERIES FORECASTING

A. S. Sizov[1] [0000-0001-8110-9929], Y. A. Khalin[2] [0000-0002-7020-8515],
A. A. Belykh[3] [0009-0005-7408-0052]

[1–3]*Southwest State University, Kursk, Russia*

[1]kafedra-ipm@mail.ru, [2]yur-khalin@yandex.ru, [3]belykhartem.a@mail.ru

Abstract

We propose a method for constructing scale-invariant representations of retail revenue time series based on three-bar Drummond Geometry (DG) computed over three adjacent periods, extended with a multi-timeframe context (day, partial calendar week, and a rolling 7-day window). Self-supervised pre-training on these “patches” is performed using a Joint-Embedding Predictive Architecture (JEPA) with spatio-temporal masking, followed by fine-tuning with output heads that quantify predictive uncertainty for next-day and next-week forecasts. The work analyzes the properties of affine invariance of the features and the identifiability of the weekly phase; empirical improvement over strong baseline models on real-world data is demonstrated.

Keywords: Drummond Geometry, Joint-Embedding Predictive Architecture (JEPA), time series, Open-High-Low-Close (OHLC), retail, short-term forecasting, self-supervised learning.

REFERENCES

1. *Fildes R., Ma S., Kolassa S.* Retail forecasting: Research and practice // International Journal of Forecasting. 2022. Vol. 38, No. 4. P. 1283–1318. https://doi.org/10.1016/j.ijforecast.2019.06.004

2. *Lim B., Arik S.O., Loeff N., Pfister T.* Temporal Fusion Transformers for interpretable multi-horizon time series forecasting // International Journal of Forecasting. 2021. Vol. 37, No. 4. P. 1748–1764.
https://doi.org/10.1016/j.ijforecast.2021.03.012

3. *Hyndman R.J., Athanasopoulos G.* Forecasting: Principles and Practice. 2nd ed. Melbourne: OTexts, 2018. 380 p. Cited pp.: 183–220, 221–274, 347–368.

4. *Oreshkin B.N., Carpov D., Chapados N., Bengio Y.* N-BEATS: Neural basis expansion analysis for interpretable time series forecasting // arXiv preprint arXiv:1905.10437. 2019. https://doi.org/10.48550/arXiv.1905.10437; *Challu C. et al.* NHITS: Neural hierarchical interpolation for time series forecasting // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. Vol. 37, No. 6. P. 6989–6997.
https://doi.org/10.1609/aaai.v37i6.25854

5. *Yue Zh. et al*. TS2Vec: Towards Universal Representation of Time Series // Proceedings of the AAAI Conference on Artificial Intelligence. 2022. Vol. 36, No. 8. P. 8980–8987. https://doi.org/10.1609/aaai.v36i8.20881

6. *Hearne T.* Drummond Geometry: Picking Yearly Highs and Lows in Interbank Forex Trading // Breakthroughs in Technical Analysis: New Thinking from the World's Top Minds / ed. by D. Keller. Princeton: Bloomberg Press, 2007. P. 1–19.
https://doi.org/10.1002/9781119204749.ch1

7. *Dawid A., LeCun Y.* Introduction to Latent Variable Energy-Based Models: A Path Towards Autonomous Machine Intelligence // Journal of Statistical Mechanics: Theory and Experiment. 2024. No. 10. Art. 104011.
https://doi.org/10.1088/1742-5468/ad292b (arXiv:2306.02572).

8. *Assran M. et al.* Self-supervised learning from images with a joint-embedding predictive architecture // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 15619–15629.
https://doi.org/10.1109/CVPR52729.2023.01499

9. *Park Y.J. et al*. A scalable and transferable time series prediction framework for demand forecasting. 2024. arXiv preprint arXiv:2402.19402.
https://doi.org/10.48550/arXiv.2402.19402

10. *Ragab M., Liu Q., Jia W., Chen M., Yun U.* Self-Supervised Learning for Time Series Analysis: Taxonomy, Progress, and Prospects // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024. Vol. 46, No. 10. P. 6775–6794. https://doi.org/10.1109/TPAMI.2024.3387317

11. *Voloshin T.A., Zaitsev K.S., Dunaev M.E.* Primenenie adaptivnykh ansamblei metodov mashinnogo obucheniya k zadache prognozirovaniya vremennykh ryadov [Application of adaptive ensembles of machine learning methods to the problem of time series forecasting] // International Journal of Open Information Technologies. 2023. Vol. 11, No. 8. P. 57–63.

12. *Diebold F.X., Mariano R.S.* Comparing Predictive Accuracy // Journal of Business & Economic Statistics. 1995. Vol. 13, No. 3. P. 253–263. https://doi.org/10.1080/07350015.1995.10524599

СВЕДЕНИЯ ОБ АВТОРАХ

СИЗОВ Алесандр Семенович – доктор технических наук, профессор, профессор кафедры «Программная инженерия», Юго-Западный государственный университет.

Alexander Semenovich SIZOV – Doctor of technical sciences. (Engineering), Professor, Professor at the Software Engineering Department, Southwest State University.

email: kafedra-ipm@mail.ru

ORCID: 0000-0001-8110-9929

ХАЛИН ЮРИЙ Алексеевич – кандидат технических наук, доцент, доцент кафедры «Программная инженерия», Юго-Западный государственный университет.

Yuri Alekseevich KHALIN – Cand. of Sci. (Engineering), Associate Professor, Associate Professor at the Software Engineering Department, Southwest State University.

email: yur-khalin@yandex.ru

ORCID: 0000-0002-7020-8515

БЕЛЫХ Артем Александрович – аспирант, Юго-Западный государственный университет.

Artem Aleksandrovich BELYKH – Postgraduate Student, Southwest State University.

email: belykhartem.a@mail.ru

ORCID: 0009-0005-7408-0052

Материал поступил в редакцию 18 декабря 2025 года

УДК 004.02

МЕТОДЫ КОГНИТИВНОГО МОДЕЛИРОВАНИЯ И ГИБРИДНЫЕ ЭВОЛЮЦИОННО-МНОГОКРИТЕРИАЛЬНЫЕ АЛГОРИТМЫ В МУЛЬТИАГЕНТНОЙ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ

В. Б. Чечнев[0009-0000-1523-3294]

Московский государственный технический университет им. Н. Э. Баумана, г. Москва, Россия

gegrev@yandex.ru

Аннотация

Предложен подход к поддержке многокритериальных решений на основе когнитивно-ориентированной мультиагентной информационно-аналитической системы. Разработаны методы когнитивного моделирования, включающие формально-онтологическое представление знаний о планировании работ и коалиционно-холоническую агентную архитектуру, а также обеспечивающие адаптивность и прозрачность вычислений. Предложен гибридный эволюционно-многокритериальный алгоритм, в рамках которого агенты генерируют альтернативные планы с помощью параллельного генетического алгоритма на локальном уровне, оптимизирующего сочетание нескольких критериев. На глобальном уровне реализован многоэтапный отбор альтернатив с фильтрацией перегрузок ресурсов и подобных решений, а также финальное агрегирование с использованием многокритериальных методов принятия решений PROMETHEE и ELECTRE.

Проведено экспериментальное исследование, сравнивающее эффективность планирования вручную и с помощью разработанной системы, а также анализ влияния динамической адаптации параметров генетического алгоритма. Полученные результаты показали, что применение системы позволяет сократить время формирования плана в 20–30 раз при сопоставимом или лучшем качестве. При этом полностью устраняются перегрузки исполнителей и обеспечивается раннее прекращение эволюционных расчетов без потери качества решений. Разработанная система и предложенные алгоритмы ориентированы на использование при планировании проектной деятельности на производственных предприятиях.

***Ключевые слова**: когнитивное моделирование, системы поддержки принятия решений, мультиагентные системы, генетический алгоритм, информационные системы, многокритериальная оптимизация, планирование загрузки персонала.*

ВВЕДЕНИЕ

Современные производственные предприятия сталкиваются с резко возрастающей сложностью задач планирования работ (ПР) и распределения ресурсов. Лица, принимающие решения (ЛПР), вынуждены действовать в условиях динамично изменяющейся среды, ограниченного времени и множества конфликтующих критериев. В таких условиях традиционные системы поддержки принятия решений (СППР), основанные на жестко запрограммированных алгоритмах и статических моделях, не обладают требуемой адаптивностью, что приводит к снижению качества разрабатываемых ими планов при повышении временных затрат.

Одним из перспективных направлений развития СППР является когнитивно-ориентированный подход, предполагающий интеграцию моделей, имитирующих когнитивные процессы ЛПР, с современными методами оптимизации. В частности, актуальной задачей является объединение когнитивного моделирования, мультиагентных технологий и адаптивных многокритериальных алгоритмов в единую СППР. Мультиагентная система (МАС) позволяет представить сложный процесс ПР в виде распределенной структуры взаимодействующих программных агентов, что повышает гибкость и масштабируемость решения. Включение механизмов динамической настройки и многокритериального анализа, в свою очередь, обеспечивает адаптацию системы к изменениям условий и учет сразу нескольких показателей эффективности.

В настоящей работе рассмотрен подход к построению когнитивно-ориентированной МАС для информационно-аналитической поддержки ПР. Целью является повышение качества и оперативности генерации решений за счет применения методов когнитивного моделирования, а также гибридного эволюционного алгоритма с многокритериальной оптимизацией. Для ее достижения решены следующие задачи:

1. разработана мультиагентная архитектура информационно-аналитической системы для ПР;

2. предложен и реализован адаптивный генетический алгоритм (ГА) генерации альтернативных решений (АР);

3. разработан двухэтапный механизм отбора и многокритериальной агрегации АР, основанный на фильтрации перегрузок и подобных вариантов, использующий многокритериальные методы принятия решений (ММПР) ELECTRE и PROMETHEE;

4. проведена экспериментальная апробация МАС на наборе сценариев различной мощности, показавшая существенное сокращение времени планирования и улучшение качества решений по сравнению с ручным подходом.

МЕТОДЫ

Когнитивное моделирование и онтологическая архитектура агентов

В качестве основы разработанной МАС сформирована формально-онтологическая модель процесса ПР. В онтологии выделены классы, соответствующие ключевым сущностям предметной области таким как: операция, продукт, контракт, работник, ресурс, период и их отношениям. С использованием фреймворка Akka.NET[1] на базе этой онтологии спроектирована многоуровневая агентная архитектура. Данная платформа реализует принципы реактивных систем [1, 2]: отзывчивость, устойчивость к сбоям, эластичность и асинхронный обмен сообщениями. Каждый класс онтологии сопоставлен определенному типу актора (адаптивного агента) со строго заданными ролью и протоколом взаимодействия. Например, класс Операция соответствует OperationActor, генерирующему альтернативные варианты расписаний для данной производственной операции, а класс Работник – WorkerActor, управляющему данными о конкретном сотруднике (см. табл. 1).

[1] Akka.NET – набор инструментов и среда выполнения для создания высокопараллельных, распределенных и отказоустойчивых событийно-управляемых приложений.

Табл. 1. Отображение онтологии на акторы Akka.NET в МАС

Класс онтологии	Актор	Описание актора
Операция	OperationActor	Выполняет генерацию альтернативных вариантов расписаний
	SupervisorActor	Организует работу всех OperationActors
Работник	WorkerActor	Управляет данными о сотрудниках
Группа работников	WorkerCoalitionActor	Управляет работниками, объединенными в одну коалицию для выполнения определенной Операции
Продукт	ProductActor	Управляет данными о Продуктах
	CoordinatorActor	Координирует результаты работы SupervisorActors на уровне Операций и отправляет консолидированную альтернативу на проверку в механизм фильтрации
Контракт	ContractActor	Управляет данными о Контрактах
План	TimeSalaryRegistersActor	Отвечает за взаимодействие с планом

Каждый актор функционирует в среде параллельного обмена сообщениями и имеет свое внутреннее состояние. Формально состояние актора $a_i \in A$ в момент времени t можно представить как

$$s(a_i, t) = \langle D_{a_i}(t), Q_{a_i}(t) \rangle,$$

где $D_{a_i}(t)$ – внутреннее состояние данных актора, включающее в себя соответствующие атрибуты онтологии, а $Q_{a_i}(t)$ – очередь сообщений, ожидающих обработки. Поведение актора задается функцией перехода состояния

$$s(a_i, t+1) = f_{a_i}\big(s(a_i, t), m_j\big),$$

где $m_j \in Q_{a_i}(t)$, а также функцией отправки сообщений другим акторам $\mathrm{out}_{a_i}(t+1)$, формирующей множество исходящих сообщений от актора a_i в каждый момент времени $t+1$.

Акторы объединены в иерархические группы – холоны, в которых специальный актор (SupervisorActor) управляет и регулирует взаимодействие акторов

между собой и внешними акторами, а также создает или отключает их для балансировки нагрузки. Кроме того, коалиции акторов на уровне класса Продукт координируются с помощью CoordinatorActor. Такая коалиционно-холоническая архитектура обеспечивает устойчивость и гибкость системы.

Взаимодействие агентов основано на асинхронной передаче сообщений. Для этого разработан их формальный словарь, представляющий собой перечень типов сообщений с их полями, и определяющий протокол коммуникаций между акторами. Реализация обмена сообщениями использует шаблон маршрутизации Akka.NET. Таким образом, онтологическая модель вместе с реактивной акторной платформой закладывает когнитивную прозрачность системы, в которой каждый агент отвечает за понятный локальный фрагмент задачи, а коммуникация между ними отражает естественные связи предметной области.

Генерация альтернатив с помощью адаптивного генетического алгоритма

Для автоматического формирования планов работ создан параллельный эволюционный механизм, запускаемый на уровне каждого OperationActor. Каждому агенту ставится подзадача составления расписания для одной операции, представленная в виде распределения требуемого объема работ L по доступным работникам R в дискретном периоде T. Такая постановка задачи ПР относится к классу задач расписания, которые являются NP-трудными и активно изучаются в современной литературе [3–8]. В научном дискурсе существует множество подходов, используемых в подобных задачах, обобщая и группируя их по базовым принципам работы, можно выделить [3–9]:

- целочисленное линейное программирование (ЦЛП) и целевое программирование (ЦП);
- жадные алгоритмы (ЖА);
- локальный поиск (ЛП);
- имитация отжига (ИО);
- муравьиный алгоритм (МА) и оптимизация на основе роя частиц (ООРЧ);
- генетический алгоритм (ГА).

Среди вышеперечисленных групп, на основе выдвинутых критериев, а также прочих технических и функциональных требований (масштабируемость,

параллелизм, отказоустойчивость и возможность адаптации параметров в режиме реального времени) был выбран один из видов эволюционных алгоритмов – ГА (см. табл. 2).

Табл. 2. Сравнение методов решения

Критерий	ЦЛП и ЦП	ЖА	ЛП	ИО	МА и ООРЧ	**ГА**
Отсутствие дискретных ограничений	±	–	±	±	±	**+**
Наличие многоцелевой оптимизации	±	–	±	±	±	**+**
Масштабируемость по числу операций	–	±	±	±	±	**+**
Параллелизм	±	±	±	±	±	**+**
Частичные результаты	–	+	+	+	±	**+**
Адаптация параметров в режиме реального времени	–	–	±	+	±	**+**
Прогнозируемость времени вычислений	Низкая	Высокая	Средняя	Средняя	Средняя	**Высокая**

Каждый OperationActor запускает локальный ГА для своей операции. Хромосома кодирует план выполнения операции O – распределение часов по сотрудникам $r \in R$ и периодам $t \in T$. Значение гена $x_{r,t}$ интерпретируется как назначенное число часов работнику r в период t. При наличии технологических связей (предшествований) накладываются дополнительные ограничения на допустимые периоды выполнения, передаваемые агентам в виде параметров.

Для оценки качества каждого $x_{r,t}$ предлагается использовать следующие критерии оптимальности.

- Соответствие задаваемому бюджету $f'(x)$ – отклонение суммарной стоимости назначенных часов от заданного диапазона бюджета по операции. Для оценки используется экспоненциальное штрафование перерасхода или недоосвоения.

- Соблюдение заданных сроков $f''(x)$ – доля часов, приходящихся на целевой период T^*.
- Равномерность распределенной загрузки $f'''(x)$ – дисперсия показателей занятости у работников по периодам относительно общей доступности.

Интегральная целевая функция ГА представляет собой взвешенную сумму баллов, которую нужно максимизировать:

$$F(x) = W_b f'(x) + W_t f''(x) + W_c f'''(x), \quad (1)$$

где W_b – вес критерия соответствия целевому бюджету, W_t – вес критерия соответствия целевому периоду, W_c – вес критерия равномерности распределения. Все вышеупомянутые веса определяются ЛПР в момент задания первоначальных значений.

С учетом специфики задачи реализованы следующие операторы ГА.

- Инициализация. Начальная популяция формируется с учетом целевого периода T^*: большинству работников назначаются небольшие нагрузки в T^*, а вне его – лишь случайные малые или нулевые назначения. Это обеспечивает наличие популяций x, заведомо близких к оптимальному $f''(x)$, что ускоряет накопление подходящих решений.
- Селекция. Применяется турнирный отбор с сохранением лучших решений.
- Кроссовер. Используются равномерный и арифметические кроссоверы с высокой вероятностью. В случае, если «потомок» практически идентичен одному из «родителей», применяются дополнительные локальные мутации для поддержания разнообразия.
- Мутация. Оператор мутации реализует адаптивное перераспределение времени. Изменения нацелены на увеличение доли часов в T^* и снижение загрузки вне него.
- Ремонт решений. После применения вышеперечисленных операторов выполняется коррекция x для строгого соблюдения ограничения требуемого объема L.

Для управления эволюцией в режиме реального времени введен специальный актор-конфигуратор, осуществляющий динамическую адаптацию параметров ГА. Он анализирует промежуточные результаты и регулирует:

- размер популяции при перерасходе времени;
- вероятность мутации при стагнации/деградации значения целевой функции (1);
- жесткость штрафов за выход за бюджет;
- число поколений при систематическом превышении временных лимитов и при стагнации/деградации несмотря на увеличение вероятности мутации.

Тем самым реализуется гибридная схема, объединяющая эволюционный поиск и управляющую логику высшего уровня.

Отбор и многокритериальная агрегация АР

Локальные ГА, запущенные параллельно на всех OperationActor, генерируют множество частных альтернатив. Декартово произведение всех локальных вариантов приводит к экспоненциальному росту числа глобальных АР, поэтому применяется двухэтапная фильтрация:

1) фильтр перегрузки ресурсов – недопустимые комбинации, в которых нарушаются ограничения по рабочему времени, отбрасываются;

2) фильтр сходных АР – для устранения почти идентичных планов используется локально-чувствительное хеширование. АР с совпадающими хеш-кодами группируются, в каждой группе остается один представитель. Тем самым достигается компрессия пространства решений до управляемого объема.

Оставшееся множество АР подвергается многокритериальному анализу. В отличие от жесткой свертки критериев используются методы превосходства, которые специально разработаны для задач, где критерии разнородны, а предпочтения ЛПР выражаются через пороги и зоны безразличия [10]. В рамках МАС реализованы два метода, позволяющих путем попарного сравнения альтернатив α_1, α_2 упорядочить все АР:

- ELECTRE III, который позволяет учитывать пороги безразличия, предпочтения и вето по каждому критерию [11]. Для каждого критерия задаются пороги: порог безразличия q_i, порог предпочтения p_i и порог вето v_i для каждого критерия k_i, при этом $q_i \leq p_i \leq v_i$. Эти пороги могут также калиброваться на основе обратной связи от ЛПР. Затем вычисляются локальные индексы согласия и несогласия, на основе которых определяется глобальный индекс достоверности

$\sigma(\alpha_1, \alpha_2)$. Отношения превосходства α_1 и α_2 принимаются, если $\sigma(\alpha_1, \alpha_2)$ больше порога достоверности, влияющего на строгость отбора, автоматически устанавливаемый, а также динамически корректируемый специальным актором. Это дает возможность работать с ситуациями, когда небольшое ухудшение по одному критерию несущественно, а крупное – блокирует превосходство альтернативы;

- PROMETHEE II, который вычисляет положительные и отрицательные потоки предпочтений для каждого АР, обеспечивая полное ранжирование [12]. При этом, как и в предыдущем методе, для каждой пары формируется кусочно-линейная функция безразличия с задаваемыми порогами q_i и p_i.

Таким образом, выбранная стратегия сочетает в себе два взаимодополняющих метода. Такое сочетание в совокупности с предлагаемым механизмом фильтрации позволяет реализовать отбор и агрегацию АР, используя контролируемое количество вычислений, значит, и условно контролируемый промежуток времени. При этом обеспечиваются прозрачность и адаптивность с помощью возможности тонкой настройки критериев фильтрации и ранжирования.

РЕЗУЛЬТАТЫ

Экспериментальная апробация разработанных моделей и алгоритмов проведена на примере задач планирования производственных работ, приближенных к реальным. Для оценки эффективности системы было организовано сравнительное тестирование на двух вариантах процесса ПР:

- базовом (ручном), в ходе которого ЛПР самостоятельно составляли план с использованием электронных таблиц и форм;
- с использованием МАС, в рамках которого ЛПР вводили исходные данные в разработанную систему, после чего МАС автоматически генерировала, фильтровала и ранжировала АР. Эксперт анализировал предложенные варианты и утверждал окончательное решение.

В рамках эксперимента каждый участник выполнял планирование по каждому сценарию в обоих режимах, порядок прохождения сценариев был сбалансирован. Использовались пять тестовых сценариев различной мощности (2 малых, 2 средних и 1 большой), отличающихся числом операций (12, 25, 45), количеством работников (12, 22, 40), длиной планового периода (3, 6, 12) и отношением суммарной трудоемкости к доступному фонду рабочего времени (9–49%).

Каждый раз фиксировались:

- время, затраченное на подготовку плана T_{plan} (минут) при заранее подготовленных исходных данных;
- предложенные критерии оптимальности плана: f', f'' и f''';
- число перегрузок работников N_{over} – среднее арифметическое от количества перегрузок работников в данном режиме у всех участников.

Табл. 3. Полученные показатели эффективности

Набор данных	Метод	T_{plan}, мин.	f', %	f'', %	f''', %	N_{over}
Большой А	Базовый	122	2.8	100	92.5	0
	МАС	4	1.0	100	99.9	0
Средний А	Базовый	91	2.1	100	94.3	0.4
	МАС	3	0.9	100	97.6	0
Средний Б	Базовый	82	3.5	100	93.1	0
	МАС	3	1.0	100	96.9	0
Малый А	Базовый	30	1.2	100	94.3	0
	МАС	2	0.3	100	98.1	0
Малый Б	Базовый	55	1.9	100	82.8	1.2
	МАС	2	1.4	100	93.6	0

На больших и средних сценариях применение МАС позволило сократить время подготовки планов примерно в 30 раз по сравнению с ручным режимом. На малых сценариях ускорение составило 20–25 раз. При этом критерий f'' был на уровне 100% как при ручном, так и автоматизированном планировании. Однако критерии f' и f''' были заметно лучше в планах, сгенерированных с помощью МАС: отклонение от бюджета оставалось в пределах 1.5% против 3.5% в ручном режиме, показатель равномерности был в пределах 93.6–99.9% против 82.8–94.3% в базовом сценарии. В планах, построенных МАС, N_{over} всегда было нулевым, тогда как в ручном режиме фиксировались отдельные случаи превышений, что согласуется с общими выводами о потере контроля многочисленных ограничений при ручном планировании.

Таким образом, разработанная система не только радикально снижает временные затраты на планирование, но и обеспечивает строгое соблюдение ограничений и улучшение ключевых критериев качества расписаний.

ОБСУЖДЕНИЕ

Полученные результаты демонстрируют, что объединение когнитивного моделирования, мультиагентной архитектуры и гибридного эволюционно-многокритериального алгоритма позволяет существенно повысить эффективность и качество планирования в сложных производственных системах.

Когнитивный аспект проявляется на уровне представления знаний и структуры системы. Онтологическая модель предметной области и ее программная реализация в виде сети агентов обеспечивают прозрачность и интерпретируемость решений аналогично тому, как онтологии используются для организации знаний и сервисов в цифровых библиотеках и научных информационных системах [13]. Пользователь имеет возможность соотнести элементы плана с понятными сущностями, видеть вклад отдельных агентов и прослеживать причинно-следственные связи, что соответствует тенденции к созданию объяснимых СППР [14].

Мультиагентный подход позволяет рассматривать задачу планирования как распределенный процесс, где за фрагменты решения отвечают независимые акторы. Подобные подходы успешно применяются в имитационных моделях сложных систем, например, при анализе сценариев управления эпидемиями [15]. В данном случае мультиагентная архитектура обеспечивает масштабируемость и устойчивость к сбоям, а также естественный механизм параллелизации вычислений.

Гибридный эволюционно-многокритериальный алгоритм сочетает в себе сильные стороны эвристических методов и строгих методов многокритериального анализа. Локальный ГА дает гибкий механизм генерации альтернатив, хорошо зарекомендовавший себя в задачах составления расписания и управления персоналом [7]. При этом поиск решений происходит в соответствии с представлениями ЛПР об оптимуме согласно современным рекомендациям о методах обработки информации в информационно-аналитических системах поддержки интеллектуальной деятельности [16].

Сопоставив предлагаемый подход с существующими решениями, можно отметить, что большинство традиционных СППР ориентировано либо на информационную поддержку без оптимизации, либо на жестко заданные алгоритмы без учета когнитивных предпочтений и ограничений ЛПР [14]. В то же время работы по мультиагентному моделированию и онтологическому описанию предметных областей показывают возможности повышения интеллектуальности и гибкости систем, но не всегда объединяют эти идеи с многокритериальными алгоритмами. Разработанная система призвана заполнить данный пробел, предлагая вышеописанное комплексное решение.

Ограничением текущей реализации является отсутствие автоматического обучения на исторических данных. Однако ее архитектура допускает интеграцию методов машинного обучения, что согласуется с современными тенденциями развития интеллектуальных СППР [8]. Перспективным направлением также является дальнейшее исследование влияния системы на когнитивную нагрузку ЛПР с целью апробации ее эффективности в контексте снижения данной нагрузки.

ЗАКЛЮЧЕНИЕ

Представлен и экспериментально проверен комплексный подход к поддержке принятия решений при ПР на основе МАС с гибридным эволюционно-многокритериальным ядром. Перечислим следующие основные результаты.

1. Разработана мультиагентная архитектура, обеспечивающая модульность, масштабируемость и когнитивную интерпретируемость решений.
2. Разработан гибридный эволюционно-многокритериальный алгоритм генерации АР.
3. Разработан двухэтапный механизм фильтрации и глобальный анализ методами ELECTRE и PROMETHEE, что позволяет учитывать разнородные критерии и предпочтения ЛПР.
4. Показано, что использование разработанной системы позволяет сократить время подготовки планов в 20–30 раз по сравнению с ручным планированием при сохранении или улучшении качества решений и полном соблюдении ограничений.

Полученные результаты могут быть использованы при проектировании и внедрении интеллектуальных СППР на производственных предприятиях, в проектном управлении и других областях, где требуется совместный учет множества критериев и ограничений в процессе ПР.

СПИСОК ЛИТЕРАТУРЫ

1. *Brown A.* Reactive Applications with Akka.NET. N.Y.: Manning Publications Co., 2019. 280 p.

2. The Reactive Manifesto [Электронный ресурс]. URL: https://www.reactivemanifesto.org/ru (12.11.2025).

3. *Dauzère-Pérès S., Ding J., Shen L., Tamssaouet K.* The flexible job shop scheduling problem: A review // European Journal of Operational Research. 2024. Vol. 314, No. 2. P. 409–432. https://doi.org/10.1016/j.ejor.2023.05.017

4. *Caselli G., Delorme M., Iori M., Magni C.A.* Exact algorithms for a parallel machine scheduling problem with workforce and contiguity constraints // Computers & Operations Research. 2024. Vol. 163, No. 3. https://doi.org/10.1016/j.cor.2023.106484

5. *Xiong H., Shi S., Ren D., Hu J.* A survey of job shop scheduling problem: The types and models // Computers & Operations Research. 2022. Vol. 142, No. 2. https://doi.org/10.1016/j.cor.2022.105731

6. *Gu H., Zhang Y., Zinder Y.* An efficient optimisation procedure for the workforce scheduling and routing problem: Lagrangian relaxation and iterated local search // Computers & Operations Research. 2022. Vol. 144. https://doi.org/10.1016/j.cor.2022.105829

7. *Borgonjon T., Maenhout B.* A genetic algorithm for the personnel task rescheduling problem with time preemption // Expert Systems with Applications. 2024. Vol. 238. https://doi.org/10.1016/j.eswa.2023.121868

8. *Thiruvady D., Nguyen S., Sun Y., Shiri F., Zaidi N., Li X.* Adaptive population-based simulated annealing for resource constrained job scheduling with uncertainty // International Journal of Production Research. 2024. Vol. 62, No. 17. P. 6227–6250. https://doi.org/10.1080/00207543.2024.2311183

9. *Gad A.G.* Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review // Archives of Computational Methods in Engineering. 2022. Vol. 29, No. 5. P. 2531–2561. https://doi.org/10.1007/s11831-021-09694-4

10. *Чечнев В.Б.* Анализ и классификация многокритериальных методов принятия решений // Онтология проектирования. 2024. Т. 14, №4(54). С. 607–624. https://doi.org/10.18287/2223-9537-2024-14-4-607-624

11. *Roy B.* The outranking approach and the foundations of ELECTRE methods // Theory and Decision. 1991. Vol. 31, No. 1. P. 49–73. https://doi.org/10.1007/BF00134132

12. *Brans J.P., Vincke Ph., Mareschal B.* How to select and how to rank projects: The PROMETHEE method // European Journal of Operational Research. 1986. Vol. 24, No. 2. P. 228–238. https://doi.org/10.1016/0377-2217(86)90044-5

13. *Атаева О.М., Калёнов Н.Е.,* Серебряков В.А. Онтологический подход к описанию единого цифрового пространства научных знаний // Электронные библиотеки. 2021. Т. 24, № 1. С. 3–19. https://doi.org/10.26907/1562-5419-2021-24-1-3-19

14. *Чечнев В.Б.* Использование систем поддержки принятия решений в автоматизации процессов принятия решений // Электронные библиотеки. 2025. Т. 28, № 1. С. 163–183. https://doi.org/10.26907/1562-5419-2025-28-1-163-183

15. *Балута В.И., Осипов В.П., Сивакова Т.В.* Предложения по разработке средств повышения эффективности управления в условиях эпидемий // Электронные библиотеки. 2021. Т. 24, № 1. С. 20–41. https://doi.org/10.26907/1562-5419-2021-24-1-20-41

16. *Цибизова Т.Ю., Ляпунцова Е.В., Макарова М.П. и др.* Когнитивное моделирование. М.: МГТУ им. Н.Э. Баумана, 2025. 252 с.

METHODS OF COGNITIVE MODELING AND HYBRID EVOLUTIONARY MULTI-CRITERIA ALGORITHMS IN A MULTI-AGENT INFORMATION-ANALYTICAL SYSTEM

V. B. Chechnev[0009-0000-1523-3294]

Bauman Moscow State Technical University, Moscow, Russia

gegrev@yandex.ru

Abstract

The paper proposes an approach to multi-criteria decision support based on a cognitively oriented multi-agent information-analytical system. Cognitive modeling methods are developed, including a formal ontological representation of knowledge about production planning and a coalition–holonic agent architecture that ensures adaptability and transparency of computations. A hybrid evolutionary multi-criteria algorithm is introduced, in which agents generate alternative plans at the local level using a parallel genetic algorithm that optimizes a combination of several criteria. At the global level, a multi-stage selection of alternatives is implemented with filtering of resource overloads and similar solutions, followed by final aggregation using the PROMETHEE and ELECTRE multi-criteria decision-making methods.

An experimental study is carried out comparing manual planning with planning supported by the developed system, as well as analyzing the impact of dynamic adaptation of the genetic algorithm parameters. The results show that the use of the system makes it possible to reduce plan generation time by a factor of 20–30 while maintaining or improving solution quality. At the same time, resource overloads are completely eliminated, and early termination of evolutionary computations is ensured without loss of solution quality. The system and proposed algorithms are intended for use in planning project activities at manufacturing enterprises.

Keywords*: cognitive modeling, decision support systems, multi-agent systems, genetic algorithm, information systems, multi-criteria optimization, workforce workload planning.*

REFERENCES

1. *Brown A.* Reactive Applications with Akka.NET. N.Y.: Manning Publications Co., 2019. 280 p.

2. The Reactive Manifesto. URL: https://www.reactivemanifesto.org/ru (12.11.2025).

3. *Dauzère-Pérès S., Ding J., Shen L., Tamssaouet K.* The flexible job shop scheduling problem: A review // European Journal of Operational Research. 2024. Vol. 314, No. 2. P. 409–432. https://doi.org/10.1016/j.ejor.2023.05.017

4. *Caselli G., Delorme M., Iori M., Magni C.A.* Exact algorithms for a parallel machine scheduling problem with workforce and contiguity constraints // Computers & Operations Research. 2024. Vol. 163, No. 3. https://doi.org/10.1016/j.cor.2023.106484

5. *Xiong H., Shi S., Ren D., Hu J.* A survey of job shop scheduling problem: The types and models // Computers & Operations Research. 2022. Vol. 142, No. 2. https://doi.org/10.1016/j.cor.2022.105731

6. *Gu H., Zhang Y., Zinder Y.* An efficient optimization procedure for the workforce scheduling and routing problem: Lagrangian relaxation and iterated local search // Computers & Operations Research. 2022. Vol. 144. https://doi.org/10.1016/j.cor.2022.105829

7. *Borgonjon T., Maenhout B.* A genetic algorithm for the personnel task rescheduling problem with time preemption // Expert Systems with Applications. 2024. Vol. 238. https://doi.org/10.1016/j.eswa.2023.121868

8. *Thiruvady D., Nguyen S., Sun Y., Shiri F., Zaidi N., Li X.* Adaptive population-based simulated annealing for resource constrained job scheduling with uncertainty // International Journal of Production Research. 2024. Vol. 62, No. 17. P. 6227–6250. https://doi.org/10.1080/00207543.2024.2311183

9. *Gad A.G.* Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review // Archives of Computational Methods in Engineering. 2022. Vol. 29, No. 5. P. 2531–2561. https://doi.org/10.1007/s11831-021-09694-4

10. *Chechnev V.B.* Analiz i klassifikatsiya mnogokriterial'nykh metodov prinyatiya resheniy // Ontologiya proektirovaniya. 2024. Vol. 14, No. 4(54). P. 607–624 (In Russian). https://doi.org/10.18287/2223-9537-2024-14-4-607-624

11. *Roy B.* The outranking approach and the foundations of ELECTRE methods // Theory and Decision. 1991. Vol. 31, No. 1. P. 49–73. https://doi.org/10.1007/BF00134132

12. *Brans J.P., Vincke P., Mareschal B.* How to select and how to rank projects: The PROMETHEE method // European Journal of Operational Research. 1986. Vol. 24, No. 2. P. 228–238. https://doi.org/10.1016/0377-2217(86)90044-5

13. *Ataeva O.M., Kalyonov N.E., Serebryakov V.A.* Ontologicheskiy podkhod k opisaniyu edinogo tsifrovogo prostranstva nauchnykh znaniy // Elektronnye biblioteki. 2021. Vol. 24, No. 1. P. 3–19 (In Russian). https://doi.org/10.26907/1562-5419-2021-24-1-3-19

14. *Chechnev V.B.* Ispol'zovanie sistem podderzhki prinyatiya resheniy v avtomatizatsii protsessov prinyatiya resheniy // Elektronnye biblioteki. 2025. Vol. 28, No. 1. P. 163–183 (In Russian). https://doi.org/10.26907/1562-5419-2025-28-1-163-183

15. *Baluta V.I., Osipov V.P., Sivakova T.V.* Predlozheniya po razrabotke sredstv povysheniya effektivnosti upravleniya v usloviyakh epidemiy // Elektronnye biblioteki. 2021. Vol. 24, No. 1. P. 20–41 (In Russian). https://doi.org/10.26907/1562-5419-2021-24-1-20-41

16. *Tsibizova T.Y., Lyapuntsova E.V., Makarova M.P. et al.* Kognitivnoe modelirovanie. M.: MGTU im. N.E. Baumana, 2025. 252 pp. (In Russian).

СВЕДЕНИЯ ОБ АВТОРЕ

ЧЕЧНЕВ Василий Борисович – аспирант Московского государственного технического университета им. Н.Э. Баумана, специальность «Когнитивное моделирование».

CHECHNEV Vasiliy Borisovich – Postgraduate student at Bauman Moscow State Technical University, specialty «Cognitive modeling».

email: gegrev@yandex.ru

ORCID: 0009-0000-1523-3294

Материал проступил в редакцию 19 декабря 2025 года