ISSN 2949-1215

Российская Академия Наук

ТРУДЫ

Кольского научного центра РАН

3/2025(16)

СЕРИЯ: ТЕХНИЧЕСКИЕ НАУКИ

0+

ISSN 2949-1215
0+

Российская Академия Наук

ТРУДЫ

Кольского научного центра. Серия: Технические науки

3/2025(16)

Научно-информационный журнал
Основан в 2010 году
Выходит 4 раза в год

Учредитель — Федеральное государственное бюджетное учреждение науки Федеральный исследовательский центр «Кольский научный центр Российской академии наук» (ФИЦ КНЦ РАН)

Свидетельство о регистрации СМИ ПИ № ФС77-83502 от 30 июня 2022 г. выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций

Научное издание

Редактор Е. Н. Еремеева
Технический редактор В. Ю. Жиганов

Подписано к печати 23.12.2025.
Дата выхода в свет 25.12.2025.
Формат бумаги 60×84 ⅛.
Усл. печ. л. 20,46. Заказ № 97. Тираж 300 экз.
Свободная цена.

Адрес учредителя, издателя, редакции и типографии:
Федеральное государственное бюджетное учреждение науки
Федеральный исследовательский центр
«Кольский научный центр РАН» (ФИЦ КНЦ РАН).
184209, г. Апатиты, Мурманская область, ул. Ферсмана, 14.
Тел.: 8 (81555) 7-53-50; 7-95-95, факс: 8 (81555) 7-64-25.
Электронная почта: ksc@ksc.ru
Сайт: www.ksc.ru

СОДЕРЖАНИЕ

Федоров А. М., Датьев И. О., Илясов М. О., Вишняков И. Г., Базегский М. О., Фигуркин Д. С., Любимова К. Д.	Технология оперативной разработки интеллектуальных информационных систем на основе коммуникационных возможностей больших языковых моделей	5
Ломов П. А., Маслобоев А. В., Олейник А. Г.	Подход к формированию онтологии цикла управления жизнеспособностью критических инфраструктур	22
Зуенко О. Н., Фридман О. В.	Аналитический обзор методов кластеризации в подпространствах	35
Шестаков А. В., Зуенко А. А.	Применение RAG-технологии для автоматической генерации тестов и проверки знаний с поддержкой диалога на естественном языке	56
Диковицкий В. В.	Предиктивное моделирование социальных реакций для регионального управления на базе методов объяснимого искусственного интеллекта	71
Горбунов Р. А., Вицентий А. В.	Исследование возможностей больших языковых моделей для извлечения данных из текстов на естественном языке	80
Диковицкий В. В., Шишаев М. Г.	Экспресс-технология формирования обучающей выборки для планиметрического минералогического анализа на основе методов машинного обучения	106
Таран П. В. Зуенко А. А.	Решение задач генеративного дизайна с использованием методов удовлетворения ограничений	117
Бирюков В. В., Олейник А. Г.	CFD-модель агрегирования ферромагнитных частиц под действием магнитного поля в восходящем водном потоке	131
Вицентий А. В.	Моделирование пространственных ситуаций как геосемантических изображений на основе геопространственного графа знаний	140
Мелихов М. В.	Дистанционный метод определения водонасыщенности объектов наземной горной инфраструктуры	154
Яковлев С. Ю., Шемякин А. С.	Идентификация и классификация опасных объектов	162

ISSN 2949-1215
0+

3/2025(16)

Scientific journal
Published since 2010
Publication frequency — four times a year

Russian Academy of Sciences
TRANSACTIONS
Kola Science Centre. Series: Engineering Sciences

Founder — Federal State Budget Institution of Science
Federal Research Centre “Kola Science
Centre of the Russian Academy of Sciences”

Mass Media Registration Certificate
ПИ No. ФС77-83502 issued by the Federal Service for Supervision
of Communications, Information Technology and Mass Media
on June, 30 2022

Scientific publication

Editor Ye. N. Yeremeyeva
Technical Editor V. Yu. Zhiganov

14, Fersman str., Apatity, Murmansk region, 184209, Russia.
Tel.: 8 (81555) 7-93-80. Fax: 8 (81555) 7-64-25.
E-mail: ksc@ksc.ru
www.ksc.ru

CONTENTS

Fedorov A. M., Datyev I. O., Ilyasov M. O., Vishnyakov I. G., Bazegskiy M. O., Figurkin D. S., Lyubimova C. D.	Technology for developing intelligent information systems based on the communicative capabilities of large language models…….....	5
Lomov P. A., Masloboev A. V., Oleynik A. G.	An approach to ontology design of the critical infrastructure resilience management cycle…….....	22
Zuenko O. N., Fridman O. V.	Analytical review of clustering methods in subspaces….....	35
Shestakov A. V., Zuenko A. A.	Application of RAG technology for automated test generation and knowledge assessment with natural language dialogue support.....	56
Dikovitsky V. V.	Predictive modeling of social reactions for regional management based on explainable artificial intelligence methods……..	71
Gorbunov R. A, Vicentiy A. V.	Research of the capabilities of large language models for extracting data from natural language texts……....	80
Dikovitsky V. V., Shishaev M. G.	Express technology for forming a training set for planimetric mineralogical analysis based on machine learning methods.....	106
Taran P. V., Zuenko A. A.	Solving generative design problems using constraint satisfaction methods…….....	117
Biryukov V. V., Oleynik A. G.	CFD-model for ferromagnetic particle aggregation under the magnetic field influence in an ascending water flow…….....	131
Vicentiy A. V.	Modeling of spatial situations as geosemantic images based on the graph of geospatial knowledge……..	140
Melikhov M. V.	Remote sensing method for determining water saturation of ground-based mining infrastructure facilities….....	154
Yakovlev S. Yu., Shemyakin A. S.	Identification and classification of hazardous objects……	162

Научная статья
УДК 004.89
doi:10.37614/2949-1215.2025.16.3.001

ТЕХНОЛОГИЯ ОПЕРАТИВНОЙ РАЗРАБОТКИ ИНТЕЛЛЕКТУАЛЬНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ НА ОСНОВЕ КОММУНИКАЦИОННЫХ ВОЗМОЖНОСТЕЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Андрей Михайлович Федоров[1], Игорь Олегович Датьев[2], Михаил Олегович Илясов[3], Иван Геннадьевич Вишняков[4], Марк Олегович Базегский[5], Даниил Сергеевич Фигуркин[6], Кристина Дмитриевна Любимова[7]
[1–6]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[7]Филиал федерального государственного автономного образовательного учреждения высшего образования «Мурманский арктический университет» в г. Апатиты, Апатиты, Россия
[1]a.fedorov@ksc.ru, https://orcid.org/0000-0002-2862-7994
[2]i.datyev@ksc.ru, https://orcid.org/0000-0002-8372-8704
[3]m.ilyasov@ksc.ru, https://orcid.org/0009-0006-9065-2396
[4]i.vishnyakov@ksc.ru, https://orcid.org/0009-0003-4938-5693
[5]markbazegskiy@yandex.ru, https://orcid.org/0000-0002-1543-3963
[6]DaniilSF@yandex.ru, https://orcid.org/0000-0002-9743-8222
[7]KLyubimova2002@mail.ru, https://orcid.org/0000-0003-2710-6481

Аннотация

Автоматизация процессов разработки информационных технологий является актуальной задачей. Современные тенденции развития требуют быстрой адаптации решений к новым условиям, однако традиционная практика классической разработки и модернизации прежних решений обладает рядом известных недостатков, ограничивающих производительность и конкурентоспособность проектов. Настоящее исследование направлено на разработку технологии оперативного создания информационных систем путем компонентной интеграции уже имеющихся разработок и интеллектуальных ассистентов и агентов. Основным инструментом такой интеграции предложено использовать стандартный протокол MCP, расширяющий коммуникационные возможности технологий, моделей и средств реализации искусственного интеллекта. Предложенное решение способствует существенному улучшению временных показателей концептуального проектирования и разработки прикладных и исследовательских информационных систем, а также повышает уровень их интеллектуализации и взаимной интеграции. Полученная технология представляет интерес для исследователей и разработчиков в сфере системного анализа, управления и обработки информации.

Ключевые слова:

информационные системы, большие языковые модели, tool calling, протокол MCP, программная разработка, агентные технологии

Благодарности:

исследование выполнено в рамках государственного задания Института информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук от Министерства науки и высшего образования РФ, тема научно-исследовательской работы: «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределённости и риска» (шифр темы FMEZ-2025-0053).

Для цитирования:

Технология оперативной разработки интеллектуальных информационных систем на основе коммуникационных возможностей больших языковых моделей / А. М. Федоров [и др.] // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 5–21. doi:10.37614/2949-1215.2025.16.3.001.

Original article

TECHNOLOGY FOR DEVELOPING INTELLIGENT INFORMATION SYSTEMS BASED ON THE COMMUNICATIVE CAPABILITIES OF LARGE LANGUAGE MODELS

Andrey M. Fedorov[1], Igor O. Datyev[2], Mikhail O. Ilyasov[3], Ivan G. Vishnyakov[4], Mark O. Bazegskiy[5], Daniil S. Figurkin[6], Christina D. Lyubimova[7]
[1–6]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[7]Apatity branch of Murmansk Arctic University, Apatity, Russia

[1]*a.fedorov@ksc.ru, https://orcid.org/0000-0002-2862-7994*
[2]*i.datyev@ksc.ru, https://orcid.org/0000-0002-8372-8704*
[3]*m.ilyasov@ksc.ru, https://orcid.org/0009-0006-9065-2396*
[4]*i.vishnyakov@ksc.ru, https://orcid.org/0009-0003-4938-5693*
[5]*markbazegskiy@yandex.ru, https://orcid.org/0000-0002-1543-3963*
[6]*DaniilSF@yandex.ru, https://orcid.org/0000-0002-9743-8222*
[7]*KLyubimova2002@mail.ru, https://orcid.org/0000-0003-2710-6481*

Abstract

Automating information technology development processes is a pressing issue. Current development trends require solutions to quickly adapt to new conditions, but traditional practices of classical development and modernization of existing solutions have a number of known shortcomings that limit the productivity and competitiveness of projects. This study aims to develop a technology for the rapid creation of information systems through the component integration of existing developments and intelligent assistants and agents. The proposed primary tool for such integration is the standard MCP protocol, which expands the communication capabilities of artificial intelligence technologies, models, and implementation tools. The proposed solution significantly improves the timeframe for conceptual design and development of applied and research information systems, while also increasing their intellectualization and mutual integration. The resulting technology is of interest to researchers and developers in the fields of systems analysis, management, and information processing.

Keywords:

information systems, large language models, tool calling, MCP protocol, software development, agent technologies

Acknowledgments:

The study was carried out within the framework of the state assignment of the Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Center of the Russian Academy of Sciences from the Ministry of Science and Higher Education of the Russian Federation, research topic: "Methods and technologies for creating intelligent information systems to support the development of complex dynamic systems with regional specificity in conditions of uncertainty and risk" (topic code FMEZ-2025-0053).

For citation:

Fedorov A. M., Datyev I.O., Ilyasov M. O., Vishnyakov I. G., Bazegskiy M. O., Figurkin D. S., Lyubimova C. D. Technology for developing intelligent information systems based on the communicative capabilities of large language models. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 5–21. doi:10.37614/2949-1215.2025.16.3.001.

Введение

Современные информационные технологии активно развиваются благодаря появлению новых возможностей в области технологий искусственного интеллекта (ИИ). Большие языковые модели внесли существенные коррективы в процессы разработки и взаимодействия как отдельных элементов пользовательского интерфейса информационных систем, так и их составных компонентов и протоколов внутренней и межсистемной коммуникации. Известно и вполне закономерно то, что немалая доля актуальных разработок представляет собой интеграцию уже существующих и новых решений. В общем случае эти составные компоненты гетерогенны, в некоторых уже используются элементы технологий ИИ. Эта разнородность делает процесс необходимой интеграции трудоемким и требующим неординарной настройки взаимосвязей между отдельными частями такого типа разрабатываемых информационных систем. Таким образом, исследователи и разработчики сталкиваются со сложностями в процессе оперативного создания новых систем и технологий. Этот факт повышает актуальность развития научно-практической базы, направленной на оптимизацию механизмов системной интеграции, применение эффективных интерфейсов взаимодействия компонентов и широкое использование ресурсов ИИ на определенной стандартизированной основе.

В настоящей работе представлена разработанная и опробованная технология оперативного создания прототипов информационных систем путем компонентной интеграции существующих разработок и интеллектуальных агентов с использованием современных инструментов ИИ, таких как языковые модели, и их стандартных средств обеспечения межсистемных коммуникаций на примере протокола MCP.

Значительная часть новых информационных систем и технологий создается путем объединения или расширения существующих решений и добавления в них новых функций, в том числе реализованных на базе технологий ИИ. В большом количестве случаев этот процесс сопряжен с рядом недостатков, среди которых: отсутствие единого унифицированного подхода к построению архитектурных

схем, усложняющее интеграцию компонентов; высокая продолжительность разработки от момента зарождения идеи до готовности прототипа информационной системы; использование для межкомпонентных коммуникаций разнородных технологий и средств, в том числе собственной разработки.

Устранение этих и других недостатков позволит более эффективно развивать современные информационные технологии как на концептуальном, так и на прикладном уровне.

Данное направление исследований приобрело еще большую актуальность в процессе научно-исследовательской работы коллектива авторов по теме «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределенности и риска» [1].

Развиваемое в этом исследовании тематическое направление фокусированного сбора данных [2] потребовало отдельной комплексной проработки на концептуальном и прикладном уровнях вопросов, заявленных в данной статье.

Более того, вопросы интеграции технологий ИИ в научные и практические разработки в настоящее время стоят очень остро и требуют от их реальных и потенциальных пользователей широкого спектра теоретических знаний, практического умения и опыта использования. Динамика развития в этой сфере очень высокая. Например, за время подготовки данной статьи из существовавших независимо друг от друга двух коммуникационных протоколов взаимодействия ИИ-агентов остался только один, что внесло свои коррективы в список готовых к публикации результатов.

Благодаря предложенному в данной работе компонентному интеграционному подходу и использованию коммуникационных свойств больших языковых моделей время на разработку ИС уменьшается, эффект от повторного использования компонентов повышается, затраты на замену технологического стека снижаются.

Научная новизна предложенного решения заключается в комплексном подходе к созданию информационных систем, предусматривающем сочетание традиционного опыта разработки и передовых методов интеграции технологий ИИ, а именно больших языковых моделей, протокола MCP, механизма tool calling и техник формирования контекстных инструкций (промпт-инжиниринга). Разработанная технология является развитием практики применения системного анализа и обработки информации для решения проблемы оперативного создания прототипов информационных систем с использованием интеллектуальных агентов и языковых моделей. Методы и средства, применяемые в исследовании, соответствуют современным требованиям системного анализа, теории принятия решений и искусственного интеллекта.

В данной работе решается задача разработки технологии оперативного построения прототипов информационных систем с использованием элементов и технологий ИИ. Формально задача поставлена следующим образом: дана исходная информационная система (ИС), в которую необходимо интегрировать элементы технологии ИИ. В качестве ИИ используются большие языковые модели, технические характеристики которых позволяют им взаимодействовать с внешними программно-алгоритмическими модулям через механизм tool calling.

Необходимо разработать последовательность действий (план, схему, архитектуру) оперативной интеграции в заданную ИС элементов технологии ИИ и определить перечень подходящих и доступных элементов. Важным требованием к планированию разработки является использование стандартных или уже разработанных функциональных компонентов (ФК), готовых к интеграции.

Для достижения указанных характеристик предлагается использовать способ интеграции в языковые модели с помощью универсального протокола MCP внешних функций, называемых функциональными инструментами (ФИ). Всё перечисленное выше является составной частью разработанной технологии.

Протокол MCP

Протокол контекста модели (MCP) — это развивающийся открытый стандарт, разработанный для обеспечения бесшовной, безопасной и динамической интеграции между моделями ИИ, особенно большими языковыми моделями (LLM) [3; 4], и внешними инструментами, источниками данных и различными средами. MCP был создан компанией Anthropic для стандартизации того, как модели ИИ взаимодействуют с внешними разнородными ресурсами, обеспечивая пополнение возможностей

модели в реальном времени с учетом контекста [5; 6]. MCP фокусируется исключительно на протоколе обмена контекстом — он не определяет, как приложения ИИ используют LLM или управляют предоставленным контекстом. MCP решает давние проблемы взаимодействия систем ИИ, управления контекстом и безопасного расширения возможностей моделей. Предназначен для построения мультиагентных систем для решения задач в различных предметных областях как в научных исследованиях, так и в реальном секторе экономики [7; 8].

Архитектура MCP

MCP использует архитектуру [9] клиент-сервер, где хост MCP — ИИ-приложение, устанавливает соединения с одним или несколькими серверами MCP. Хост MCP создает один клиент MCP для каждого сервера MCP. Каждый клиент MCP поддерживает выделенное индивидуальное соединение с соответствующим сервером MCP.

Ключевыми компонентами архитектуры MCP являются: хост MCP — приложение ИИ, которое координирует и управляет одним или несколькими клиентами MCP; клиент MCP — компонент, который поддерживает соединение с сервером MCP и получает контекст от сервера MCP для использования хостом MCP; сервер MCP — программа, предоставляющая контекст клиентам MCP. MCP-сервер относится к программе, которая обслуживает контекстные данные независимо от места ее работы. MCP-серверы могут работать локально или удаленно.

Два слоя протокола MCP: 1) уровень данных: определяет протокол на основе JSON-RPC для клиент-серверного взаимодействия, включая управление жизненным циклом и основные примитивы, такие как инструменты, ресурсы, запросы и уведомления; 2) транспортный уровень: определяет механизмы связи и каналы, обеспечивающие обмен данными между клиентами и серверами, включая установление транспортного соединения, формирование сообщений и авторизацию.

Уровень данных реализует протокол обмена на основе JSON-RPC 2.0 [10], который определяет структуру и семантику сообщений. Этот уровень включает в себя: 1) управление жизненным циклом: осуществляет инициализацию соединения, согласование возможностей и завершение соединения между клиентами и серверами; 2) серверные функции: позволяет серверам предоставлять основные функции, включая инструменты для действий ИИ, ресурсы для контекстных данных и запросы шаблонов взаимодействия с клиентом и от него; 3) клиентские функции: позволяет серверам запрашивать у клиента выборку из LLM-хоста, получать ввод данных от пользователя и регистрировать сообщения для клиента; 4) служебные функции: поддерживает дополнительные возможности, такие как уведомления об обновлениях в режиме реального времени и отслеживание хода выполнения длительных операций.

Транспортный уровень управляет каналами связи и аутентификацией между клиентами и серверами. Он отвечает за установление соединения, формирование сообщений и защищенную связь между участниками MCP.

MCP поддерживает два транспортных механизма: *Stdio transport*: использует стандартные потоки ввода/вывода для прямого взаимодействия между локальными процессами на одной машине, обеспечивая оптимальную производительность без сетевых издержек; *Streamable HTTP transport*: использует HTTP POST для сообщений клиент-сервер с опциональными событиями, отправленными сервером, для потоковой передачи. Этот транспорт обеспечивает взаимодействие с удаленным сервером и поддерживает стандартные методы HTTP-аутентификации, включая токены-носители, ключи API и настраиваемые заголовки. MCP рекомендует использовать OAuth для получения токенов аутентификации.

Транспортный уровень абстрагирует информацию о взаимодействии от уровня протокола, обеспечивая единый формат сообщений JSON-RPC 2.0 для всех транспортных механизмов.

Протокол уровня данных. Основной частью MCP является определение схемы и семантики между клиентами и серверами MCP. Именно уровень данных, в частности набор примитивов, определяет способы обмена контекстом между серверами MCP и клиентами MCP.

MCP использует JSON-RPC 2.0 в качестве базового протокола RPC. Клиент и серверы отправляют запросы, ответы и уведомления друг другу.

Управление жизненным циклом. Цель управления жизненным циклом — согласовать функции и операции, поддерживаемые клиентом или сервером, такие как инструменты, ресурсы или подсказки, поддерживаемые как клиентом, так и сервером.

Примитивы

Примитивы MCP — важнейший концепт MCP. Примитивы определяют типы контекстной информации, которой можно поделиться с приложениями ИИ, и диапазон доступных действий. MCP определяет три основных примитива, которые могут предоставлять серверы: 1) *инструменты (Tools)* — исполняемые функции, которые приложения ИИ могут вызывать для выполнения действий (например, файловых операций, вызовов API, запросов к базе данных); 2) *ресурсы (Resources)* — источники данных, предоставляющие контекстную информацию приложениям ИИ (например, содержимое файлов, записи базы данных, ответы API); 3) *запросы к языковой модели (Promts, промпты)* — многоразовые шаблоны, помогающие структурировать взаимодействие с языковыми моделями (например, системные промпты, «few-shot» примеры).

Каждому типу примитива соответствуют методы обнаружения (`*/list`), извлечения (`*/get`) и (в некоторых случаях) выполнения (`tools/call`). Клиенты MCP будут использовать методы `*/list` для поиска доступных примитивов. Например, клиент может сначала составить список всех доступных инструментов (`tools/list`), а затем выполнить их. Такая структура позволяет создавать динамические списки.

MCP также определяет примитивы, которые могут предоставлять клиенты. Эти примитивы позволяют разработчикам серверов MCP создавать более сложные взаимодействия.

Выборка (Sampling) позволяет серверам запрашивать дополнения языковой модели из клиентского приложения ИИ. Это полезно, когда авторы серверов хотят получить доступ к языковой модели, но при этом сохранить независимость от модели и не включать SDK языковой модели в свой сервер MCP. Они могут использовать метод `sampling/complete` для запроса дополнения языковой модели из клиентского приложения ИИ.

Извлечение (Elicitation) позволяет серверам запрашивать дополнительную информацию у пользователей. Это полезно, когда авторы серверов хотят получить дополнительную информацию от пользователя или запросить подтверждение действия. Они могут использовать метод `elicitation/request` для запроса дополнительной информации у пользователя.

Ведение журнала (Logging) позволяет серверам отправлять сообщения журнала клиентам для отладки и мониторинга.

Уведомления (Notifications)

Протокол поддерживает уведомления в режиме реального времени для обеспечения динамических обновлений между серверами и клиентами. Так, при изменении доступных инструментов сервера, например при появлении новых функций или изменении существующих инструментов, сервер может отправлять уведомления об обновлениях инструментов, чтобы информировать подключенных клиентов об этих изменениях. Уведомления отправляются в виде уведомлений JSON-RPC 2.0 (без ожидания ответа) и позволяют серверам MCP предоставлять обновления подключенным клиентам в режиме реального времени.

Инициализация

Процесс инициализации является ключевой частью управления жизненным циклом MCP и служит нескольким важным целям:

1. Согласование версии протокола. Поле `protocolVersion` (например, "2025-06-18") гарантирует, что клиент и сервер используют совместимые версии протокола. Это предотвращает ошибки связи, которые могут возникнуть при попытке взаимодействия разных версий. Если взаимосовместимая версия не согласована, соединение следует разорвать.

2. Определение возможностей. Объект `capabilities` позволяет каждой стороне указать поддерживаемые функции, включая [примитивы](#primitives), которые она может обрабатывать (инструменты, ресурсы, подсказки), а также указать, поддерживает ли она такие функции, как [уведомления](#notifications). Это обеспечивает эффективное взаимодействие и позволяет избегать неподдерживаемых операций.

3. Обмен идентификацией. Объекты `clientInfo` и `serverInfo` предоставляют информацию об идентификации и версиях для отладки и обеспечения совместимости.

Во время инициализации менеджер клиентов MCP приложения ИИ устанавливает соединения с настроенными серверами и сохраняет их возможности для последующего использования. Приложение использует эту информацию, чтобы определить, какие серверы могут предоставлять определенные типы функций (инструменты, ресурсы, подсказки) и поддерживают ли они обновления в реальном времени.

Обнаружение инструментов

Запрос «tools/list» прост и не содержит параметров.

Ответ содержит массив «инструменты», который предоставляет полные метаданные о каждом доступном инструменте. Эта структура на основе массива позволяет серверам одновременно предоставлять несколько инструментов, сохраняя при этом четкие границы между различными функциями.

Приложение AI извлекает доступные инструменты со всех подключенных серверов MCP и объединяет их в единый реестр инструментов, к которому имеет доступ языковая модель. Это позволяет LLM получать список доступных действий и автоматически генерировать соответствующие вызовы инструментов во время диалогов.

Ответ на выполнение инструмента

Ответ демонстрирует гибкую систему контента MCP:

1. Массив `content`: ответы инструмента возвращают массив объектов контента, что позволяет создавать ответы в различных форматах (текст, изображения, ресурсы и т. д.).

2. Типы контента: каждый объект контента имеет поле `type`. В этом примере `"type": "text"` обозначает простой текстовый контент, но MCP поддерживает различные типы контента для различных вариантов использования.

3. Структурированный вывод: ответ предоставляет полезную информацию, которую приложение ИИ может использовать в качестве контекста для взаимодействия с языковой моделью.

Этот шаблон выполнения позволяет приложениям ИИ динамически вызывать функции сервера и получать структурированные ответы, которые можно интегрировать в диалоги с языковыми моделями.

Когда языковая модель решает использовать инструмент во время диалога, приложение ИИ перехватывает вызов инструмента, направляет его на соответствующий сервер MCP, выполняет его и возвращает результаты обратно в LLM в рамках потока диалога. Это позволяет LLM получать доступ к данным в режиме реального времени и выполнять действия во внешнем мире.

MCP поддерживает уведомления в режиме реального времени, позволяющие серверам информировать клиентов об изменениях без явного запроса. Это демонстрирует систему уведомлений — ключевую функцию, обеспечивающую синхронизацию и быстрое реагирование подключений MCP.

Уведомления об изменении списка инструментов

При изменении доступных инструментов сервера, например при появлении новых функций, изменении существующих инструментов или их временной недоступности, сервер может заблаговременно уведомить подключенных клиентов.

Ключевые особенности уведомлений MCP:

1. Ответ не требуется: в уведомлении отсутствует поле `id`. Это соответствует семантике уведомлений JSON-RPC 2.0, где ответ не ожидается и не отправляется.

2. На основе возможностей: Это уведомление отправляется только серверами, которые объявили "listChanged": true в своих возможностях инструментов во время инициализации (как показано на шаге 1).

3. Управляемое событиями: сервер решает, когда отправлять уведомления, основываясь на изменениях внутреннего состояния, что делает соединения MCP динамичными и отзывчивыми.

Получив уведомление, клиент обычно реагирует, запрашивая обновленный список инструментов. Это создает цикл обновления, который поддерживает актуальность информации клиента о доступных инструментах:

Эта система уведомлений критически важна по нескольким причинам: 1) динамические среды: инструменты могут появляться и исчезать в зависимости от состояния сервера, внешних зависимостей или прав пользователя; 2) эффективность: клиентам не нужно спрашивать об изменениях, они

уведомляются об обновлениях; 3) консистентность: обеспечивает клиентам постоянную точную информацию о доступных возможностях сервера; 4) совместная работа в реальном времени: обеспечивает адаптивность ИИ-приложений к изменяющимся контекстам.

Этот шаблон уведомлений распространяется не только на инструменты, но и на другие примитивы MCP, обеспечивая комплексную синхронизацию в реальном времени между клиентами и серверами. Когда ИИ-приложение получает уведомление об изменении инструментов, оно обновляет свой реестр инструментов и доступные возможности LLM. Это гарантирует доступ к актуальному набору инструментов, а LLM может динамически адаптироваться к новым функциям по мере их появления.

Вызов инструментов (tool calling)

В последние годы большие языковые модели (LLM) стали широко использоваться в различных приложениях, включая генерацию дополненного поиска (RAG), системы рекомендаций и корпоративные среды. Одной из ключевых возможностей LLM является вызов инструментов (tool calling), который позволяет моделям взаимодействовать с внешними системами и расширять свои возможности.

Tool calling — это механизм, благодаря которому большие языковые модели (LLM) взаимодействуют с внешними системами и инструментами, а также выполняют задачи, в которых требуется доступ к внешним ресурсам, таким как базы данных, API и другие сервисы. Tool calling обычно реализуется через интерфейс, который позволяет моделям отправлять запросы внешним системам и получать ответы в формате, который может быть обработан моделью.

К преимуществам tool calling следует отнести: 1) расширение возможностей LLM: tool calling позволяет моделям выполнять задачи, в которых требуется доступ к внешним ресурсам, таким как базы данных и API; 2) улучшение производительности: tool calling может значительно улучшить производительность моделей, так как они могут использовать внешние системы для выполнения задач, которые требуют больших вычислительных ресурсов; 3) повышение точности: tool calling позволяет моделям получать более точные ответы на запросы, так как они могут использовать внешние системы для проверки и уточнения информации.

Однако при использовании tool calling необходимо учитывать доступность, надежность и безопасность внешних систем, к которым будут происходить обращения и которые будут возвращать результаты вызова функций.

Современные исследования и разработки в области tool calling направлены в том числе на решение вопросов, связанных с интеграцией, производительностью и безопасностью. Например, в [11] авторы предлагают онлайн-оптимизированный RAG, который непрерывно адаптирует векторы (эмбеддинги) извлекаемых данных из реальных взаимодействий, используя минимальную обратную связь. В работе [12] представлен CoreThink Agentic Reasoner — фреймворк, дополняющий LLM-модели легковесным слоем символьных рассуждений для структурной декомпозиции и адаптивного управления инструментами. Авторы [13] предлагают инструменты естественного языка (NLT), которые заменяют программный вызов инструментов JSON на вывод на естественном языке. В работе [14] исследуется вызов инструментов для арабского языка и рассмотрены три ключевых вопроса: необходимость наличия данных на арабском языке, влияние настройки инструкций общего назначения на производительность вызова инструментов и ценность тонкой настройки для конкретных высокоприоритетных инструментов. Авторы [15] изучают вызов инструментов в регулируемых корпоративных средах, таких как финтех. В [16] предлагают унифицированный подход к интеграции инструментов, который абстрагирует различия протоколов и оптимизирует производительность выполнения.

Таким образом, tool calling является критически важной возможностью для больших языковых моделей, позволяющей им взаимодействовать с внешними системами и расширять свои возможности. Механизм tool calling стал активно развиваться в последние годы усилиями исследователей и компаний OpenAI, Google и других. Например, OpenAI активно использует tool calling в своих моделях, таких как GPT-4, для взаимодействия с внешними API и сервисами.

Tool calling существенно повышает объяснимость и детерминированность полученных с помощью технологий ИИ результатов, поскольку основан на вызове «классических» функций. Основная сложность — это вероятностный характер процедуры выбора языковой моделью функций, соответствующих поставленной задаче. Обычно эта сложность преодолевается за счет предварительного

создания полных стандартизированных описаний вызываемых функций и правильно подобранных контекстных инструкций (промптов). Более подробно особенности использования tool calling представлены в работах [11; 17].

Технология оперативного создания прототипов информационных систем на основе ИИ

В соответствии с поставленной во введении задачей разработана технология, используя которую можно оперативно создавать прототипы интеллектуальных ИС. В рамках данной разработки представлен концепт архитектуры интеллектуальной ИС на основе большой языковой модели, определен набор средств функционального расширения и наполнения ее возможностей, а также алгоритм их совместной интеграции в качестве основного связующего звена. Алгоритмическая блок-схема решения поставленной задачи показана на рис. 1 в виде стандартной UML-диаграммы активности.

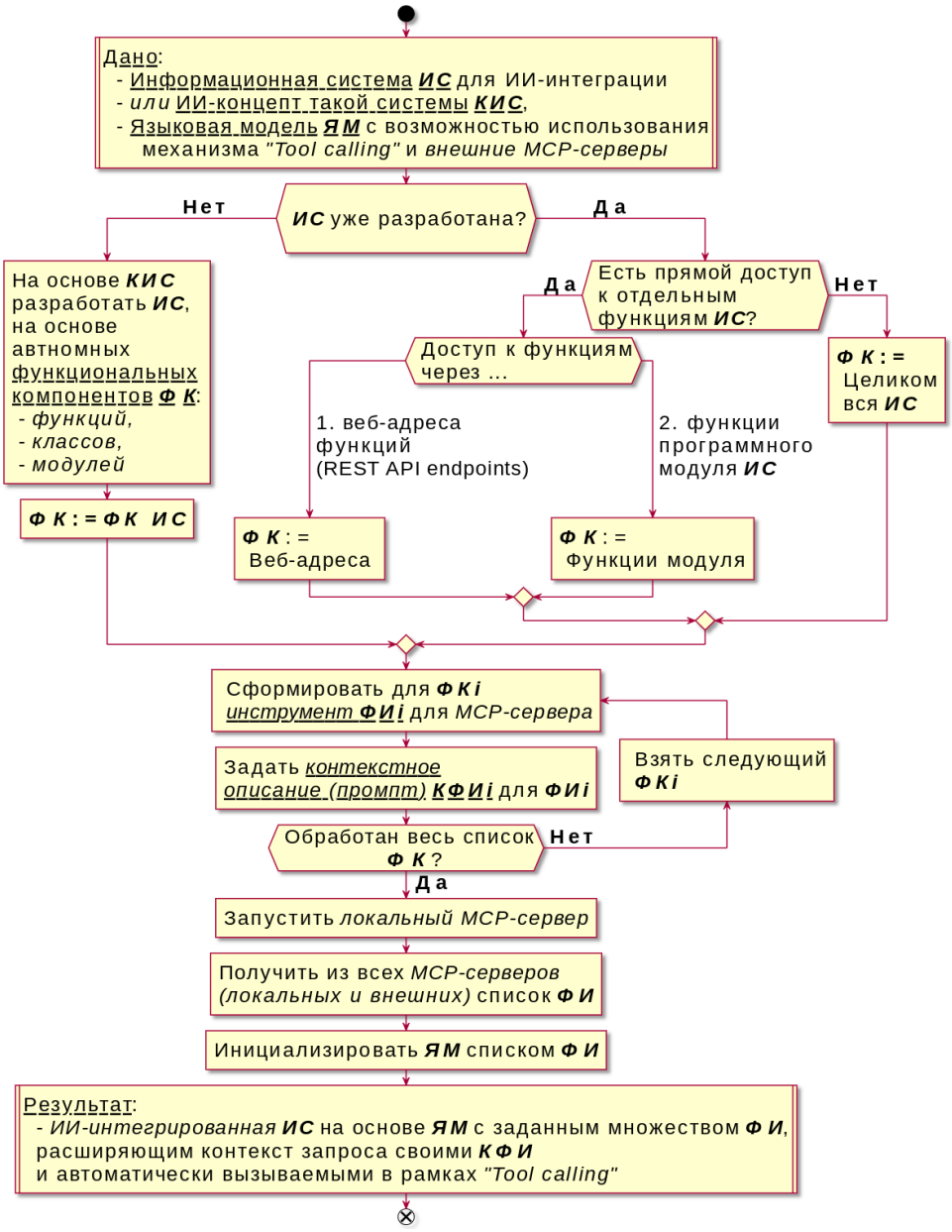


Рис. 1. Технология оперативного создания прототипов ИС на основе ИИ

Таким образом, оперативно создаваемые прототипы интеллектуальных систем архитектурно являются модульными агентами на основе больших языковых моделей, к которым по стандартному протоколу MCP подключаются внешние функциональные ресурсы — инструменты. Само интеллектуальное ядро такого агента выполняет связующую коммуникационную роль, обеспечивающую универсальный человеко-машинный интерфейс на основе естественного языка. Преимущество такого подхода в том, что большое количество типовых MCP-инструментов можно выбрать из специализированных репозиториев (например, fastmcp.me). Недостающие в работе специфические модули необходимо разработать самостоятельно. Такую разработку можно целенаправленно вести с учетом спецификации MCP. Однако, если есть готовые (унаследованные) модули, построенные без учета требований MCP, их можно интегрировать в предложенную архитектуру на уровне конфигурирования локального MCP-сервера. В данной работе учтены подобные варианты, и на схеме представлены отдельные решения как для разной степени открытости исходного кода консольных (скриптовых) модулей, так и для веб-разработок, реализованных на базе архитектуры REST API, где доступ к функциям производится по веб-адресам (маршрутам).

Практическая реализация технологии оперативного создания прототипов ИС на основе ИИ

В данном разделе представлен пример реализации технологии оперативного создания прототипов ИС на основе ИИ применительно к разработанной ранее ИС автоматизированного мониторинга чатов «ВКонтакте». Функциями данной ИС являются: 1) авторизация в ВК Мессенджер через сохраненный пользовательский профиль барузера Firefox; 2) открытие чатов по их названию через поисковую систему «ВК»; 3) отправка текстовых сообщений в активный чат; 4) создание тестовых сообщений для анализа структуры DOM; 5) прокрутка (динамическая подгрузка) истории сообщений; 6) извлечение контента сообщений (парсинг, сбор данных).

Эффективность и результативность работы основной функции системы (п. 6) обеспечивается остальными вспомогательными функциями. Такая функциональная специфика на практике отражает суть исследуемого и развиваемого коллективом авторов подхода фокусированного сбора данных.

На рисунке 2 в виде UML-диаграммы вариантов использования представлены способы реализации имеющейся ИС до и после преобразования в интеллектуальную систему. Набор функций исходной системы остался прежним, а пользовательский интерфейс реализован на основе запросов в свободной форме на естественном языке посредством создания интеллектуального агента.

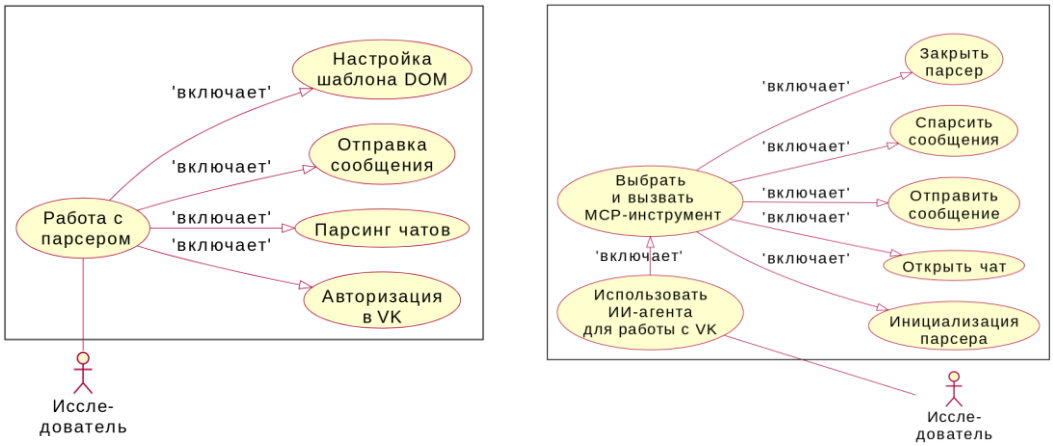


Рис. 2. Варианты действий пользователя ИС, построенной без использования ИИ (слева) и варианты действий пользователя ИС, построенной на базе ИИ (справа)

На рисунке 3 в виде UML-диаграммы компонентов представлена структура интеллектуальной ИС: ИИ-агента и MCP-сервера с реализованными в нем функциональными инструментами. Применяемая технология интеграции позволяет оставить полностью неизменными структурные элементы существующей (до преобразования) ИС. С помощью сервера создается оболочка вокруг работающих функций ИС, что делает их доступными через стандартизированный программный интерфейс.

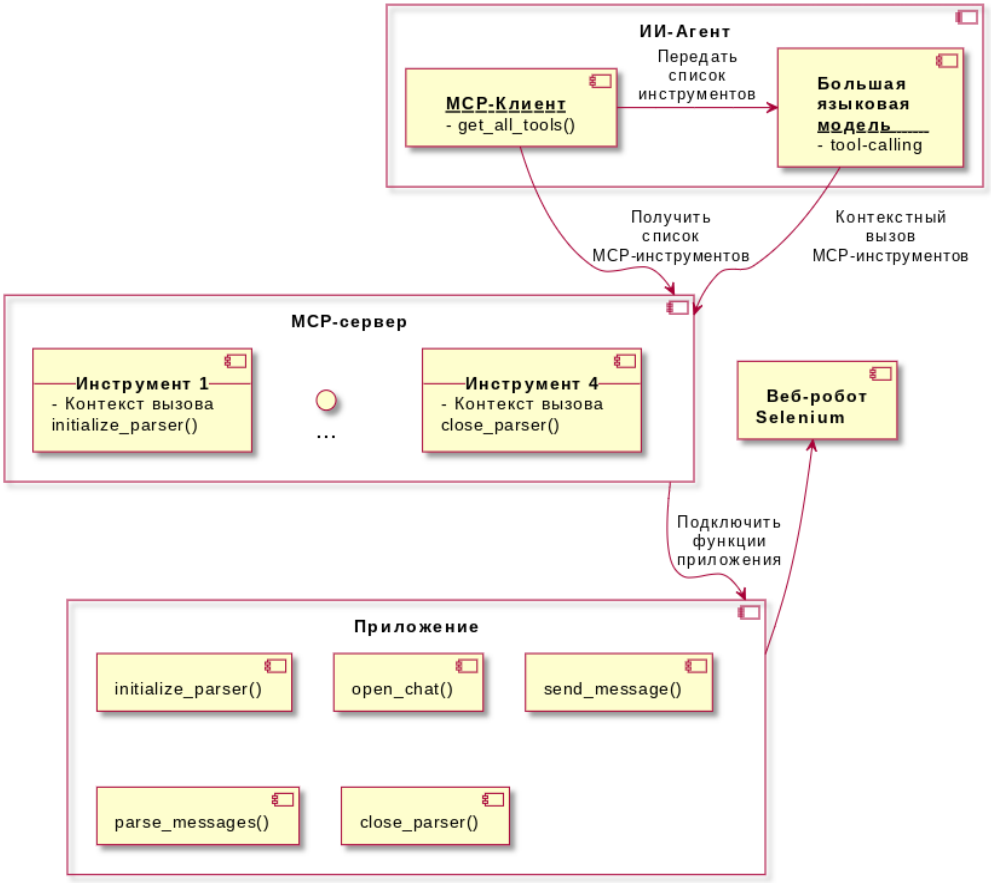


Рис. 3. Компоненты интеллектуальной ИС, построенной на основе технологий ИИ

В программной реализации MCP-сервер использует фреймворк FastMCP для создания набора инструментов. Сервер хранит экземпляр ИС (VKMessageParser) в глобальной переменной и инициализирует его при первом обращении через специальную функцию-обертку (wrapper). Создано пять инструментов-оберток, каждый из которых соответствует конкретному методу исходной ИС, в том числе инициализация сессии, открытие чатов, отправка сообщений, парсинг истории сообщений и завершение сессии. Каждый инструмент вызывает соответствующий метод ИС и возвращает полученный результат. Сервер использует порт 8000 и HTTP-протокол для обмена данными.

Таким образом, MCP-сервер предоставляет языковой модели функции для работы с мессенджером «ВКонтакте», реализованные в ИС ранее (до преобразования). При запуске интеллектуальная ИС устанавливает соединение с сервером и получает перечень доступных инструментов. Каждый инструмент соответствует определенной функции исходной ИС. Например, инструмент для открытия чатов представлен на листинге 1. Декоратор @mcp.tool указывает, что эта функция определяется как MCP-инструмент. Аналогично оформлены и другие инструменты-функции: initialize_parser(), send_message(), parse_messages(), close_parser().

```
@mcp.tool
def open_chat(chat_name: str) -> str:
    """Открыть чат в VK по его названию"""
    parser = VKMessageParser()
    success = parser.open_chat(chat_name)
return f"Чат '{chat_name}' успешно открыт" if success else f"Не удалось открыть чат
'{chat_name}'"
```

Листинг 1. Функция-инструмент для открытия чатов

Часть ИС, выступающая в роли интеллектуального агента, реализована как LangGraph-приложение [18] с архитектурой ReAct (Reasoning + Acting). В нем создается MCP-клиент с использованием интерфейса MultiServerMCPClient. Таким образом, агент запускает MCP-сервер как дочерний процесс и общается с ним через стандартные потоки ввода-вывода (stdio). Пример конфигурации представлен на листинге 2.

```
client = MultiServerMCPClient(
  {
    "vk_server": {
      "transport": "stdio",
      "command": sys.executable,
      "args": [str(Path(__file__).parent / "mcp_vk.py")],
    }
  }
)
```

Листинг 2. Инициализация клиента описанием целевого MCP-сервера

Соединение агента с большой языковой моделью происходит с помощью конфигурации, представленной на листинге 3. Языковая модель развернута на локальном сервере ИИММ. При необходимости можно подключать языковые модели, размещенные как на локальных, так и на облачных ресурсах. Для более детерминированных ответов модели в настройках подключения к модели параметр температуры (temperature) установлен в нулевое значение. Интеграция языковой модели и полученных из MCP-сервера инструментов происходит в процессе формирования специализированного реактивного объекта.

```
model = ChatOpenAI(
  model="reedmayhew/claude-3.7-sonnet-reasoning-gemma3-12B",
  base_url="http://адрес-сервера-иимм:11434/v1/",
  api_key="none",
  temperature=0,
  max_tokens=4096)
tools = await client.get_tools()
agent = create_react_agent(model, tools)
```

Листинг 3. Инициализация подключения агента к большой языковой модели

На рисунке 4 в виде UML-диаграммы последовательности представлено взаимодействие компонентов ИС в процессе отправки запроса пользователя агенту на открытие чата и парсинг сообщений. Агент, получив запрос, обращается к большой языковой модели для выбора подходящих инструментов. Языковая модель анализирует запрос и возвращает агенту информацию о необходимых инструментах. В данном случае это «open_chat» и «parse_messages». Этот вызов передается через «MultiServerMCPClient» к MCP-серверу, который выполняет соответствующую функцию, ранее представленную на листинге 1. В свою очередь, MCP-сервер вызывает ранее реализованные методы «open_chat» и «parse_messages», которые с помощью модуля веб-роботизации Selenium выполняют операции в веб-версии мессенджера «ВКонтакте».

На рисунке 5 в виде UML-диаграммы развертывания представлена ИС автоматизированного мониторинга чатов «ВКонтакте» с использованием интеллектуального агента.

ИС разворачивается в двух изолированных докер-контейнерах. Контейнер «mcp-vk-server» содержит MCP-сервер, предоставляющий инструменты парсинга, а также исходную ИС, в которой реализованы функции для работы веб-робота в браузере с помощью библиотеки Selenium.

В контейнере «mcp-client» реализован интеллектуальный агент на базе фреймворка «LangChain», который отвечает за планирование действий, обработку диалога, и в нем же располагается MCP-клиент для подключения к MCP-серверу, чтобы получить список доступных агенту инструментов.

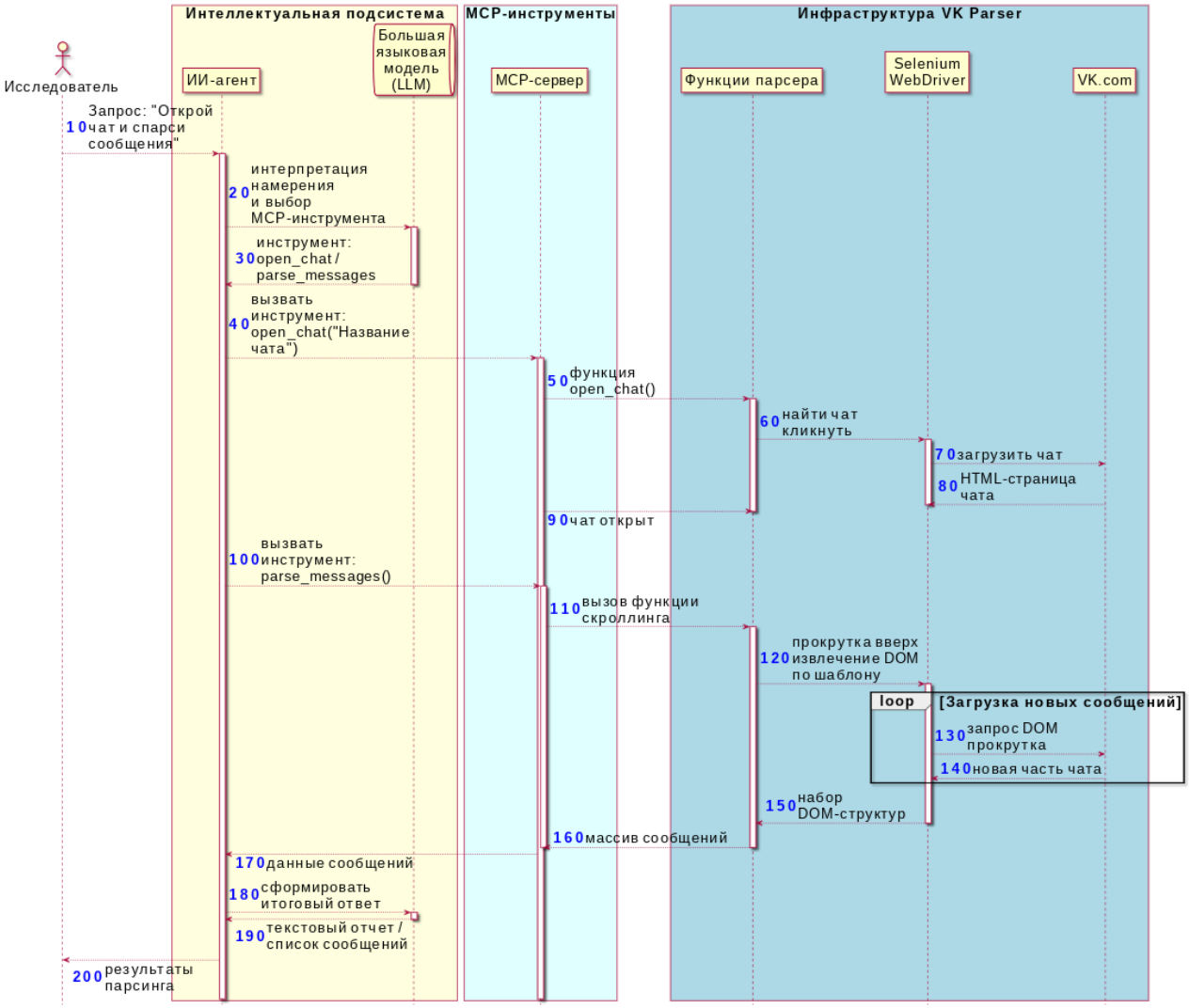


Рис. 4. Последовательность взаимодействий компонентов интеллектуальной ИС

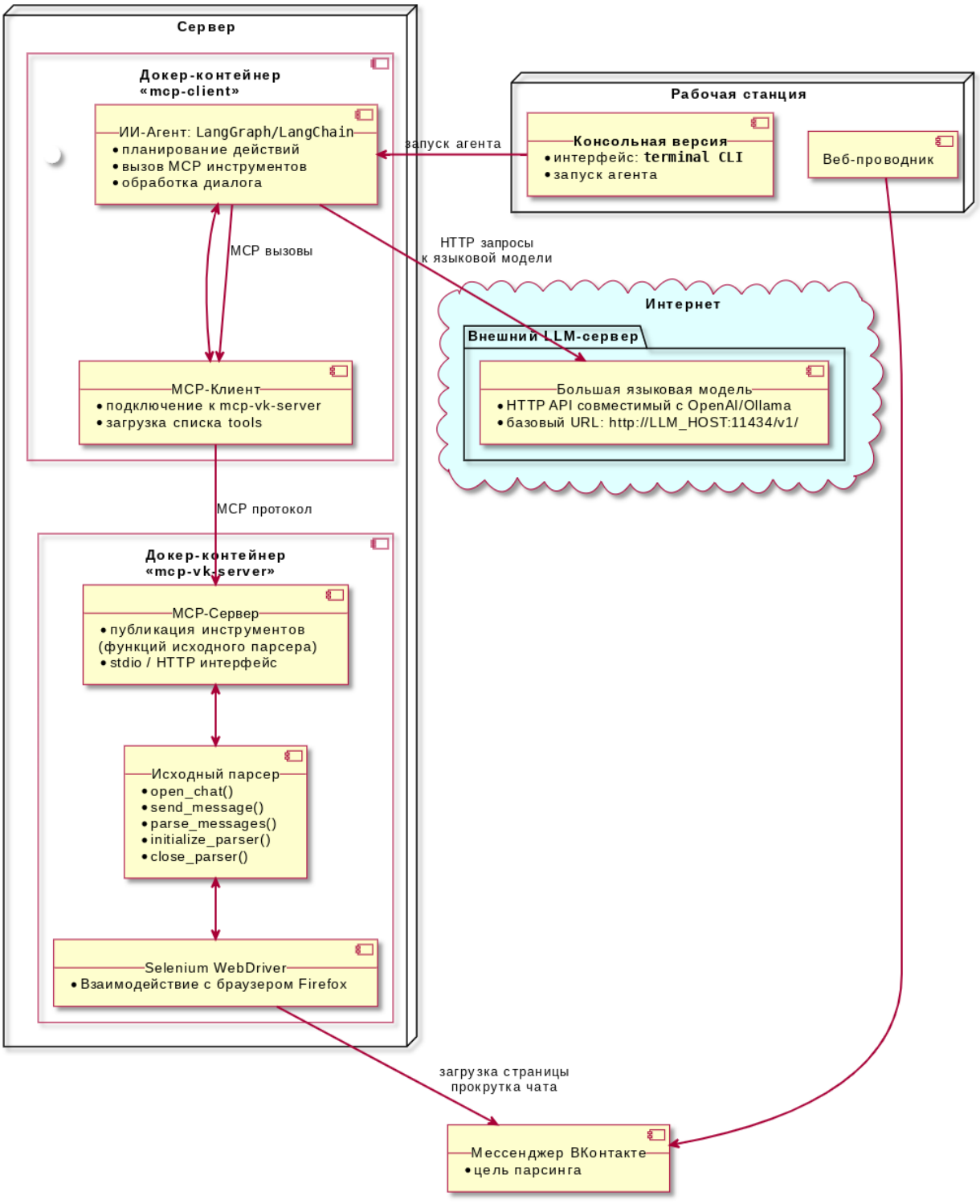


Рис. 5. Архитектура ИС автоматизированного мониторинга чатов «ВКонтакте» на основе ИИ

ИС взаимодействует с языковой моделью, развернутой на локальном сервере ИИММ, для обработки естественного языка, а также с веб-версией мессенджера «ВКонтакте». Пользователь взаимодействует с системой через консольный интерфейс на рабочей станции, формулируя запросы на естественном языке.

Обсуждение

Для оценки работоспособности и эффективности прототипа ИИ-интегрированной системы проведена серия экспериментов. Экспертно оценена способность агента понимать и правильно интерпретировать пользовательские запросы на естественном языке. В данном исследовании, в отличие от классического общераспространенного варианта использования, результаты генеративной работы языковой модели реализуют функции управления. Именно проявление этих коммуникационных возможностей ИИ стало предметом данной исследовательской работы. Максимально корректное восприятие языковой моделью пользовательских запросов обеспечивает правильный выбор и активацию подключенных к ней MCP-инструментов. С учетом отмеченной специфики, первоочередными стали вопросы обеспечения устойчивости используемых моделей к галлюцинациям и минимизация ошибочных действий.

Выбранная и развернутая на локальном сервере ИИММ языковая модель (claude-3.7-sonnet-reasoning-gemma3-12B) в целом справилась с поставленными на данном этапе тестовыми задачами и продемонстрировала исчерпывающие для этого возможности. Тестовая замена указанной версии модели на более простые варианты значительно снижала эффективность работы ИС. Однако, в силу нетривиальности выбора адекватных метрик, соответствующие формальные сравнительные оценки проведенных экспериментов авторы планируют представить в следующих своих работах.

Таким образом, по результатам проведенных работ можно отметить, что предложенная технология обладает рядом достоинств: 1) принципиальная возможность интеллектуализации практически любых проектируемых и разрабатываемых ИС; 2) доступные методы интеграции и унификации подходов к разработке прототипов информационных систем; 3) возможность оперативного внесения изменений и расширения функциональности системы — масштабируемость; 4) снижение времени на проектирование и тестирование решений; 5) использование уже имеющихся программных компонентов — повторное использование; 6) возможность предварительной оценки функционального наполнения проекта интеллектуальной ИС и его принципиальной работы без необходимости создания больших объемов программного кода.

Разработана технология, позволяющая создавать прототипы интеллектуальных информационных систем, в которых пользовательский интерфейс реализован на естественном языке и позволяет гибко и динамично формировать функциональное наполнение системы. Управление всей системой и отдельными ее компонентами производится с помощью контекстных инструкций (промптов).

В общем случае разработанные на основе представленной технологии интеллектуальные системы и их компоненты могут быть преобразованы в автономных агентов. Известны методы построения из таких элементов целых мультиагентных систем, в которых взаимодействие также строится на основе коммуникационных возможностей больших языковых моделей. В частности, к таким методам относятся протоколы A2A и ANP. Результаты по этому направлению исследований будут представлены авторами в будущих публикациях.

К недостаткам предложенного решения относятся конфабуляции (галлюцинации) и недетерминированность результатов работы больших языковых моделей. Это негативное свойство моделей переносится и на основной рабочий механизм функциональной интеграции (tool calling), на базе которого реализованы те самые коммуникационные возможности больших языковых моделей.

В данном случае в качестве возможных решений рассматриваются: 1) формирование точных описаний MCP-инструментов; 2) подбор хороших контекстных инструкций (промтов); 3) организация промежуточных контрольных проверок и метрик правильности ответов; 4) использование актуальных и максимально производительных вариантов больших языковых моделей.

В целом потенциал и возможности разработанной технологии во многом перекрывают указанные недостатки, которые на текущем этапе можно считать временными и вполне разрешимыми.

Заключение

Представленные в данной работе теоретические и прикладные технологические аспекты позволяют контролируемо интегрировать элементы технологии ИИ в разрабатываемые классическим способом информационные системы. Таким образом, продемонстрирована принципиальная возможность использования практически в любых информационных системах в качестве интерфейсно-коммуникационного компонента интеллектуальных агентов и ассистентов на основе больших языковых моделей. При желании и необходимости в рамках описанной технологии функциональная и целевая нагрузка на эти интеллектуальные компоненты может быть скорректирована. В целом способ ИИ-интеграции функций ИС на основе протокола MCP и возможность выбора нужной глубины этой интеграции позволяют решать задачу контролируемого и объяснимого использования современных интеллектуальных технологий.

Полученные результаты коллектив авторов планирует непосредственно использовать для развития своих исследований в области фокусированного сбора и интеллектуализированного извлечения информации из открытых онлайн-источников. В уже начатых и активно продолжающихся по этому направлению работах для ранее сформулированных теоретических и концептуальных аспектов разрабатываются прикладные прототипы, предназначенные для исследовательского тестирования и опробования компонентов обобщающей информационной технологии фокусированного сбора.

Список источников

1. Информационная карта НИОКТР. URL: https://gisnauka.ru/nioktr/detail/3A4F1NQ3Q3O68L885FV6A9XO (дата обращения: 20.10.2025).
2. Датьев И. О., Федоров А. М., Ревякин А. А. Фокусированный сбор и обработка открытых данных социальных медиа // Онтология проектирования. 2024. Т. 14, № 4 (54). С. 569–581. doi:10.18287/2223-9537-2024-14-4-569-581.
3. Han, Junxiao; Yu, Zheng,; Bao, Lingfeng; Liu, Jiakun; Wan, Yao; Yin, Jianwei; Deng, Shuiguang; Han, Song (2025). From LLMs to Agents: A Comparative Evaluation of LLMs and LLM-based Agents in Security Patch Detection. 10.48550/arXiv.2511.08060.
4. Li, Rui; Gu, Jia-Chen; Kung, Po-Nien; Heming, Xia; liu, Junfeng; Kong, Xiangwen; Sui, Zhifang; Peng, Nanyun (2025). LLM-REVal: Can We Trust LLM Reviewers Yet? 10.48550/arXiv.2510.12367.
5. Krishnan, Naveen Kumar (2025). Advancing Multi-Agent Systems Through Model Context Protocol: Architecture, Implementation, and Applications. Preprint. https://arxiv.org/abs/2504.21030.
6. Ayyagari, Vallikranth. Model Context Protocol for Agentic AI: Enabling Contextual Interoperability Across Systems // International Journal of Computational and Experimental Science and Engineering. 2025. Vol. 11, no. 3, Aug. doi:10.22399/ijcesen.3678.
7. Pankaj Agrawal. Model context protocol: Architectural framework for reducing AI dependency conflicts in financial services. *World Journal of Advanced Engineering Technology and Sciences*. 2025. 15 (03). P. 1916–1923. doi:10.30574/wjaets.2025.15.3.1128.
8. Zhang, Qian; Xie, Le. PowerAgent: A Road Map Toward Agentic Intelligence in Power Systems: Foundation Model, Model Context Protocol, and Workflow // IEEE Power and Energy Magazine. 2025. 23. P. 93–101. 10.1109/MPE.2025.3579718.1-8.
9. Model Context Protocol Architecture. URL: https://modelcontextprotocol.io/docs/learn/architecture (дата обращения: 20.10.2025).
10. JSON-RPC 2.0 Specification. URL: https://www.jsonrpc.org/ (дата обращения: 20.10.2025).
11. Pan, Yu; Li, Xiaocheng; Wang, Hanzhao (2025). Online-Optimized RAG for Tool Use and Function Calling. 10.48550/arXiv.2509.20415.
12. Bhat, Vishvesh; Ghugarkar, Omkar; McAuley, Julian (2025). On Generalization in Agentic Tool Calling: CoreThink Agentic Reasoner and MAVEN Dataset. 10.48550/arXiv.2510.22898.

13. Johnson, Reid; Pain, Michelle; West, Jordan (2025). Natural Language Tools: A Natural Language Approach to Tool Calling In Large Language Agents. 10.48550/arXiv.2510.14453.
14. Ersoy, Asim; Altinisik, Enes; Sencar, Husrev; Darwish, Kareem (2025). Tool Calling for Arabic LLMs: Data Strategies and Instruction Tuning. 10.48550/arXiv.2509.20957.
15. Osuagwu, Richard; Cook, Thomas; Masoud, Maraim; Ghosal, Koustav; Mattivi, Riccardo (2025). ScaleCall—Agentic Tool Calling at Scale for Fintech: Challenges, Methods, and Deployment Insights. 10.48550/arXiv.2511.00074.
16. Doh, Seungheon; Choi, Keunwoo; Nam, Juhan (2025). TalkPlay-Tools: Conversational Music Recommendation with LLM Tool Calling. 10.48550/arXiv.2510.01698.
17. Ross, Hayley; Mahabaleshwarkar, Ameya; Suhara, Yoshi (2025). When2Call: When (not) to Call Tools. 10.48550/arXiv.2504.18851.
18. Taulli, T., Deshmukh, G. Introduction to LangGraph // Building Generative AI Agents. Apress, Berkeley, CA, 2025. https://doi.org/10.1007/979-8-8688-1134-0_9.

References

1. R&D Information Card. Available at: https://gisnauka.ru/nioktr/detail/3A4F1NQ3Q3O68L885FV6A9XO (accessed 20.10.2025).
2. Datyev I. O., Fedorov A. M., Reviakin A. A. Focused collection and processing of open social media data. *Ontology of designing*, 2024, 14 (4), pp. 569–581. (In Russ.). doi:10.18287/2223-9537-2024-14-4-569-581.
3. Han, Junxiao & Yu, Zheng & Bao, Lingfeng & Liu, Jiakun & Wan, Yao & Yin, Jianwei & Deng, Shuiguang & Han, Song. (2025). From LLMs to Agents: A Comparative Evaluation of LLMs and LLM-based Agents in Security Patch Detection. 10.48550/arXiv.2511.08060.
4. Li, Rui & Gu, Jia-Chen & Kung, Po-Nien & Heming, Xia & liu, Junfeng & Kong, Xiangwen & Sui, Zhifang & Peng, Nanyun. (2025). LLM-REVal: Can We Trust LLM Reviewers Yet?. 10.48550/arXiv.2510.12367.
5. Krishnan, Naveen Kumar. (2025). Advancing Multi-Agent Systems Through Model Context Protocol: Architecture, Implementation, and Applications. Preprint. https://arxiv.org/abs/2504.21030.
6. Ayyagari, Vallikranth. Model Context Protocol for Agentic AI: Enabling Contextual Interoperability Across Systems. *International Journal of Computational and Experimental Science and Engineering*, 2025, vol. 11, no. 3, Aug. doi:10.22399/ijcesen.3678.
7. Pankaj Agrawal. Model context protocol: Architectural framework for reducing AI dependency conflicts in financial services. *World Journal of Advanced Engineering Technology and Sciences*, 2025, 15 (03), pp. 1916–1923. doi: 10.30574/wjaets.2025.15.3.1128.
8. Zhang, Qian & Xie, Le. PowerAgent: A Road Map Toward Agentic Intelligence in Power Systems: Foundation Model, Model Context Protocol, and Workflow. *IEEE Power and Energy Magazine*, 2025, 23, pp. 93–101. 10.1109/MPE.2025.3579718.1-8.
9. Model Context Protocol Architecture. Available at: https://modelcontextprotocol.io/docs/learn/architecture (accessed 20.10.2025).
10. JSON-RPC 2.0 Specification. Available at: https://www.jsonrpc.org/ (accessed 20.10.2025).
11. Pan, Yu & Li, Xiaocheng & Wang, Hanzhao. (2025). Online-Optimized RAG for Tool Use and Function Calling. 10.48550/arXiv.2509.20415.
12. Bhat, Vishvesh & Ghugarkar, Omkar & McAuley, Julian. (2025). On Generalization in Agentic Tool Calling: CoreThink Agentic Reasoner and MAVEN Dataset. 10.48550/arXiv.2510.22898.
13. Johnson, Reid & Pain, Michelle & West, Jordan. (2025). Natural Language Tools: A Natural Language Approach to Tool Calling In Large Language Agents. 10.48550/arXiv.2510.14453.
14. Ersoy, Asim & Altinisik, Enes & Sencar, Husrev & Darwish, Kareem. (2025). Tool Calling for Arabic LLMs: Data Strategies and Instruction Tuning. 10.48550/arXiv.2509.20957.
15. Osuagwu, Richard & Cook, Thomas & Masoud, Maraim & Ghosal, Koustav & Mattivi, Riccardo. (2025). ScaleCall - Agentic Tool Calling at Scale for Fintech: Challenges, Methods, and Deployment Insights. 10.48550/arXiv.2511.00074.
16. Doh, Seungheon & Choi, Keunwoo & Nam, Juhan. (2025). TalkPlay-Tools: Conversational Music Recommendation with LLM Tool Calling. 10.48550/arXiv.2510.01698.
17. Ross, Hayley & Mahabaleshwarkar, Ameya & Suhara, Yoshi. (2025). When2Call: When (not) to Call Tools. 10.48550/arXiv.2504.18851.
18. Taulli, T., Deshmukh, G. Introduction to LangGraph. In: *Building Generative AI Agents*. Apress, Berkeley, CA, 2025. https://doi.org/10.1007/979-8-8688-1134-0_9.

Информация об авторах

А. М. Федоров — кандидат технических наук, ведущий научный сотрудник;
И. О. Датьев — кандидат технических наук, старший научный сотрудник;
М. О. Илясов — программист;
И. Г. Вишняков — аспирант ФИЦ КНЦ РАН, системный администратор ИИММ КНЦ РАН;
М. О. Базегский — стажер-исследователь;
Д. С. Фигуркин — стажер-исследователь;
К. Д. Любимова — студент.

Information about the authors

A. M. Fedorov — Candidate of Science (Tech.), Leading Research Officer;
I. O. Datyev — Candidate of Science (Tech.), Senior Research Officer;
M. O. Ilyasov — Programmer;
I. G. Vishnyakov — Postgraduate Student of the Federal Research Center of the Kola Science Center of the Russian Academy of Sciences, System Administrator of the IIMM KSC RAS;
M. O. Bazegskiy — Research Intern;
D. S. Figurkin — Research Intern;
C. D. Lyubimova — Student.

Статья поступила в редакцию 22.10.2025; одобрена после рецензирования 27.11.2025; принята к публикации 02.12.2025.
The article was submitted 22.10.2025; approved after reviewing 27.11.2025; accepted for publication 02.12.2025.

Научная статья
УДК 004.9
doi:10.37614/2949-1215.2025.16.3.002

ПОДХОД К ФОРМИРОВАНИЮ ОНТОЛОГИИ ЦИКЛА УПРАВЛЕНИЯ ЖИЗНЕСПОСОБНОСТЬЮ КРИТИЧЕСКИХ ИНФРАСТРУКТУР

Павел Андреевич Ломов[1], Андрей Владимирович Маслобоев[2], Андрей Григорьевич Олейник[3]✉
[1, 2, 3]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[1]p.lomov@ksc.ru, https://orcid.org/0000-0002-0924-0188
[2]a.masloboev@ksc.ru, https://orcid.org/0000-0002-1231-9225
[3]a.oleynik@ksc.ru✉, https://orcid.org/0000-0002-7612-5999

Аннотация

В работе представлен вариант решения задачи формирования онтологии поддержки цикла управления жизнеспособностью критических инфраструктур. Онтология формируется на основе разработанной формализованной концептуальной модели процесса управления. В концептуальной модели представляются множества информационных сущностей и отношений, задействованных в различных фазах жизненного цикла процесса управления, а также необходимые атрибуты, характеризующие сущности и отношения. Концептуальное описание позволяет сформировать соответствующую систему онтологических паттернов содержания. Использование этих паттернов существенно упрощает создание прикладных онтологий для конкретных ситуаций, возникающих при функционировании критических инфраструктур. Для генерации паттернов реализуется двухэтапная процедура с использованием больших языковых моделей.

Ключевые слова:

критическая инфраструктура, жизнеспособность, цикл управление, онтология, онтологические паттерны содержания, большая языковая модель

Благодарности:

работа выполнена в рамках научно-исследовательской работы «Методы и информационные технологии мониторинга и управления региональными критическими инфраструктурами Арктической зоны Российской Федерации» (проект № FMEZ-2025-0054).

Для цитирования:

Ломов П. А., Маслобоев А. В., Олейник А. Г. Подход к формированию онтологии цикла управления жизнеспособностью критических инфраструктур // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 22–34. doi:10.37614/2949-1215.2025.16.3.002.

Original article

AN APPROACH TO ONTOLOGY DESIGN OF THE CRITICAL INFRASTRUCTURE RESILIENCE MANAGEMENT CYCLE

Pavel A. Lomov[1], Andrey V. Masloboev[2], Andrey G. Oleynik[3]✉
[1, 2, 3]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[1]p.lomov@ksc.ru, https://orcid.org/0000-0002-0924-0188
[2]a.masloboev@ksc.ru, https://orcid.org/0000-0002-1231-9225
[3]a.oleynik@ksc.ru✉, https://orcid.org/0000-0002-7612-5999

Abstract

A solution for the task of ontology forming for supporting the critical infrastructure resilience management cycle is presented in the article. The ontology is formed on the basis of the developed formalized conceptual model of the management process. The conceptual model represents sets of information entities and relationships involved in various phases of the life cycle of the management process, as well as the necessary attributes characterizing the entities and relationships. The conceptual description allows to form a corresponding system of Content Ontology Design Patterns. The use of these patterns significantly simplifies the creation of applied ontologies for specific situations arising during the operation of critical infrastructures. A two-step procedure using large language models is implemented for generate patterns.

Keywords:

critical infrastructure, resilience, management cycle, ontology, content ontology design patterns, large language models

Acknowledgments:
The study was carried out within the framework of the research project “Methods and information technologies for monitoring and managing regional critical infrastructures of the Arctic zone of the Russian Federation” (project No. FMEZ-2025-0054).
For citation:
Lomov P. A., Masloboev A. V., Oleynik A. G. An approach to ontology design of the critical infrastructure resilience management cycle. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 22–34. doi:10.37614/2949-1215.2025.16.3.002.

Введение

В состав Российской Федерации входят довольно обширные территории, которые характеризуются неблагоприятными климатическими условиями, высокими экологическими рисками, труднодоступностью и, как следствие, ограниченным развитием инфраструктурной обеспеченности. Поэтому на этих территориях особого внимания требуют надежность, безопасность и эффективность функционирования существующей инфраструктуры обеспечения различных видов деятельности, а также рациональное планирование новых инфраструктурных проектов. Обладающие определенной спецификой задачи обеспечения устойчивого функционирования инфраструктурных объектов не менее важны и для густонаселенных территорий с развитой промышленностью. Использование современных методов и информационных технологий мониторинга и управления инфраструктурами различного типа позволяет повысить их устойчивость к различным угрозам, быстро реагировать на аварийные ситуации и оптимизировать процессы управления.

Проводимые исследования направлены на разработку методов и средств информационной поддержки управления жизнеспособностью критических инфраструктур (КИ). КИ определяется как многокомпонентная распределенная система, состоящая из множества взаимозависимых подсистем, нарушение работоспособности хотя бы одной из которых может привести к существенному ухудшению безопасности жизнедеятельности населения (в широком смысле) и оказать значительное влияние на функционирование других подсистем и на жизнеспособность КИ в целом [1]. Под жизнеспособностью понимается защитное свойство системы [2], подвергающейся негативным воздействиям, позволяющее ей сопротивляться, поглощать, приспосабливаться и восстанавливаться после последствий воздействия, в том числе путем сохранения и восстановления ее основных структур и функций.

Управление жизнеспособностью КИ представляет собой многоэтапный и многофункциональный процесс, в общем случае включающий в себя такие функции управления, как целеполагание, стратегическое планирование, оперативное реагирование, а также функции контроля, учета, мониторинга, прогнозирования и координации. На каждом этапе этого процесса для реализации соответствующих функций управления необходимы ресурсы, силы и средства, адекватные оперативному контексту ситуации и решаемым задачам ситуационного управления.

Базовый цикл управления жизнеспособностью КИ включает последовательность взаимосвязанных фаз ситуационного управления устойчивым функционированием системы в условиях возникновения, развития и воздействия деструктивных событий различной природы и масштаба. Как правило, такой цикл включает в себя фазу идентификации, восприятия и предупреждения потенциальных угроз и опасностей, соответствующую режиму готовности (режим нормального функционирования системы), фазу поглощения негативного воздействия угроз и реагирования на возмущения и фазу восстановления работоспособности компонентов системы после сбоя. После фазы восстановления наступает фаза адаптации к новым условиям функционирования, в рамках которой система трансформируется и накапливает опыт противодействия угрозам, порожденным деструктивным воздействием на систему, тем самым повышая свою устойчивость (жизнеспособность) к подобного рода событиям в будущем.

Формализация задачи информационной поддержки управления жизнеспособностью

Для реализации средств информационно-аналитической поддержки управления жизнеспособностью КИ разрабатываются варианты формальных моделей, учитывающие специфику различных задач управления. Систематизация и примеры таких моделей подробно рассматриваются в работе [2].

В настоящей статье разработка варианта формального представления модели жизненного цикла управления жизнеспособностью системы опиралась на опыт использования технологии концептуального моделирования, основанной на функционально-целевом подходе [3], наработок в сфере ситуационного управления [4], онтологического моделирования [5; 6], методов анализа рисков нарушения безопасности КИ [7], кризисного управления [8], модельного инструментария и средств обеспечения жизнеспособности сложных систем [9].

В первую очередь были определены основные объекты (сущности), которые должны быть представлены в модели. К сущностям верхнего уровня относятся: объект управления, субъект управления и воздействия, несущие угрозы функционированию объекта управления. Каждая из этих сущностей имеет некоторую структуру и обладает определенными характеристиками. В рамках задач поддержки управленческих решений не всегда есть необходимость полностью рассматривать все элементы структуры и характеристики этих сущностей. Целесообразность учета компонентов структуры и характеристик, как правило, будет зависеть от конкретной ситуации, имеющей место на объекте управления. Ситуация, в свою очередь, определяется состоянием объекта управления и влиянием на его состояние некоторого деструктивного воздействия, нарушающего нормальное функционирование объекта управления. Возникновение нарушения порождает проблему. Устранение проблемы является целью субъекта управления и должно обеспечиваться им путем решения комплекса задач. Для решения задач субъект управления реализует определенные функции и использует необходимые ресурсы. В общем случае задачи и функции их решения распределяются между структурными компонентами субъекта управления, на основе которых формируется организационная структура управления решением задач устранения проблемы. Таким образом, организационная структура включает множество агентов управления, реализующих необходимые функции. В результате анализа наличия ресурсов и времени, необходимого для реализации каждой функции, формируется функциональная структура устранения проблемы. По аналогии с концептуальными моделями [3; 10] последовательность выполнения функций (отношения следования функций) может определяться двумя способами:

1) явно — задаются директивно, когда на основе нормативных документов или собственных предпочтений субъект управления явно указывает, что некоторая функция управления cf_k может выполняться только после завершения выполнения функции cf_i;

2) неявно — определяются по потоку ресурсов — функция управления cf_k должна выполняться после завершения выполнения функции cf_i, если необходимые для выполнения cf_k ресурсы генерируются в результате выполнения функции cf_i.

Специальные процедуры автоматизированного анализа отношений следования на функциях, включенных в модель организационной структуры, позволяют проверить непротиворечивость явных и неявных отношений следования и генерировать функциональную модель управления. При определении порядка следования функций также учитывается время, необходимое на их выполнение.

В работе [11] впервые предложена формализованная структура модели субъекта управления, основанная на применении процессного подхода к анализу и моделированию ситуационного управления жизнеспособностью КИ и отличающаяся полнотой формального представления задач и функций обеспечения жизнеспособности и связанных с этими задачами процессов ситуационного управления.

На основании приведенных выше рассуждений в модели системы управления жизнеспособностью представлены следующие компоненты и множества:

$CF^i, i = \overline{1,n}$ — множество фаз жизненного цикла управления жизнеспособностью (n — количество фаз цикла);

$\{cf_x\}$ — множество функций управления, элементы которого соотносятся с определенными фазами, т. е. каждой фазе соответствует некоторое подмножество из m функций обеспечения жизнеспособности cf_{ij}, $i = \overline{1,n}$, j=$\overline{1,m}$. Каждая функция предполагает использование определенных ресурсов r_{ij} и времени t_{ij}, необходимых для ее реализации; m — количество функций, реализуемых в рамках i-й фазы.

cf_{i0} и cf_{ie} — начальная и конечная функции цикла, обеспечивающие подготовку входных ресурсов (например, «Мониторинг» или «Восприятие рисков») и получение результатов реализации цикла управления («Обучение» или «Оценка нового состояния») соответственно;

MA = $\{ma_x\}$ — множество агентов управления жизнеспособностью, выполняющих определенные функции в рамках этапов жизненного цикла жизнеспособности;

R = {r_q}, $q = \overline{1,l}$ — множество разнотипных ресурсов, необходимых для реализации функций управления cf_{ij} на этапах жизненного цикла управления жизнеспособностью. Для каждого ресурса в качестве обязательных атрибутов указывается его объем (*Vol*), тип (*Typ*) и стоимость (*C*).

На множестве MA могут быть заданы отношения подчиненности, обеспечивающие отражение иерархической структуры субъекта управления. Данные отношения могут быть полезны при разработке и использовании автоматизированных процедур координации действий агентов управления различных уровней в процессе решении задач, обусловленных конкретной ситуацией. Подобные процедуры представлены, например, в [12].

Каждому элементу множества агентов управления MA должно быть сопоставлено некоторое, реализуемое им, непустое подмножество функций управления $\{cf_{ij}\}_{ma_i}$. Следует отметить, что в общем случае одна и та же функция может реализовываться разными агентами управления, т. е. возможны ситуации, когда $\{cf_{ij}\}_{ma_i} \cap \{cf_{kx}\}_{ma_k} \neq \emptyset$.

Выше уже было указано, что между функциями, реализуемыми агентами управления, существуют отношения следования, которые могут определяться двумя способами - директивно, либо по потоку ресурсов. Явно в модели целесообразно представлять только отношения следования, задаваемые директивно: CR = $\{cr_{kj}\}$ = {<cf_{ik}, cf_{ij}>}, где кортеж <cf_{ik}, cf_{ij}> соответствует утверждению, что функция cf_{ik} может выполняться только после завершения выполнения функции cf_{ij}. При этом для корректного формирования модели управления должна быть предусмотрена процедура, проверяющая непротиворечивость отношений следования функций, определенных различными способами.

Между ресурсами и функциями определяются отношения использования и порождения ресурсов. Ресурсы, необходимые для выполнения функции (используемые функцией), относятся ко входным ресурсам данной функции и определяются в модели отношением $in(r_q,cf_{ij})$. Ресурсы, генерируемые в результате выполнения функции, относятся к выходным: $out(cf_{ij},r_q)$. Аналогичным образом определяются входные и выходные ресурсы каждой фазы процесса управления. $IN^R(CF^i)$ — входные ресурсы, существующие (или доступные) к моменту начала фазы и необходимые для реализации выполняемых в рамках фазы функций. $OUT^R(CF^i)$ — выходные ресурсы, доступные по завершении фазы. Также для фазы имеет смысл рассматривать внутренние ресурсы $DOM^R(CF^i)$, которые не входят в $IN^R(CF^i)$, а генерируются функциями фазы. В зависимости от типов ресурсов указанные множества могут иметь непустое пересечение. Например, поступающие на вход фазы информационные ресурсы могут использоваться входящими в нее функциями, но передаваться на выход без изменений, а материальные ресурсы в процессе реализации фазы обязательно претерпевают изменения.

В связи с тем, что объект управления (критическая инфраструктура) обладает структурной и функциональной сложностью, попытки построения ее полной модели представляются малопродуктивными. Однако формирование адекватной модели фрагмента КИ, рассмотрение которого будет достаточно для поиска решения проблем, возникающих в конкретной ситуации, вполне возможно. Например, в работе [13] предложена информационная структура для информационной системы категорирования критически важных объектов (КВО), предусматривающая представление структуры и функций КВО, возможных угроз (деструктивных воздействий) и оценок последствий реализации угроз на компоненты и функции КВО. Аналогично в разрабатываемой модели должна быть реализована возможность представления фрагмента КИ, подвергающегося определенному воздействию в конкретной ситуации. Рассматриваемый фрагмент КИ может быть описан аналогично тому, как это реализовалось в системах концептуального моделирования [3; 10]. Для этого в модель включаются следующие множества:

$SC = \{sc_x\}$ — множество принимаемых в рассмотрение структурных компонентов КИ;

$F^{sc} = \{f_x^{sc}\}$ — множество принимаемых в рассмотрение функций и/или процессов, реализуемых элементами множества *SC*;

$R^{Fsc} = \{r_x^{Fsc}\}$ — множества ресурсов, используемых и производимых элементами множества F^{sc}.

На множестве *SC* задаются отношения иерархии, определяющие структурную организацию принимаемых в рассмотрение компонентов КИ. Использование отношений иерархии типа «И» позволяет представить компонент КИ как композицию его составных частей. Использование типа иерархии «ИЛИ» дает возможность представить альтернативные варианты составных частей компонента.

Подобно описанным выше отношениям, задаваемым на множествах системы управления жизнеспособностью, задаются отношения и в модели фрагмента КИ. С каждым элементом множества *SC* связывается непустое подмножество множества F^{sc}, что определяет, какие функции реализует конкретный элемент множества *SC*. Равносильным является установка связей от элементов F^{sc} к элементам *SC*, показывающих, какими компонентами КИ может реализовываться конкретная функция. Также устанавливаются отношения «вход/выход» между элементами множеств F^{sc} и R^{Fsc} и отношения следования на множестве F^{sc}, определяющие порядок их выполнения. Отношения на модели фрагмента КИ позволяют более детально определить угрозы или результаты деструктивного воздействия на КИ и идентифицировать порожденную проблему и комплекс задач, обеспечивающих ее устранение.

Проблемам классификации, моделирования, оценки рисков и предупреждения чрезвычайных ситуаций посвящено довольно много научных работ. В частности, издательством «МГФ Знание» выпускается серия книг «Безопасность России», в которых систематизируются результаты широкого спектра исследований и разработок в области анализа проблем риска и безопасности. Существует обширная нормативная база, регулирующая деятельность по предотвращению опасностей и рисков, ликвидации или снижению негативных последствий от чрезвычайных ситуаций различной природы. На основе имеющейся информации можно сформировать модель деструктивного воздействия на КИ, необходимую и достаточную для выявления порождаемой им в конкретной ситуации проблемы и определения комплекса задач по ее устранению.

Структура представления в модели деструктивных воздействий на КИ включает множества идентификаторов событий воздействия DA = $\{da_i\}$, типов воздействий TA = $\{ta_j\}$ и угроз/рисков, порождаемых воздействиями, TR = $\{tr_k\}$. Для каждого принимаемого в рассмотрение деструктивного воздействия da_i, кроме собственно идентификатора (наименования), задается время и место события. Путем установления связи da_i с элементами множества TA определяются типы порождаемых этим событием воздействий. На основе нормативных документов и других литературных источников (например, [14]) в системе информационно-аналитической поддержки управления может быть создан «справочник» типов воздействий, а связь ta_j с элементами множества TR определяет, какие риски возникают вследствие данного воздействия. Для связи между воздействием и риском ($<ta_j, tr_k>$) задаются атрибуты, определяющие величину и вероятность риска в конкретной рассматриваемой ситуации.

Формирование модели для поддержки управления жизнеспособностью КИ в конкретной ситуации осуществляется путем задания соответствующих экземпляров элементов описанных выше множеств и текущих значений атрибутов этих элементов. Затем устанавливаются связи, определяющие воздействие элементов, характеризующих деструктивное событие, на элементы, характеризующие состав и функции компонентов КИ. Для этих связей могут быть заданы атрибуты, определяющие степень и вероятность воздействия. На основе анализа влияния текущего события на компоненты КИ определяется проблема и комплекс задач для ее устранения. Анализ может проводится как экспертным путем, так и с использованием автоматизированных процедур, основанных на различных моделях оценки рисков. Затем для требующих решения задач выбираются «покрывающие» [10] функции управления cf_{ij} и акторы управления, которые могут реализовать эти функции ma_i. При необходимости на выбранном подмножестве функций задаются директивные отношения следования. Таким образом будут определены все элементы организационной структуры для решения возникшей в конкретной ситуации проблемы. Следует отметить, что при установлении связей между элементами описаний деструктивного воздействия, компонентов КИ и системы управления состав принимаемых к рассмотрению экземпляров элементов может изменяться. Это обусловлено как выявлением наличия неучтенных изначально элементов, так и определением избыточности начальных описаний.

Построение онтологической модели цикла управления жизнеспособностью

Для компьютерной реализации концептуальных моделей предметных областей и решаемых задач в последние десятилетия широко используются формальные онтологии [6; 15]. Созданы и развиваются программные системы построения и обработки онтологий [16; 17]. Однако построение концептуальной модели сложной системы или задачи «вручную» даже с использованием специальных редакторов онтологий — весьма трудоемкая процедура.

Существенно упрощает процесс построения онтологии использование онтологических паттернов, которые представляют собой детально описанные и проверенные на практике решения регулярно возникающих проблем онтологического моделирования [18; 19]. Они позволяют повторно использовать удачные проектные решения и обеспечивают единообразие структуры и семантики создаваемых онтологий. В зависимости от решаемых задач различают несколько типов онтологических паттернов [20].

В настоящей работе, как и в [21], используются онтологические паттерны содержания (Content Ontology Design Patterns, CDP), так как они непосредственно определяют структуру понятийной системы разрабатываемой онтологии. Как было отмечено ранее, они представляют собой фрагменты онтологий, предназначенные для повторного использования при моделировании распространенных ситуаций, возникающих между объектами, процессами и явлениями в различных предметных областях, таких как, например, участие объектов в событиях, отношения часть — целое, последовательности действий, причинно-следственные зависимости и др.

Фактически паттерны содержания можно рассматривать как структурированные фрагменты онтологий верхнего уровня. Они содержат общие классы (Агент, Последовательность), свойства (идентификатор, имя) и отношения (зависит от, имеет часть), которые могут быть конкретизированы. После такой специализации содержащихся в них элементов и добавления новых элементов CDP становятся неотъемлемыми составными частями прикладных онтологий.

Использование CDP в процессе проектирования онтологий способствует унификации их моделей и согласованности. Каждый паттерн содержания отражает определенную точку зрения на понятие (conceptual viewpoint), которая должна коррелировать с задачами моделирования конкретной предметной области или задачи. Для выбора наиболее подходящего паттерна применяются так называемые компетентностные вопросы (Competency Questions, CQ) [20]. Эти вопросы определяют, какие сведения можно получить из онтологии, если в ее структуре используется данный паттерн. Таким образом, набор CQ служит инструментом для проверки релевантности CDP целям моделирования и помогает разработчику выбрать паттерн, представляющий подходящую схему описания знаний.

Сами паттерны содержания целесообразно генерировать с использованием больших языковых моделей (Large Language Models, LLM). В настоящее время существуют работы, посвященные применению языковых моделей [22; 23] в онтологическом моделировании, однако они преимущественно ориентированы на создание и/или наполнение полноценных онтологий. Такой подход требует значительных вычислительных и когнитивных ресурсов, а также сложной верификации.

В отличие от этого, генерация онтологических паттернов содержания представляет собой намного более простую задачу, так как паттерн концептуально проще полной онтологии. Он описывает решение одной конкретной задачи моделирования и, как правило, включает ограниченное число сущностей и отношений. Благодаря этому при генерации описание требуемого паттерна можно задать в виде краткой инструкции (промпта) нескольких предложений, содержащих спецификацию типовой ситуации, которую необходимо отразить в онтологии.

Для повышения контроля над процессом генерации, а также снижения сложности инструкций для языковой модели и тем самым уменьшения вероятности получения синтаксических ошибок и/или так называемых галлюцинаций имеет смысл проводить генерацию в два этапа.

На первом этапе выполняется формирование набора функций на выбранном языке программирования, которые принимают на вход простые структуры данных — массивы (списки), ассоциативные массивы (словари), вложенные структуры, которые содержат естественно языковые имена понятий, отношений и атрибутов — и текущую онтологию. Каждая функция реализует определенный тип модификации онтологической модели (например, добавление класса, отношения, атрибута, ограничения). На выходе функции возвращают модифицированную онтологию, включающую новые элементы, описанные во входных данных. Таким образом, язык программирования выступает в роли прослойки между LLM и онтологическим форматом, обеспечивая контроль корректности синтаксиса и структуры создаваемых фрагментов.

В качестве примера рассмотрим промпт для генерации функции, добавляющей список понятий в онтологию:

Напиши на Python функцию для наполнения онтологии OWL с использованием библиотеки owlready2. Функция должна принимать следующие аргументы:

** объект онтологии (Ontology) из библиотеки owlready2;*

** словарь, в котором ключ — имя понятия, а значение — его определение.*

Функция должна добавлять понятия в онтологию как классы, а их определения — как аннотации rdfs:comment.

Добавь подробный комментарий (docstring), описывающий назначение функции и её аргументы.

Результатом генерации с помощью языковой модели GigaChat будет следующая функция:

```
from owlready2 import *

def populate_ontology_with_classes_and_comments(onto: Ontology, concepts_dict: dict):
    """
    Функция наполняет онтологию новыми классами и добавляет аннотации к ним.

    Параметры:
    onto (owlready2.Ontology): Объект онтологии, куда будут добавляться новые классы.
    concepts_dict (dict): Словарь, где ключи - имена понятий (названия классов),
                          а значения - их описания (аннотации).

    Действия:
    1. Для каждого элемента словаря создается новый класс с именем ключа.
    2. Классу добавляется аннотация 'rdfs:comment' с текстом, указанным в значении словаря.
    """
    for concept_name, definition in concepts_dict.items():
        # Создаем новый класс в онтологии с заданным именем
        new_class = types.new_class(concept_name, (Thing,), ontology=onto)

        # Добавляем аннотацию rdfs:comment с описанием класса
        new_class.comment.append(definition)
```

На втором этапе формируется инструкция для языковой модели, включающая запрос на генерацию набора элементов будущего паттерна — классов, свойств, отношений и ограничений — представленных в виде структур данных, принимаемых функциями, созданными на первом этапе. В этом случае LLM не формирует онтологию напрямую, а лишь порождает описание паттерна в виде структурированных данных, которые используются для получения окончательного результата с помощью функций на языке программирования.

Например, с помощью такого промпта можно извлечь из текста — описания паттерна понятия, которые далее можно поместить в его онтологию с помощью ранее приведенной функции:

Из представленного текста необходимо извлечь информацию для построения OWL-онтологии.

Построй структуру онтологии, соответствующую содержанию текста, и представь результат в виде трёх Python-структур:

concepts — словарь, где — имя понятия, значение — определение понятия.

object_properties — список кортежей, где каждый кортеж имеет вид: (subject_concept, property_name, range_concept, описание_свойства) и отражает отношения между понятиями (object properties).

datatype_properties — список кортежей, где каждый кортеж имеет вид: (subject_concept, property_name, тип значения, описание_свойства) и отражает атрибуты (data properties) понятий.

При указании типов значений используй корректные OWL/XSD-обозначения:

xsd:string, xsd:integer, xsd:float, xsd:boolean, xsd:dateTime.

В результате будет получен список понятий в виде python-словаря:

concepts = {

"ResilienceCycle": "Формальная модель управления устойчивостью системы, включающая взаимосвязанные фазы, функции, ресурсы и агентов.",

"Phase": "Фаза жизненного цикла устойчивости: anticipation, absorption или adaptation.",

"ControlFunction": "Управляющая функция, выполняемая в определённой фазе цикла для поддержания устойчивости.", … }

Такой подход позволяет четко разделить фазу концептуализации знаний (осуществляемую LLM и разработчиком онтологии) и фазу формальной спецификации (выполняемую средствами программного представления и сериализации онтологий). Это обеспечивает: 1) синтаксическую корректность результата, поскольку ответственность за генерацию формальных структур возлагается на проверенные программные библиотеки для работы с онтологическими форматами (например, OWL API, RDFLib, Owlready2, rdflib.js); 2) снижение вероятности семантических и логических ошибок, так как языковая модель оперирует ограниченным числом понятий и отношений, представленных в небольшом текстовом описании паттерна; 3) получение разработчиком возможности легко проверять и уточнять результат, поскольку результат концептуализации, сгенерированный языковой моделью, будет представлен в виде относительно простых структур языка программирования.

Разработку онтологии с использованием приведенного подхода удобнее осуществлять в так называемых электронных блокнотах (Computational Notebooks), которые можно запускать в программных средах Google Colab, Jupyter Notebook и Apache Zeppelin. Такие блокноты позволяют комбинировать форматированные текстовые заметки с блоками программного кода, а также легко запускать последние в необходимом разработчику порядке.

С помощью такого подхода на основе соответствующих фрагментов приведенной выше формализации могут быть построены паттерны, описывающие ключевые аспекты системы управления жизнеспособностью. Для примера ниже приведены схемы трех сгенерированных паттернов (рис. 1–3).

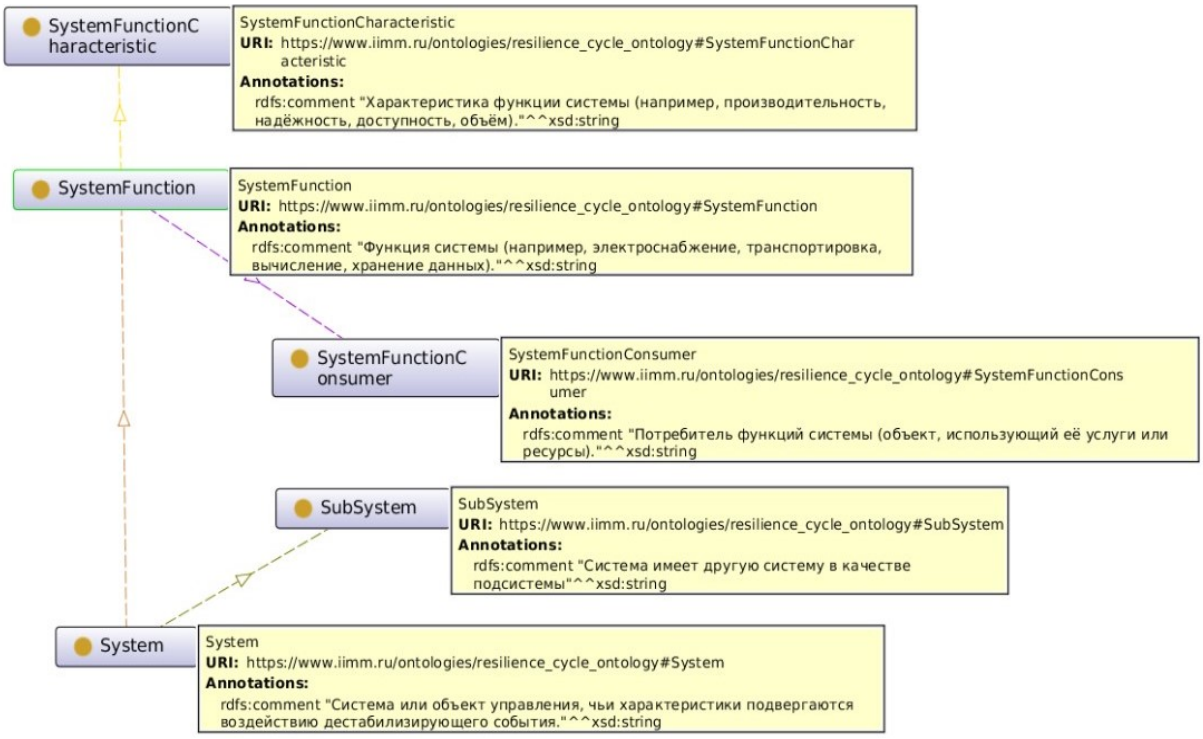


Рис. 1. Онтологический паттерн «Объект управления»

Паттерн «Объект управления» (см. рис. 1) определяет структуру онтологического описания объекта (системы), относительно которого осуществляется решение задач обеспечения жизнеспособности. Наряду с функциями, или «услугами», которые реализует данный объект, и их характеристиками, в паттерн включаются «потребители» услуг (класс *"SystemFunctionConsumer"*). Это дает возможность при использовании данного паттерна для формирования онтологической модели конкретной ситуации достаточно полно представить «сферу влияния» [13] данного объекта на другие объекты и системы, зависящие от его функционирования, что, в свою очередь, способствует более адекватной оценке критичности как самого объекта управления, так и сложившейся в результате деструктивного воздействия ситуации.

Паттерн описания деструктивного события (см. рис. 2) предусматривает указание места его возникновения, или проявления (класс *"EventPosition"*), а также то, как данное событие проявляется (класс *"EventManifestation"*). Отметим, что в общем случае одно событие может иметь несколько проявлений, оказывающих разное влияние на функциональность объекта управления. Для каждого проявления задается сила его проявления в данном событии.

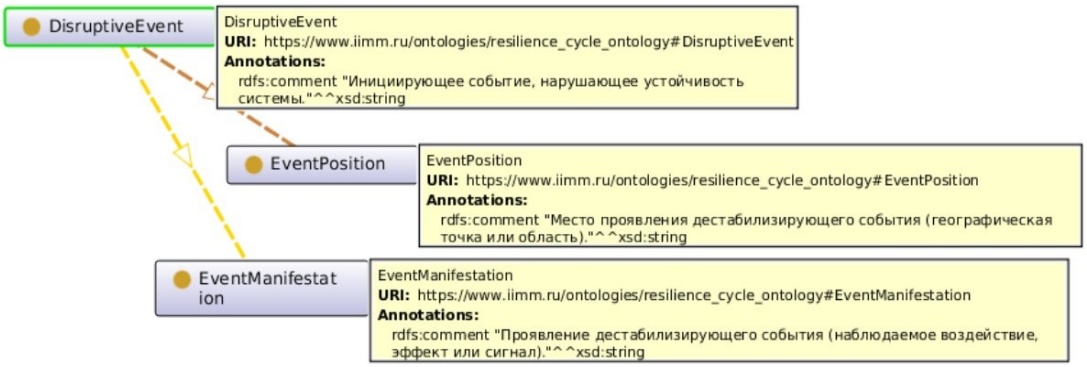


Рис. 2. Онтологический паттерн «Деструктивное событие»

Паттерн «Факт воздействия» (см. рис. 3) позволяет сформировать онтологическое описание воздействия конкретного деструктивного события на объект управления путем указания, какие проявления события, в какой степени влияют на определенные функции объекта. При этом существует возможность указать значение силы влияния, которое может быть отличным от данной характеристики проявления в описании деструктивного события.

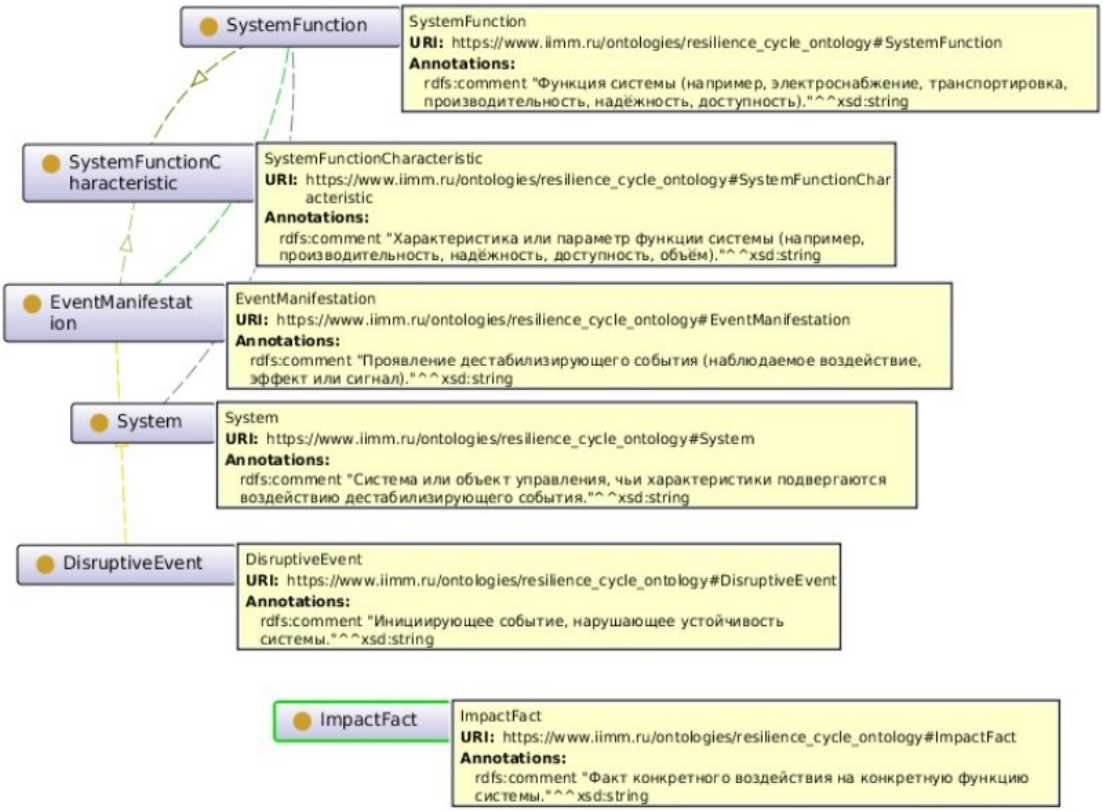


Рис. 3. Онтологический паттерн «Факт воздействия»

Представленные в качестве примеров паттерны можно условно отнести к категории элементарных. Они включают небольшое число компонентов (классов и отношений) и просты в использовании.

Путем комбинирования элементарных паттернов формируются композитные паттерны. Они представляют структуру взаимодействия элементарных паттернов, а также сложные объекты и взаимодействия предметной области.

Заключение

Предложенная формализованная структура концептуальной модели поддержки управления жизнеспособностью критических инфраструктур определяет состав классов информационных сущностей и связей между ними, задействованных в реализации всех фаз жизненного цикла процесса управления. Она включает три основных блока, представляющих систему управления жизнеспособностью, собственно объект управления, в качестве которого рассматривается критическая инфраструктура, и различные события, или процессы, оказывающие деструктивное воздействие на объект управления.

В качестве компьютерной реализации модели предложено создать онтологию на основе онтологических паттернов содержания (CDP). Для генерации паттернов использованы возможности больших языковых моделей (LLM), которые генерировали программный код паттерна в соответствии с задаваемой текстовой инструкцией, содержащей спецификацию концептуальной структуры, требующей представления в онтологии. Генерация осуществляется в два этапа, что повышает контроль над процессом генерации и снижает вероятность получения синтаксических ошибок и/или так называемых галлюцинаций в генерируемых LLM кодах.

Полный комплекс паттернов, содержащих представление всех компонентов формальной концептуальной модели процесса поддержки управления жизнеспособностью КИ, предоставит пользователю необходимый набор инструментов для предметного описания подлежащих анализу проблемных ситуаций. Онтологическая модель для конкретной ситуации формируется путем задания экземпляров представленных в паттернах сущностей, значений атрибутов этих экземпляров и определенных шаблонами связей между ними. Процедуры вывода на полученной онтологии обеспечивают формирование организационной и функциональной модели управления решением задач по устранению представленной в онтологии проблемы. В результате будет получена онтология, являющаяся экземпляром онтологии задачи управлением поддержания жизнеспособности конкретной критической инфраструктуры при определенном деструктивном воздействии.

Несмотря на то, что в общем случае ситуации нарушения нормального функционирования критических инфраструктур имеют свои особенности, сохранение уже построенных онтологических моделей полезно, так как их фрагменты в различном объеме могут быть использованы при построении онтологических моделей управления поддержанием жизнеспособности КИ и в других ситуациях. Например, могут рассматриваться аналогичные деструктивные воздействия, иметь место нарушение подобных компонентов КИ или в решении возникших проблем могут участвовать такие же акторы управления. Учитывая это, целесообразно обеспечить повторное использование не только CDP-паттернов, но и разработанных на их основе фрагментов прикладных онтологий. Для этого и каталог CDP-паттернов, и фрагменты онтологий должны быть размещены в публичном репозитории с регламентированным доступом.

Список источников

1. Маслобоев А. В., Цыгичко В. Н. Адаптация и развитие риск-ориентированной методологии обеспечения безопасности критически важных объектов к управлению жизнеспособностью критических инфраструктур // Надежность и качество сложных систем. 2024. № 4. С. 140–159.
2. Маслобоев А. В. Формальные модели жизнеспособности региональных критических инфраструктур // Труды Института системного анализа РАН. 2022. Т. 72, № 3. С. 59–80.
3. Олейник А. Г., Путилов В. А. Развитие технологии концептуального моделирования, основанной на функционально-целевом подходе // История науки и техники. 2014. № 1. С. 37–52.

4. Фридман А. Я., Олейник А. Г. Методы и средства поддержки принятия решений по обеспечению устойчивого функционирования промышленно-природных комплексов в Арктической зоне РФ // История науки и техники. 2019. № 4. С. 26–34.
5. Смирнов А. В., Пашкин М. П., Шилов Н. Г., Левашова Т. В., Кашевник А. М. Контекстно-управляемая поддержка принятия решений в распределенной информационной среде // Информационные технологии и вычислительные системы. 2009. № 1. С. 38–48.
6. Husáková M., Bureš V. Formal Ontologies in Information Systems Development: A Systematic Review // Information. 2020. Vol. 11, Iss. 2. Article no.: 0066.
7. Цыгичко В. Н., Черешкин Д. С., Смолян Г. Л. Безопасность критических инфраструктур. М.: УРСС, 2019. 200 с.
8. Pursiainen C. The Crisis Management Cycle. UK, London: Routledge, 2017. 194 p.
9. Critical Infrastructure Security and Resilience: Theories, Methods, Tools and Technologies // Advanced Sciences and Technologies for Security Applications / ed. by D. Gritzalis, M. Theocharidou, G. Stergiopoulos. Springer Cham, Springer Nature Switzerland AG, 2019. 313 p.
10. Емельянов С. В., Олейник А. Г., Попков Ю. С., Путилов В. А. Информационные технологии регионального управления. М.: Едиториал УРСС, 2004. 400 с.
11. Masloboev A. V. A ternary conceptual framework of critical infrastructure resilience management cycle // Reliability and Quality of Complex Systems. 2025. no. 3. P. 109–125.
12. Fridman A. Planning and Coordination in Hierarchies of Intelligent Dynamic Systems // TELKOMNIKA. December 2016. Vol. 14, No. 4. P. 1408–1416.
13. Яковлев С. Ю., Шемякин А. С., Олейник А. Г. Регулирование техногенно-экологической безопасности критически важных объектов инфраструктуры: обновление нормативной базы // Труды Кольского научного центра РАН. Серия: Технические науки. 2022. Т. 13, № 2. С. 93–102.
14. Анализ риска и проблем безопасности: в 4 ч. // 4.1. Основы анализа и регулирования безопасности; науч. руковод. К. В. Фролов. М.: МГФ «Знание», 2006. 640 с.
15. Deckers R., Lago P. Systematic Literature Review of Domain-Oriented Specification Techniques // Journal of Systems and Software. 2022. Vol. 192. Article no.: 111415.
16. Gehrke B., Braun M., Schenk P., West R., Michie S., Hastings J. OntoSpreadEd: Developing Ontologies with a Dedicated Online Template Spreadsheet Editor // Wellcome Open Research. 2025. Vol. 10. Article no.: 360.
17. Mateiu P., Groza A. Ontology Engineering with Large Language Models // 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). 2023. P. 226–229.
18. Gangemi A. Ontology Design Patterns for Semantic Web Content // The Semantic Web—ISWC 2005. Lecture Notes in Computer Science. 2005. Vol. 3729. P. 262–276.
19. Presutti V., Gangemi A., David S., de Cea G., Suárez-Figueroa M. C., Montiel-Ponsoda E., Poveda M. Ontology Design Patterns for Ontology Reuse and Re-engineering // Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2008). 2008. P. 128–142.
20. Blomqvist E., Presutti V., Daga E., Gangemi A. Experimenting with Ontology Design Patterns in an Ontology Engineering Context // Proceedings of the 5th Workshop on Ontology and Semantic Web Patterns (WOP 2014). CEUR Workshop Proceedings. 2014. Vol. 1302. 83 p.
21. Олейник А. Г., Ломов П. А. Разработка онтологии интегрированного пространства знаний // Онтология проектирования. 2016. Т. 6, № 4 (22). С. 465–474.
22. Lo A., Jiang A. Q., Li W., Jamnik M. End-to-End Ontology Learning with Large Language Models // Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024). 2024. Article No.: 2767. P. 87184–87225.
23. Lippolis A. S., Saeedizade M. J., Keskisärkkä R., Zuppiroli S., Ceriani M., Gangemi A., Blomqvist E., Nuzzolese A. G. Ontology Generation using Large Language Models // Proceedings of the 22nd European Semantic Web Conference, ESWC 2025, Portoroz, Slovenia, June 1–5, 2025. 2025. Part I. P. 321–341.

References

1. Masloboev A. V., Tsygichko V. N. Adaptatsiya i razvitie risk-orientirovannoy metodologii obespecheniya bezopasnosti kriticheski vazhnykh ob"ektov k upravleniyu zhiznesposobnost'yu kriticheskikh infrastruktur [Adaptation and development of a risk-based methodology for ensuring the security of critical facilities for managing the resilience of critical infrastructures]. *Nadezhnost' i kachestvo slozhnykh sistem* [Reliability and Quality of Complex Systems], 2024, no. 4, pp. 140–159. (In Russ.).

2. Masloboev A. V. Formal'nye modeli zhiznesposobnosti regional'nykh kriticheskikh infrastruktur [Formal models of the resilience of regional critical infrastructures]. *Trudy Instituta sistemnogo analiza RAN* [Proceedings of the Institute of Systems Analysis of the Russian Academy of Sciences], 2022, Vol. 72, no. 3, pp. 59–80. (In Russ.).
3. Oleynik A. G., Putilov V. A. Razvitie tekhnologii kontseptual'nogo modelirovaniya, osnovannoy na funktsional'no-tselevom podkhoda [Development of conceptual modeling technology based on a functional-target approach]. *Istoriya nauki i tekhniki* [History of Science and Technology], 2014, no. 1, pp. 37–52. (In Russ.).
4. Fridman A. Ya., Oleynik A. G. Metody i sredstva podderzhki prinyatiya resheniy po obespecheniyu ustoychivogo funktsionirovaniya promyshlenno-prirodnykh kompleksov v Arkticheskoy zone RF [Methods and means of supporting decision-making to ensure sustainable functioning of industrial and natural complexes in the Arctic zone of the Russian Federation]. *Istoriya nauki i tekhniki* [History of Science and Technology], 2019, no. 4, pp. 26–34. (In Russ.).
5. Smirnov A. V., Pashkin M. P., Shilov N. G., Levashova T. V., Kashevnik A. M. Kontekstno-upravlyaemaya podderzhka prinyatiya resheniy v raspredelennoy informatsionnoy srede [Context-driven decision support in a distributed information environment]. *Informatsionnye tekhnologii i vychislitel'nye sistemy* [Information technology and computing systems], 2009, no. 1, pp. 38–48. (In Russ.).
6. Husáková M., Bureš V. Formal Ontologies in Information Systems Development: A Systematic Review. *Information*, 2020, Vol. 11, Iss. 2, Article no.: 0066.
7. Tsygichko V. N., Chereshkin D. S., Smolyan G. L. *Bezopasnost' kriticheskikh infrastruktur* [Critical Infrastructure Security]. Moscow, URSS, 2019, 200 p. (In Russ.).
8. Pursiainen C. *The Crisis Management Cycle*. UK, London, Routledge, 2017, 194 p.
9. Critical Infrastructure Security and Resilience: Theories, Methods, Tools and Technologies. *Advanced Sciences and Technologies for Security Applications* / ed. by D. Gritzalis, M. Theocharidou, G. Stergiopoulos. Springer Cham, Springer Nature Switzerland AG, 2019, 313 p.
10. Emel'yanov S. V., Oleynik A. G., Popkov Yu. S., Putilov V. A. *Informatsionnye tekhnologii regional'nogo upravleniya* [Information technologies for regional management]. Moscow, Editorial URSS, 2004, 400 p. (In Russ.).
11. Masloboev A. V. A ternary conceptual framework of critical infrastructure resilience management cycle. *Reliability and Quality of Complex Systems*, 2025, no. 3, pp. 109–125.
12. Fridman A. Planning and Coordination in Hierarchies of Intelligent Dynamic Systems. *TELKOMNIKA*. December 2016, Vol. 14, no. 4, pp. 1408–1416.
13. Yakovlev S. Yu., Shemyakin A. S., Oleynik A. G. Regulirovanie tekhnogenno-ekologicheskoy bezopasnosti kriticheski vazhnykh ob"ektov infrastruktury: obnovlenie normativnoy bazy [Regulation of technogenic and environmental safety of critical infrastructure facilities: updating the regulatory framework]. *Trudy Kol'skogo nauchnogo tsentra RAN. Seriya: Tekhnicheskie nauki* [Proceedings of the Kola Science Center of the Russian Academy of Sciences. Series: Technical Sciences], 2022, Vol. 13, no. 2, pp. 93–102. (In Russ.).
14. *Analiz riska i problem bezopasnosti: v 4 ch. 4.1. Osnovy analiza i regulirovaniya bezopasnosti* [Analysis of Risks and Safety Issues. In 4 Parts. 4.1. Fundamentals of Safety Analysis and Regulation]. Moscow, MGF "Znanie", 2006, 640 p. (In Russ.).
15. Deckers R., Lago P. Systematic Literature Review of Domain-Oriented Specification Techniques. *Journal of Systems and Software*, 2022, Vol. 192, Article no.: 111415.
16. Gehrke B., Braun M., Schenk P., West R., Michie S., Hastings J. OntoSpreadEd: Developing Ontologies with a Dedicated Online Template Spreadsheet Editor. *Wellcome Open Research*, 2025, Vol. 10. Article no.: 360.
17. Mateiu P., Groza A. Ontology Engineering with Large Language Models. *25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 2023, pp. 226–229.
18. Gangemi A. Ontology Design Patterns for Semantic Web Content. *The Semantic Web—ISWC 2005. Lecture Notes in Computer Science*, 2005, Vol. 3729, pp. 262–276.
19. Presutti V., Gangemi A., David S., de Cea G., Suárez-Figueroa M. C., Montiel-Ponsoda E., Poveda M. Ontology Design Patterns for Ontology Reuse and Re-engineering. *Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2008)*, 2008, pp. 128–142.
20. Blomqvist E., Presutti V., Daga E., Gangemi A. Experimenting with Ontology Design Patterns in an Ontology Engineering Context. *Proceedings of the 5th Workshop on Ontology and Semantic Web Patterns (WOP 2014). CEUR Workshop Proceedings*, 2014, Vol. 1302, 83 p.
21. Oleynik A. G., Lomov P. A. Razrabotka ontologii integrirovannogo prostranstva znaniy [Development of an ontology for an integrated knowledge space]. *Ontologiya proektirovaniya* [Ontology of design], 2016, Vol. 6, no. 4 (22), pp. 465–474. (In Russ.).

22. Lo A., Jiang A. Q., Li W., Jamnik M. End-to-End Ontology Learning with Large Language Models. *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024, Article No.: 2767, pp. 87184–87225.
23. Lippolis A. S., Saeedizade M. J., Keskisärkkä R., Zuppiroli S., Ceriani M., Gangemi A., Blomqvist E., Nuzzolese A. G. Ontology Generation using Large Language Models. *Proceedings of the 22nd European Semantic Web Conference, ESWC 2025, Portoroz, Slovenia, June 1–5, 2025*, 2025, Part I, pp. 321–341.

Информация об авторах

П. А. Ломов — кандидат технических наук, старший научный сотрудник;
А. В. Маслобоев — доктор технических наук, ведущий научный сотрудник;
А. Г. Олейник — доктор технических наук, главный научный сотрудник.

Information about the authors

P. A. Lomov — Candidate of Science (Tech.), Senior Research Fellow;
A. V. Masloboev — Doctor of Science (Tech.), Leading Research Fellow;
A. G. Oleynik — Doctor of Science (Tech.), Chief Research Fellow.

Статья поступила в редакцию 15.09.2025; одобрена после рецензирования 20.09.2025; принята к публикации 08.10.2025.
The article was submitted 15.09.2025; approved after reviewing 20.09.2025; accepted for publication 08.10.2025.

Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 35–55.
Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2025. Vol. 16, No. 3. P. 35–55.

Обзорная статья
УДК 004.832
doi:10.37614/2949-1215.2025.16.3.003

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ КЛАСТЕРИЗАЦИИ В ПОДПРОСТРАНСТВАХ

Ольга Николаевна Зуенко[1]✉, Ольга Владимировна Фридман[2]
[1, 2]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[1]o.zuenko@ksc.ru✉, https://orcid.org/0000-0001-5431-7538
[2]o.fridman@ksc.ru, https://orcid.org/0000-0003-1897-4922

Аннотация

В статье приведен аналитический обзор методов кластеризации в подпространствах, которые позволяют обрабатывать данные высокой размерности, характеризующиеся большим количеством признаков и их значений. Методы обеспечивают возможность анализа данных с пропусками и зашумленных данных. Разбиение на кластеры осуществляется не в полном пространстве признаков, а в отдельных его проекциях без замены исходного набора признаков их линейными комбинациями. Это позволяет снизить размерность анализируемого признакового пространства при сохранении возможности интерпретации пользователем результатов кластеризации. Выделены и подробно описаны основные этапы процесса кластеризации в рамках рассматриваемых методов. Уделено внимание вопросу использования дополнительных пользовательских ограничений для повышения точности результирующих разбиений. Проанализированные методы находят широкое применение при решении различных задач интеллектуального анализа данных: при распознавании и обработке изображений, видео, текста, исследованиях генома.

Ключевые слова:

интеллектуальный анализ данных, кластеризация в подпространствах, дополнительные пользовательские ограничения, высокая размерность признакового пространства

Благодарности:

работа выполнена в рамках темы научно-исследовательской работы «Методы и информационные технологии мониторинга и управления региональными критическими инфраструктурами Арктической зоны Российской Федерации» (FMEZ-2025-0054). Авторы благодарят А. А. Зуенко за предложения, которые позволили повысить качество работы.

Для цитирования:

Зуенко О. Н., Фридман О. В. Аналитический обзор методов кластеризации в подпространствах // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 35–55. doi:10.37614/2949-1215.2025.16.3.003.

Survey article

ANALYTICAL REVIEW OF CLUSTERING METHODS IN SUBSPACES

Olga N. Zuenko[1]✉, Olga V. Fridman[2]
[1, 2]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[1]o.zuenko@ksc.ru✉, https://orcid.org/0000-0001-5431-7538
[2]o.fridman@ksc.ru, https://orcid.org/0000-0003-1897-4922

Abstract

The article provides an analytical overview of subspace clustering methods that allow to process high-dimensional data characterized by a large number of features and their values. The methods provide the ability to analyze missing and noisy data. Clustering is performed not in the full feature space, but in its projections, without replacing the original set of features with their linear combinations. This allows reducing the dimensionality of the feature space under consideration while maintaining the ability for the user to interpret the clustering results. The main stages of the clustering process within the considered methods are highlighted and described in detail. Attention is paid to the use of additional user constraints to improve the accuracy of the resulting partitions. The analyzed methods are widely used in various data mining problems, such as image and video recognition, text processing, and genome research.

Keywords:

data mining, subspace clustering, additional user constraints, feature space of high dimension

Acknowledgments:

The study was carried out within the framework of the Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences state assignment of the Ministry of Science and Higher Education of the Russian Federation, research topic "Methods and information technologies for monitoring and managing regional critical infrastructures in the Arctic zone of the Russian Federation" (registration number of the research topic FMEZ-2025-0054). The authors thank A. A. Zuenko for the suggestions that improved the quality of the work.

For citation:

Zuenko O. N., Fridman O. V. Analytical review of subspace clustering methods. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 35–55. doi:10.37614/2949-1215.2025.16.3.003.

Введение

Кластеризация является описательной задачей, нацеленной на идентификацию однородных групп объектов на основе значений их атрибутов (измерений) [1; 2]. Методы кластеризации широко изучались в статистике [3], распознавании образов [4; 5] и машинном обучении [6; 7].

Методы кластеризации можно в целом разделить на две категории [1; 2]: плоские и иерархические. При наличии набора объектов и критерия кластеризации плоские методы кластеризации обеспечивают разделение объектов на кластеры таким образом, что объекты в кластере более похожи друг на друга, чем объекты в разных кластерах. Иерархическая кластеризация представляет собой вложенную последовательность разбиений объектов. Агломеративная иерархическая кластеризация начинается с помещения каждого объекта в его собственный кластер, а затем осуществляется объединение этих атомарных кластеров в более крупные кластеры. Разделительная (дивизимная), иерархическая кластеризация представляет собой обратный процесс, который начинается с помещения всех объектов в один кластер с дальнейшим разбиением его на более мелкие части [1].

Стремительный рост объемов данных, характеризующихся сложностью, разрозненностью и нелинейностью, требует разработки новых методов для извлечения из них знаний, включая методы кластеризации данных высокой размерности. Каждый объект данных может характеризоваться десятками атрибутов (измерений), и каждому атрибуту может соответствовать домен, содержащий большое количество значений. Такая ситуация часто возникает при распознавании и обработке изображений [8], видео [9], текста [10] или при исследованиях генома [11]. Большое количество атрибутов, соответствующих каждому объекту, приводит к тому, что данные становятся разреженными, так как множества объектов не могут совпадать по всем измерениям [12]. Усугубляет эту проблему то, что многие измерения или комбинации измерений могут содержать шум или значения, которые равномерно распределены. Поэтому метрики расстояния, которые используют все измерения данных, могут быть неэффективными. Гораздо проще ориентироваться в структуре данных, если выполнять кластеризацию по подпространствам. Этот процесс называется кластеризацией в подпространствах [13].

При использовании данной методологии сходство объектов ищется не по всем атрибутам, а лишь по некоторым их подмножествам. Выборка должна включать семантически значимые атрибуты, что важно для интерпретации результатов кластеризации. Вычисление сходства объектов и распределение их на кластеры выполняются в подпространствах, причем в каждом подпространстве производится отдельная кластеризация, которая не зависит от атрибутов других подпространств.

Главная сложность кластеризации данных высокой размерности состоит в адекватной оценке сходства объектов. Идея методов кластеризации в подпространствах заключается в том, что объекты кластеров не обязательно должны иметь сходство по всем атрибутам, но должны обладать сходством по некоторому подмножеству атрибутов, при этом считается, что остальные не имеют отношения к структуре кластера [14]. Нет смысла искать кластеры в таком многомерном пространстве, поскольку средняя плотность точек в любом месте пространства данных, скорее всего, будет довольно низкой [15].

Кластеризация — это задача обучения без учителя, то есть объекты группируются в кластеры при отсутствии какой-либо априорной информации об их распределении. Кластеризация в подпространствах — это задача кластеризации, в которой отсутствует какая-либо априорная информация о количестве подпространств, содержащих кластеры, размерности этих подпространств и количестве кластеров, скрытых в каждом подпространстве [16].

Решение задачи кластеризации в подпространствах включает ряд этапов, затратных с точки зрения вычислительных ресурсов [17]. Первый из них заключается в отборе существенных атрибутов и выявлении интересных подпространств, в которых можно разделить объекты на требуемое количество кластеров. Этот этап обеспечивает предварительную подготовку данных. Затем следует этап кластеризации объектов внутри подпространств. Наконец, на последнем этапе требуется сжато описать полученные в результате кластеры с использованием характеристик объектов.

Первый этап является наиболее важным и сложным, поскольку для разных кластеров подходят разные подмножества атрибутов. Это явление, при котором различные признаки или сочетания признаков могут быть релевантны для разных кластеров, называется релевантностью локальных признаков или корреляцией локальных признаков [10].

Один из традиционных способов решения проблемы высокой размерности заключается в применении методов снижения размерности набора данных. Такие методы, как анализ главных компонент или преобразование Карунена-Лоева [4; 5], оптимально снижают размерность исходного пространства, формируя измерения, которые являются линейными комбинациями заданных атрибутов. Новое пространство обладает тем свойством, что расстояния между точками остаются примерно такими же, как и раньше. Хотя эти методы помогают снизить размерность пространства поиска, у них есть два недостатка. Во-первых, новые измерения могут вызывать затруднения для интерпретации результирующих кластеров. Во-вторых, эти методы неэффективны при идентификации кластеров, которые могут существовать в различных подпространствах исходного пространства данных.

Группа подходов к кластеризации данных высокой размерности основывается на концепции поиска частых паттернов [14]. Разным кластерам соответствуют разные подпространства, чтобы несоответствующие атрибуты не зашумляли кластеры. В такой постановке задачи прослеживается аналогия с поиском частых паттернов, и концепции алгоритмов, первоначально созданные для поиска частых паттернов, могут быть использованы в качестве основы парадигмы кластеризации в подпространствах.

Систематизация методов кластеризации в подпространствах

В [18] представлен обзор многих методов кластеризации в подпространствах. Эти методы можно разделить на две категории. Методы кластеризации подпространства «Сверху вниз» начинаются с начального приближения кластеров в полном пространстве признаков и итеративно уточняют кластеризацию, назначая вес каждому из измерений. Эти методы не гарантируют нахождение наилучшей кластеризации и, как правило, находят только гиперсферические кластеры [18].

Напротив, алгоритмы кластеризации подпространства «Снизу вверх» начинаются с обнаружения «интересных» кластеров в низкоразмерных подпространствах в соответствии с критериями плотности [17; 19–22] или мерой расстояния [23]. Кластеры в подпространстве итеративно объединяются для формирования кластеров подпространства более высокой размерности. «Узким местом» этих алгоритмов является NP-полнота перечисления кластеров подпространства. Чтобы сделать алгоритмы «Снизу вверх» более эффективными, целесообразно использовать пользовательские ограничения в процессе перечисления, чтобы отсечь большие части пространства поиска.

В [24] разделяют три группы методов кластеризации в подпространствах:

1. Методы на основе сетки. При таком подходе пространство данных делится на ячейки, параллельные осям [4]. Затем ячейки, количество объектов в которых превышает заданное пороговое значение, объединяются для формирования кластеров подпространства. Количество интервалов — это еще один входной параметр, который определяет диапазон значений в каждой сетке. Для удаления неперспективных ячеек и повышения эффективности используется свойство антимонотонности, как в алгоритме поиска частых паттернов Apriori. Если ячейка оказывается плотной в $k - 1$ измерении, то она учитывается при поиске плотной ячейки в k измерениях. Если границы сетки строго соблюдаются для разделения объектов, то точность результатов кластеризации снижается, поскольку могут теряться соседние объекты, которые разделены границей сетки. В данных алгоритмах качество кластеризации сильно зависит от входных параметров.

2. Методы на основе окна [25]. Кластеризация в подпространствах на основе окон устраняет недостатки кластеризации на основе ячеек, которые могут привести к пропуску важных результатов [5]. Здесь окно перемещается по значениям атрибутов, и получаются перекрывающиеся интервалы, которые используются для формирования кластеров подпространства. Размер скользящего окна является одним из параметров. Эти алгоритмы генерируют кластеры подпространств, параллельные оси.

3. Методы на основе плотности [11]. Подход не использует сетки. Кластер определяется как совокупность объектов, образующих цепочку, которые находятся на заданном расстоянии и превышают заданный порог количества объектов. Затем соседние плотные области объединяются в более крупные кластеры. Эти алгоритмы могут находить в подпространствах кластеры произвольной формы. Кластеры создаются путем объединения объектов из смежных областей с плотной структурой. Подходы на основе плотности используют значение параметра расстояния, применяемого для вычисления меры плотности, которая адаптируется к размерности подпространства.

На рисунке 1 представлена иерархия алгоритмов кластеризации в подпространствах на основе стратегии поиска.

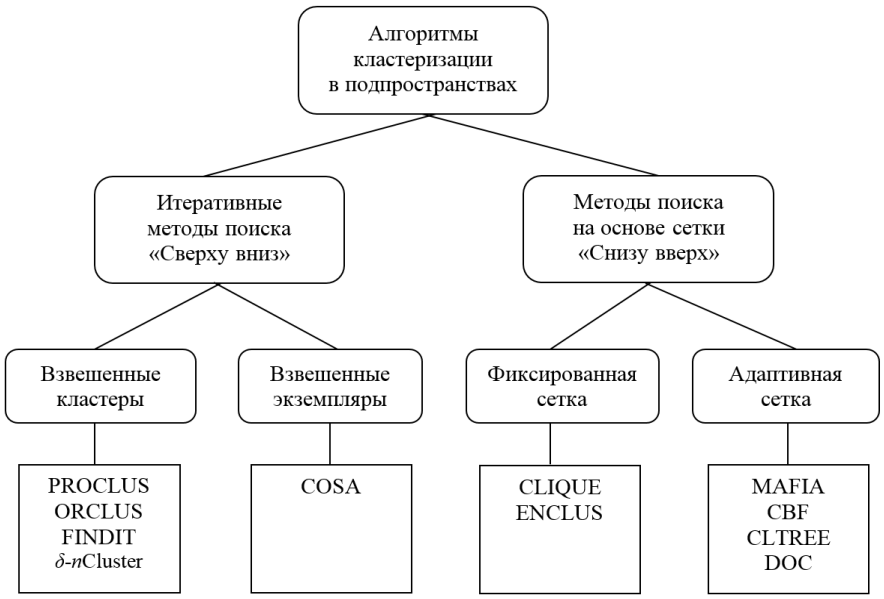


Рис. 1. Иерархия алгоритмов кластеризации в подпространствах на основе стратегии поиска

Далее рассмотрим данные подходы и алгоритмы, их реализующие, более подробно.

Алгоритмы PROCLUS [26], ORCLUS [27] и FINDIT [28] являются алгоритмами кластеризации в подпространствах, основанными на расстояниях и разбиениях. Они начинают с поиска начального приближения кластеров в полноразмерном пространстве с одинаково взвешенными измерениями, а затем каждому измерению присваивается вес для каждого кластера. Обновленные веса используются в следующей итерации для восстановления кластеров. ORCLUS [27] отличается от других подходов тем, что использует метод главных компонент (PCA) и, следовательно, проекции, в которых он ищет кластеры, не обязательно параллельны осям. Алгоритмы, основанные на PCA, также называются «корреляционной кластеризацией» [29]. Для нисходящих алгоритмов требуется два параметра: количество кластеров и средний размер подпространств, которые часто трудно определить, но которые также критически важны для производительности алгоритмов [18].

Основное отличие модели *n*Cluster [30] от подходов, основанных на плотности и сетке, заключается в способе определения кластеров. В модели *n*Cluster каждая пара объектов близка друг к другу по всем измерениям подпространств. В некотором смысле модель на основе плотности можно рассматривать как кластеризацию с одной связью, а модель *n*Cluster — как кластеризацию

с полной связью. Более того, подходы, основанные на сетке, делят область каждого измерения на неперекрывающиеся ячейки; здесь же используется подход скользящего окна для сохранения значимых кластеров. Основное различие между моделью *n*Cluster и подходами, основанными на расстоянии и разбиении, такими как PROCLUS и ORCLUS, заключается в том, что модель *n*Cluster допускает перекрытие между кластерами, в то время как подходы на основе разбиении не допускают.

Алгоритм COSA (Clustering On Subsets of Attributes) [31] уникален тем, что использует ближайших соседей для каждого экземпляра в наборе данных, чтобы определить веса всех измерений данного конкретного экземпляра. Этот итеративный алгоритм назначает веса каждому измерению для каждого экземпляра, а не кластера. После кластеризации веса измерений элементов кластера сравниваются и вычисляется итоговое значение важности для каждого измерения каждого кластера.

Модель, основанная на плотности, направлена на поиск областей с высокой плотностью в подпространствах, разделенных областями с меньшей плотностью. CLIQUE [17], MAFIA [21], CBF [32] и CLTree [33] — это алгоритмы кластеризации в подпространствах, основанные на плотности и сетке. Они дискретизируют пространство данных на неперекрывающиеся прямоугольные ячейки, разбивая каждое измерение на несколько интервалов, а затем используют поиск по алгоритму Apriori для нахождения перекрывающихся кластеров. И CLIQUE, и ENCLUS используют сетку фиксированного размера для разбиения каждого измерения на интервалы.

Другие алгоритмы применяют стратегии, основанные на данных, для определения границ каждого измерения. MAFIA и CBF используют гистограммы для анализа плотности данных в измерениях.

Стратегия CLTree основана на дереве решений. Другой алгоритм кластеризации на основе плотности — SUBCLU [11] — вместо сеточного подхода использует алгоритм DBSCAN [34] для поиска кластеров произвольной формы в отдельных подпространствах послойным способом, что может быть очень затратно.

Метод DOC [35] (Density-Based Optimal Projective Clustering) — это алгоритм Монте-Карло, который вычисляет с высокой вероятностью хорошее приближение проективного кластера. Алгоритм выполняется с помощью итераций, каждая из которых генерирует один новый кластер. Итерация останавливается, когда некоторый заданный критерий выполнен. Проективный кластер — это параллельный осям куб, который имеет максимальную длину ребра и содержит больше некоторого значения от общего числа точек. Для работы DOC также нужно задать фактор баланса, который представляет собой выбор пользователем отношения относительной важности числа точек к числу измерений в кластере.

Основные этапы кластеризации в подпространствах

Кластеризация в подпространствах направлена на выявление проекций подпространства исходного набора данных, т. е. наборов атрибутов или интервалов в пределах диапазона атрибутов, где можно найти соответствующие кластеры объектов.

Большинство алгоритмов «Снизу вверх» состоят из трех основных этапов: 1) этап предварительной обработки, на котором производится подготовка данных; 2) этап анализа данных, на котором производится поиск кластеров; 3) этап постобработки, на котором производится объединение кластеров или удаление избыточных.

Этап предварительной обработки

Этот этап в основном применяется в алгоритмах, которые сначала разбивают необработанный набор данных на (гибкую) сетку, подходящую для использования в качестве входных данных стандартными алгоритмами интеллектуального анализа наборов элементов (паттернов), такими как Apriori [26], Eclat [36], FPGrowth [37]. В этом случае создается новая бинарная таблица данных (бинарная объектно-признаковая таблица). Измерения необработанных данных разбиваются на интервалы, называемые ячейками, которые являются атрибутами этой новой таблицы.

Процесс разбиения зависит от конкретных алгоритмов кластеризации в подпространствах. Например, алгоритмы, основанные на сетке, такие как CLIQUE [17] и его модификации

(например, [21; 22; 38]), начинаются с дискретизации каждого измерения на ячейки. Количество ячеек может быть определено пользователем [20; 22; 38] или основано на распределении данных [39]. Пусть j-я ячейка измерения d_i обозначается как b_{ij}, $i \in [1, n]$, $j \in [1, p_i]$, где p_i — количество ячеек измерения d_i. Полученное разбиение исходного пространства данных можно рассматривать как многомерную сетку, где пересечение ровно одного интервала из каждого измерения образует ячейку подпространства: $u = b_{1l} \times \ldots \times b_{nq}$, где $l \in [1, p_1]$, $q \in [1, p_n]$. k-мерная ячейка — это ячейка, созданная путем пересечения k интервалов из k различных измерений. После разбиения необработанный набор данных кодируется в булевый набор данных, который указывает на наличие или отсутствие каждого объекта в каждой из ячеек.

Алгоритмы кластеризации в подпространствах на основе расстояний, такие как nCluster [23], сначала вычисляют расстояние между каждой парой объектов в каждом измерении. Для каждого измерения d_i два разных объекта находятся в одном наборе, если расстояние между ними в этом измерении ниже определенного пользователем порога $\delta \times R$ (где R — длина диапазона d_i). Количество максимальных наборов, которые могут быть созданы для каждого измерения d_i, определяет количество ячеек d_i.

Этап анализа данных

Алгоритмы «Снизу вверх» начинаются с перечисления соответствующих одномерных подпространств и итеративно генерируют подпространства более высоких размерностей, которые удовлетворяют критериям, предъявляемым к получаемым кластерам [17; 22; 23; 38–40]. Чтобы сделать процесс перечисления эффективным, критерии должны состоять из антимонотонных ограничений, которые используются для обрезки пространства поиска.

Если рассматривать предварительно обработанный набор данных как транзакционные данные, то поиск для k-мерных плотных единиц сводится к задаче поиска частых k-элементных наборов [26]. Набор транзакций сопоставляется с набором объектов O, а элементы транзакций соответствуют ранее вычисленным интервалам измерений. Тогда k-мерная ячейка с плотностью, большей, чем τ, может рассматриваться как частый набор размера k с частотой, большей, чем τ. Поскольку ячейки не пересекаются, два разных набора одинаковой размерности не могут встречаться в одном и том же частом наборе, и, таким образом, вычисленный частый набор может рассматриваться как плотная k-мерная ячейка. CLIQUE использует свойство антимонотонности ограничения на плотность: если k-мерное пространство плотное, то любое из его $(k - 1)$-мерных подпространств также плотное [17].

В случае алгоритма на основе расстояния, ячейки могут перекрываться. Это может привести к большему количеству ячеек, чем в подходах на основе плотности, что увеличивает сложность алгоритма. Чтобы преодолеть эту проблему, nCluster напрямую ищет максимальные кластеры, принимая стратегию поиска замкнутого набора элементов [41]. nCluster использует следующее антимонотонное свойство: если набор объектов O и набор ячеек D образуют δ-n-кластер (т. е. для каждых двух объектов o_w и o_z из O и каждого интервала $b_{ij} \in D$, o_w и o_z являются соседями относительно меры расстояния на d_i и порога δ), то O образует δ-n-кластер с каждым подмножеством D, а D образует $\delta - \delta$-n-кластер с каждым подмножеством O. Кроме того, он рассматривает кластер подпространства как имеющий смысл, если он состоит из m_r объектов и m_c измерений (m_c и m_r задаются предварительно). Добавление дополнительных ограничений к этому шагу интеллектуального анализа данных может сделать процесс кластеризации не только более эффективным, но и более точным.

Этап постобработки

Этап постобработки может иметь несколько целей. Одна из них — идентификация максимальных кластеров и генерация их соответствующих описаний. Например, CLIQUE генерирует k-мерно-максимальные кластеры, соединяя k-мерные плотные единицы (полученные на предыдущем шаге), имеющие общие грани. В случае nCluster применяемый алгоритм интеллектуального анализа замкнутых множеств вычисляет замкнутый набор ячеек, связанных с замкнутым набором объектов, но некоторые из них фактически могут быть не максимальными при рассмотрении измерений вместо ячеек,

вычисленных на них. Поэтому немаксимальные кластеры подпространства удаляются на этапе постобработки. В конце этого этапа максимальные кластеры описываются как пара (O, D), где O — набор объектов, а D — набор ячеек разных измерений (подпространство). В случае *n*Cluster ячейки b_{ij} заменяются соответствующей исходной размерностью d_i. Этот шаг также можно использовать для обрезки избыточных подпространственных кластеров или тех, которые, например, не соответствуют определенным ограничениям, которые не могут быть эффективно реализованы в процессе поиска (например, немонотонные ограничения) [23].

Алгоритм CLIQUE [17]

Приложения для интеллектуального анализа данных предъявляют особые требования к алгоритмам кластеризации, включая: способность находить кластеры, встроенные в подпространства данных высокой размерности, масштабируемость, понятность результатов для конечного пользователя, отсутствие предположения о каком-либо каноническом распределении данных и нечувствительность к порядку входных записей. Алгоритм CLIQUE удовлетворяет каждому из этих требований. Он идентифицирует плотные кластеры в подпространствах максимальной размерности, генерирует описания кластеров в виде выражений ДНФ, которые минимизированы для простоты понимания, выдает идентичные результаты независимо от порядка, в котором представлены входные записи.

В алгоритме CLIQUE используется подход, основанный на плотности: кластер — это область, которая имеет более высокую плотность точек, чем окружающая ее область. Необходимо автоматически идентифицировать проекции входных данных, то есть подмножество атрибутов, предполагая, что эти проекции включают области высокой плотности. Чтобы аппроксимировать плотность точек данных, пространство данных разбивается и находится количество точек, которые лежат внутри каждой ячейки (единицы) разбиения. Это достигается путем разбиения каждого измерения на одинаковое количество интервалов равной длины. Каждая ячейка имеет одинаковый объем, и поэтому количество точек внутри нее может быть использовано для аппроксимации плотности ячейки.

После нахождения соответствующих подпространств задача состоит в том, чтобы найти кластеры в этих проекциях. Кластеры представляют собой объединения связанных ячеек высокой плотности в подпространстве. Чтобы упростить их описания, кластеры ограничиваются осями — параллельными гиперпрямоугольниками (брусами).

Каждая ячейка в k-мерном подпространстве может быть описана как конъюнкция неравенств, так как она является пересечением $2k$ осей. Поскольку каждый кластер является объединением таких ячеек, его можно описать с помощью выражения ДНФ. Компактное описание получается путем покрытия кластера минимальным числом максимальных, возможно, перекрывающихся прямоугольников и описания кластера как объединения этих прямоугольников.

Кластеризация в подпространствах позволяет обрабатывать отсутствующие значения во входных данных. Точка данных считается принадлежащей определенному подпространству, если значения атрибутов в этом подпространстве не отсутствуют, независимо от значений остальных атрибутов. Это позволяет использовать записи с отсутствующими значениями для кластеризации и получать более точные результаты, чем в случае замены отсутствующих значений значениями, взятыми из распределения [17].

Постановка задачи

Пусть $\mathcal{A} = \{A_1, A_2, \ldots, A_d\}$ — набор полностью упорядоченных доменов, а $S = A_1 \times A_2 \times \ldots \times A_d$ – d-мерное числовое пространство, $A_1, A_2, \ldots, A_d$ — измерения (атрибуты) S. Входные данные состоят из набора d-мерных точек $\mathcal{V} = \{v_1, v_2, \ldots, v_m\}$, где $v_i = \langle v_{i1}, v_{i2}, \ldots, v_{id}\rangle$. j-й компонент v_i взят из домена A_j. Пространство данных S разбивается на неперекрывающиеся прямоугольные ячейки. *Ячейки* получаются путем разбиения каждого измерения на интервалы равной длины. Длина ξ является одним из входным параметров.

Каждая ячейка u является пересечением одного интервала из каждого атрибута. Она имеет вид $\{u_1, u_2, \ldots, u_d\}$, где $u_i = [l_i, h_i)$ — открытый справа интервал в разбиении A_i. Точка $v = \langle v_1, v_2, \ldots, v_{id}\rangle$

содержится в ячейке $u = \{u_1, u_2,\ldots, u_d\}$, если $l_i \le v_i < h_i$ для всех u_i. *Селективность ячейки* определяется как доля всех точек данных, содержащихся в ячейке. Ячейку u называют *плотной*, если селективность (u) больше, чем τ, где порог плотности τ является другим входным параметром.

Аналогично определяются ячейки во всех подпространствах исходного d-мерного пространства. Рассмотрим проекцию набора данных V в $A_{t1} \times A_{t2} \times\ldots\times A_{tk}$, где $k < d$ и $t_i < t_j$, если $i < j$. Ячейка в подпространстве является пересечением интервала из каждого из k атрибутов.

Кластер — это максимальный набор связанных плотных ячеек в k-измерениях. Две k-мерные ячейки u_1, u_2 связаны, если они имеют общую грань или если существует другая k-мерная ячейка u_3, такая, что u_1 связана с u_3 и u_2 связана с u_3. Ячейки $u_1 = \{r_{t1},\ldots, r_{tk}\}$ и $u_2 = \{r'_{t1},\ldots, r'_{tk}\}$ имеют общую грань, если есть $k - 1$ измерений $A_{t1},\ldots, A_{tk}$, таких, что $r_{tj} = r'_{tj}$ и либо $h_{tk} = l'_{tk}$, либо $h'_{tk} = l_{tk}$. *Область* в k измерениях — это параллельный осям прямоугольный k-мерный набор. Интересны только те области, которые могут быть представлены как объединения ячеек. Область может быть описана в виде дизъюнктивной нормальной формы (ДНФ) на интервалах доменов A_i. Область R, содержащаяся в кластере C, называется максимальной, если ни одно собственное надмножество R не содержится в C. Минимальное описание кластера — это неизбыточное покрытие кластера максимальными областями. То есть минимальное описание кластера C — это множество R максимальных областей, такое, что их объединение равно C, но объединение любого собственного подмножества R не равно C [17].

Задача кластеризации в данном случае состоит в том, что, при заданном множестве точек данных и входных параметрах ξ и τ, требуется найти кластеры во всех подпространствах исходного пространства данных и представить минимальное описание каждого кластера в виде выражения ДНФ. Приведем пример такой задачи.

Пример 1 [17]. На рис. 2 двумерное пространство (возраст, зарплата) разделено сеткой 10 × 10. Ячейка — пересечение интервалов; примером может служить ячейка $f = (30 \le \text{возраст} < 35) \wedge (1 \le \text{зарплата} < 2)$. Область представляет собой прямоугольное объединение ячеек. A и B являются областями: $A = (30 \le \text{возраст} < 50) \wedge (4 \le \text{зарплата} < 8)$ и $B = (40 \le \text{возраст} < 60) \wedge (2 \le \text{зарплата} < 6)$. Плотные ячейки затемнены, и очевидно, что $A \cup B$ является кластером. A является максимальной областью, содержащейся в этом кластере, тогда как $A \cap B$ таковой не является. Минимальным описанием для этого кластера может служить следующее выражение ДНФ: $((30 \le \text{возраст} < 50) \wedge (4 \le \text{зарплата} < 8)) \vee ((40 \le \text{возраст} < 60) \wedge (2 \le \text{зарплата} < 6))$.

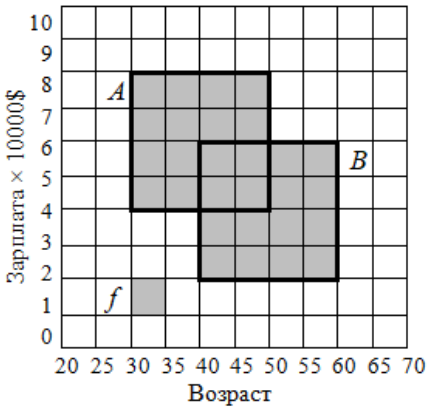


Рис. 2. Иллюстрация примера

Первый шаг алгоритма CLIQUE заключается в поиске интересных подпространств. Для этого выполняется поиск плотных ячеек, из которых в дальнейшем конструируются подпространства. Среди подпространств отбираются те, которые содержат достаточное количество точек. Для поиска плотных ячеек используется алгоритм «Снизу вверх». Алгоритм работает по уровням. Сначала он определяет

1-мерные плотные ячейки, выполняя проход по данным. Этот алгоритм похож на алгоритм Apriori для поиска частых паттернов [42]. Определив (k – 1)-мерные плотные ячейки, кандидаты k-мерных клеток определяются с помощью процедуры генерации кандидатов, описанной в [42]. Проход по данным выполняется для поиска тех кандидатов, которые являются плотными.

Следующим шагом алгоритма CLIQUE является поиск кластеров в отобранных подпространствах. Входные данные для этого шага CLIQUE — это набор плотных ячеек D, все в одном и том же k-мерном пространстве S. Выходные данные будут разбиением D на $D^1,\ldots, D^q$ таким образом, что все ячейки в D^i связаны и никакие две ячейки $u^i \in D^i$, $u^j \in D^j$ с $i \neq j$ не связаны. Каждое такое разбиение является кластером согласно определению, приведенному в [17]. Задача эквивалентна поиску компонент связности в графе, определенном следующим образом: вершины графа соответствуют плотным ячейкам и между двумя вершинами есть ребро тогда и только тогда, когда соответствующие плотные ячейки имеют общую грань. Для выявления связанных компонент графа в [43] применяется алгоритм поиска в глубину.

Входными данными для следующего шага является множество C связанных плотных ячеек в том же k-мерном пространстве S. Выходными данными будет множество $\mathcal{R}$ максимальных областей, таких, что $\mathcal{R}$ является покрытием C. Для решения этой задачи используется алгоритм жадного роста. Начинают с произвольной плотной ячейки $u_1 \in C$ и жадно наращивают (как описано ниже) максимальную область R_1, которая покрывает u_1. Добавляют R_1 к $\mathcal{R}$. Затем находят другую единицу $u_2 \in C$, которая еще не покрыта ни одной из максимальных областей в $\mathcal{R}$. Далее жадно наращивают максимальную область R_2, которая покрывает u_2. Эта процедура повторяется, пока все ячейки в C не будут покрыты некоторой максимальной областью в $\mathcal{R}$. Чтобы получить максимальную область, покрывающую плотную ячейку u, начинают с u и наращивают ее вдоль измерения a_1 как слева, так и справа от ячейки. Увеличивают u как можно больше в обоих направлениях, используя связанные плотные ячейки, содержащиеся в C. Результатом является прямоугольная область. Эта процедура повторяется для всех измерений, что дает максимальную область, покрывающую u. Порядок, в котором измерения рассматриваются для увеличения плотной области, определяется случайным образом.

Последний шаг CLIQUE принимает в качестве входных данных покрытие для каждого кластера и находит минимальное покрытие. Минимальность определяется в терминах количества максимальных областей (прямоугольников), необходимых для покрытия кластера. В [17] предлагается следующая жадная эвристика: из покрытия удаляется наименьшая (по количеству ячеек) максимальная область, которая является избыточной (т. е. каждый блок также содержится в некоторой другой максимальной области). Связи разрываются произвольно. Процедура повторяется до тех пор, пока не останется ни одной максимальной области. Также предлагается следующая эвристика сложения: кластер рассматривается как пустое пространство. К покрытию добавляется максимальная область, которая покроет максимальное количество еще непокрытых ячеек в кластере. Связи разрываются произвольно. Процедура повторяется до тех пор, пока весь кластер не будет покрыт.

На заключительном шаге генерируются описания кластеров в форме выражений ДНФ, которые минимизированы для простоты понимания (см. пример 1).

При проектировании CLIQUE были объединены разработки из нескольких областей, включая интеллектуальный анализ данных, стохастическую сложность, распознавание образов и вычислительную геометрию [17]. Он нечувствителен к порядку входных записей и не предполагает некоторого канонического распределения данных.

Алгоритм *n*Cluster

В работе [23] предложена модель кластеризации подпространства на основе расстояния, называемая *n*Cluster и использующая гибкий метод разделения измерений, который допускает перекрытие между различными ячейками атрибута. Потенциально это способно привести к большему количеству ячеек, чем в алгоритмах на основе сетки, что увеличивает сложность задачи. Чтобы решить задачу кластеризации рассматриваются только те кластеры, которые содержат нетривиальное

количество объектов и атрибутов, и извлекаются только максимальные *n*Clusters, чтобы избежать создания слишком большого количества кластеров.

Пусть O — множество объектов. Каждый объект имеет множество атрибутов A, а домены атрибутов в A ограничены. Используем x, y для обозначения объектов в O, а буквы a, b — для обозначения атрибутов в A, R_a — для обозначения диапазона значений атрибута a и v_{xa} — для обозначения значения объекта x по атрибуту a.

Расстояние между двумя объектами x и y по атрибуту a определяется как $|v_{xa} - v_{ya}|$. Если расстояние меньше предопределенного порога, то x и y называются соседями по атрибуту a. Аналогично определяются соседи объекта по подмножеству атрибутов в A, и они называются соседями по подпространству. Атрибуты обычно не имеют одинаковых диапазонов значений. Поэтому вместо использования постоянного порога для всех атрибутов используется относительный порог расстояния [23].

Определение 1 (δ-сосед в подпространстве). Пусть x, y — два объекта, а $G \subseteq A$ — подмножество атрибутов. Если для каждого номинального атрибута $a \in G$, имеется $v_{xa} = v_{ya}$, и для каждого непрерывного атрибута $a \in G$, имеется $|v_{xa} - v_{ya}| \leq \delta \cdot R_a$, где δ — предопределенный порог, то x и y называются δ-соседями друг друга в подпространстве G. Если объекты из множества T являются δ-соседями друг друга на наборе атрибутов G, то эти объекты образуют кластер в подпространстве G и называются его δ-n-кластером.

Определение 2 (подпространственный δ-n-кластер). Пусть $T \subseteq O$ — множество объектов, а $G \subseteq A$ — множество атрибутов. Если для каждых двух объектов x, $y \in T$ и каждого атрибута $a \in G$, объекты x и y являются δ-соседями по атрибуту a, то (T, G) — это подпространственный δ-n-кластер или просто δ-n-кластер.

Пример 2 [23]. В таблице 1 представлен набор данных с 4 атрибутами и 8 объектами. Диапазоны значений атрибута a, b, c и d равны [0, 20], [-50, 50], [0, 100] и [0, 30] соответственно. Если установить δ на 0,1, то кортеж ({1, 2, 4, 6, 8}, {a}) образует δ-n-кластер, так же как и кортеж ({1, 6}, {a, b, c}). Если в табл. 1 установить δ в 0,2, можно найти кластер подпространства ({2, 3, 4}, {a, b}), который не может быть найден с помощью подхода на основе сетки.

Таблица 1

Пример набора данных

№ объектов	a	b	c	d
1	5	0	27	0
2	6	50	75	24
3	3	-29	53	13
4	5	-2	51	30
5	0	1	100	7
6	6	4	29	19
7	20	27	23	1
8	7	-50	0	2

Пусть имеется два n-кластера (T_1, G_1) и (T_2, G_2), и если $T_1 \subseteq T_2$ и $G_1 \subseteq G_2$, то (T_1, D_1) является подкластером (T_2, G_2), а (T_2, G_2) является суперкластером (T_1, G_1). Если либо $T_1 \subset T_2$, либо $G_1 \subset G_2$, то (T_1, G_1) является собственным подкластером (T_2, G_2). δ-n-кластеры обладают следующим свойством, основанным на их определении.

Свойство 1 (свойство антимонотонности). Пусть $T \subseteq O$ — набор объектов, а $G \subseteq A$ — набор атрибутов. Если T и G образуют δ-n-кластер, то T образует δ-n-кластер с каждым подмножеством G, а G образует δ-n-кластер с каждым подмножеством T [23].

Чтобы кластер был осмысленным и полезным, он должен содержать нетривиальное количество объектов и атрибутов. Используется два порога m_r и m_c для ограничения минимального количества объектов и атрибутов, необходимо найти только δ-n-кластеры, содержащие не менее m_r объектов и m_c атрибутов.

Ограничение минимального количества объектов и атрибутов отфильтровывает незначительные δ-n-кластеры, но все еще может получаться слишком большое количество δ-n-кластеров, и многие из них являются избыточными в том смысле, что они могут быть включены в некоторые более крупные δ-n-кластеры. На основании свойства 1, если набор объектов T и набор атрибутов G могут образовывать δ-n-кластер, то любой его подкластер может образовывать δ-n-кластер. Чтобы избежать создания слишком большого количества δ-n-кластеров, рассматриваются только максимальные δ-n-кластеры.

Определение 3 (максимальный δ-n-кластер). Пусть $T \subseteq O$ — набор объектов, а $G \subseteq A$ — набор атрибутов, причем T и G образуют δ-n-кластер. Если не существует δ-n-кластера (T', G'), такого, что (T', G') является собственным подкластером (T, G), то (T, G) называется максимальным δ-n-кластером.

Пусть $\delta = 0{,}1$. В наборе данных из табл. 1 δ-n-кластер ($\{1, 6\}$, $\{a, b\}$) не является максимальным, потому что его набор атрибутов может быть расширен атрибутом c, а его набор объектов может быть расширен объектом 4. В свою очередь, δ-n-кластеры ($\{1, 4, 6\}$, $\{a, b\}$) и ($\{1, 6\}$, $\{a, b, c\}$) являются максимальными δ-n-кластерами, потому что ни их наборы объектов не могут быть расширены без сокращения их наборов атрибутов, ни их наборы атрибутов не могут быть расширены без сокращения их наборов объектов [23].

Поиск δ-n-кластеров с одним атрибутом

Необходимо найти максимальные δ-n-кластеры, поэтому для каждого подпространства G находятся максимальные наборы объектов, которые могут образовывать δ-n-кластеры с G. Атрибут может образовывать δ-n-кластеры с несколькими максимальными наборами объектов. Их поиск производится на основе следующего наблюдения.

Лемма 1 [23]. При наличии атрибута a и набора объектов T, кортеж (T, $\{a\}$) является δ-n-кластером тогда и только тогда, когда $max\{v_{xa} | x \in T\} - min\{v_{xa} | x \in T\} \leq \delta R_a$.

На основе вышеприведенной леммы определяют максимальные наборы объектов атрибута, используя метод, аналогичный методу, использованному в [44] для поиска наборов максимальной размерности (MDS).

Далее объекты в O сортируются в порядке возрастания их значений по атрибуту a, а затем ищутся пары позиций p_1 и p_2 ($p_1 < p_2$) в отсортированной последовательности, такие, что разность значений в двух позициях не больше δR_a, но разница между значениями в ($p_1 - 1$) и p_2 или в p_1 и ($p_2 + 1$) больше δR_a.

Если количество различных значений атрибута очень большое, то количество сгенерированных списков максимальных объектов также может быть очень большим. Это может создать трудности для алгоритма интеллектуального анализа. Чтобы избежать генерации слишком большого количества сильно перекрывающихся списков максимальных объектов по одному и тому же атрибуту, используют порог ω для управления перекрытием. Порог ω используется следующим образом. Пусть T — текущий максимальный набор объектов, обнаруженный по атрибуту a, а R_T — диапазон T, т. е. $R_T = [min_{x \in T} \{v_{xa}\}, max_{x \in T} \{v_{xa}\}]$. Тогда диапазон следующего максимального набора объектов не может иметь более $\omega |R_T|$ перекрытия с R_T. Когда $\omega = 0$, атрибуты делятся на неперекрывающиеся ячейки, как в подходе на основе сетки.

В таблице 2 показаны максимальные наборы объектов для всех атрибутов, сформированные на основе табл. 1. Каждый атрибут и его максимальный набор объектов образуют δ-n-кластер. Эти δ-n-кластеры используются в качестве отправных точек для поиска δ-n-кластеров, содержащих большее количество атрибутов.

Таблица 2

Максимальные наборы объектов атрибутов

Атрибут	Максимальный набор объектов
a_1	{1, 3, 4}
a_2	{1, 2, 4, 6, 8}
b_1	{1, 4, 5, 6}
c_1	{1, 6, 7}
c_2	{3, 4}
d_1	{1, 7, 8}

Поиск максимальных δ-n-кластеров, содержащих более одного атрибута

Для δ-n-кластера (T, G) и атрибута $a \notin G$, если есть по крайней мере m_r объектов в T, являющихся δ-соседями друг друга по атрибуту a, то атрибут a можно добавить в G, чтобы сформировать δ-n-кластер с еще одним атрибутом. Чтобы найти все такие атрибуты a, мы поддерживаем список атрибутов для каждого объекта. Список атрибутов объекта x содержит все атрибуты, по которым у объекта x имеется по крайней мере $m_r - 1$ δ-соседей. Например, в табл. 2 атрибут a имеет два максимальных набора объектов. Если просто добавить a к спискам атрибутов всех объектов, содержащихся в его двух максимальных наборах объектов, то нельзя определить, какие объекты находятся в одних и тех же максимальных наборах объектов.

Чтобы решить эту проблему, при добавлении имени атрибута к спискам атрибутов объектов следует добавить нижний индекс к имени атрибута. Списки атрибутов объектов в одном и том же максимальном наборе объектов получают атрибут с тем же нижним индексом. Имя атрибута с нижним индексом называется символом атрибута, чтобы отличать его от самого атрибута. Количество символов атрибута равно количеству максимальных наборов объектов атрибута, а частота символа атрибута в списках атрибутов равна размеру максимального набора объектов, который представляет символ. В приведенном выше примере атрибут a имеет два символа a_1 и a_2, списки атрибутов объектов 1, 3 и 4 содержат a_1, а списки атрибутов объектов 1, 2, 4, 6 и 8 содержат a_2. Списки атрибутов всех объектов в табл. 1 показаны в табл. 3.

Таблица 3

Списки атрибутов объектов

Объект	Список атрибутов
1	a_1, a_2, b_1, c_1, d_1
2	a_2
3	a_1, c_2
4	a_1, a_2, b_1, c_2
5	b_1
6	a_2, b_1, c_1
7	c_1, d_1
8	a_2, d_1

Лемма 2. Два объекта являются δ-соседями по атрибуту a тогда и только тогда, когда списки атрибутов двух объектов содержат один и тот же символ атрибута a.

Поскольку списки атрибутов содержат полную информацию, списки атрибутов используются для обнаружения максимальных δ-n-кластеров в последующем поиске. Алгоритм поиска основан на следующем наблюдении.

Лемма 3. Набор атрибутов G образует δ-n-кластер с набором объектов T тогда и только тогда, когда списки атрибутов объектов в T содержат один и тот же символ каждого атрибута в G [39].

Если рассматривать символы атрибутов как элементы, наборы символов атрибутов как наборы элементов, а списки атрибутов как транзакции, то поиск δ-n-кластеров можно представить как поиск частых наборов элементов (паттернов) из базы данных транзакций [16]. Концепция максимальных δ-n-кластеров используется в статье для удаления избыточных δ-n-кластеров, и она похожа на концепцию частых замкнутых паттернов [45], которая используется для удаления избыточных наборов элементов при поиске паттернов.

Набор элементов (паттерн) замкнут, если он максимален в наборе транзакций, содержащих его. Если δ-n-кластер максимален, то соответствующий ему набор символов атрибутов является замкнутым паттерном относительно списка атрибутов. Таким образом, появляется возможность применять алгоритмы поиска частых замкнутых паттернов для получения максимальных δ-n-кластеров.

В работе [23] используется LCM [46] для поиска максимальных n-кластеров. Замкнутый набор элементов (паттерн) не всегда соответствует максимальному δ-n-кластеру. Например,

$\{a_1, a_2, b_1\}$ — это замкнутый набор элементов в табл. 3, а соответствующий ему набор атрибутов — $\{a, b\}$, а набор объектов — $\{1, 4\}$, но ($\{1, 4\}$, $\{a, b\}$) не является максимальным n-кластером, поскольку один из его суперкластеров ($\{1, 4, 6\}$, $\{a, b\}$) также является δ-n-кластером. Немаксимальные δ-n-кластеры удаляются на этапе постобработки.

Использование дополнительных ограничений в процедурах кластеризации в подпространствах

Ограничения на экземпляры кластера

Все существующие алгоритмы кластеризации в подпространствах используют входные параметры, которые можно рассматривать как ограничения. Например, CLIQUE [17] использует пороговое значение минимальной плотности, которую может иметь кластер. Хотя небольшие изменения этих параметров могут полностью изменить результат кластеризации, значения этих пороговых значений обычно никогда заранее не известны пользователю. С другой стороны, более интуитивные ограничения, такие как знание априорной группировки некоторых объектов внутри кластеров, могут принести существенную пользу с точки зрения повышения эффективности алгоритмов кластеризации.

Эти ограничения, известные как ограничения уровня экземпляра, были введены в [30] и успешно применялись к различным традиционным плоским алгоритмам, алгоритмам агломеративной кластеризации [47–49] и деревьям прогнозирующей кластеризации [50]. Несмотря на то, что анализ ограничений позволяет получать более значимые кластеры, зачастую их применение приводит к увеличению сложности алгоритмов [17]. Таким образом, актуальной проблемой является разработка методов кластеризации, которые использовали бы ограничения для усечения пространства поиска и в конечном счете ускорения получения решения. Также с точки зрения уменьшения вычислительной сложности процедур кластеризации в подпространствах желательно, чтобы пользователь изначально указывал только интересующие его измерения, где следует производить кластеризацию. Однако в реальности это далеко не всегда возможно.

Существуют различные типы предпочтений пользователя и фоновых знаний предметной области, которые целесообразно учитывать и анализировать в процессе кластеризации, например: ожидаемое количество кластеров, минимальный или максимальный размеры кластера, веса для различных объектов и измерений, ограничения на параметры кластеризации (порог плотности, выбранная функция расстояния, порог энтропии и т. д.), а также ограничения на уровне экземпляра, такие как *must-link* (два объекта должны быть в одном кластере) и *cannot-link* (два объекта должны быть в разных кластерах). При наличии двух последних ограничений алгоритм становится алгоритмом с частичным привлечением учителя [15].

В работе [51] представлен алгоритм кластеризации в подпространствах на основе ограничений SC-MINER, который находит кластеры в подпространствах с учетом ограничений на экземпляры объектов. Авторы предлагают расширить общую структуру для методов кластеризации в подпространствах «Снизу вверх», интегрировав ограничения *must-link* и *cannot-link* в процедуры кластеризации на этапе анализа данных, чтобы повысить не только эффективность алгоритмов, но и их точность. Эти два ограничения позволяют конечному пользователю влиять на результаты кластеризации в подпространствах, добавляя некоторые экспертные знания. В [51] эти ограничения определяются следующим образом.

Определение 4 (ограничение *cannot-link*). Ограничение *cannot-link* на объекты o_i и o_j, записанное $CL(o_i, o_j)$, удовлетворяется в процессе кластеризации в подпространстве SC тогда и только тогда, когда для каждого кластера подпространства $(O, D) \in SC$, $\{o_i, o_j\} \not\subseteq O$.

Определение 5 (ограничение *must-link*). Ограничение *must-link* на объекты o_i и o_j, записанное $ML(o_i, o_j)$, удовлетворяется в процессе кластеризации в подпространстве SC тогда и только тогда, когда для каждого кластера подпространства $(O, D) \in SC$, $\{o_i, o_j\} \subseteq O$ или $\{o_i, o_j\} \cap O = \emptyset$.

Следующие свойства показывают, как ограничения *must-link* и *cannot-link* можно использовать для эффективного сокращения перечисления кластеров в подпространствах.

Свойство 2. Ограничение *cannot-link* $CL(o_i, o_j)$ является антимонотонным относительно $\subseteq$: $\forall P \subseteq O$: $\{o_i, o_j\} \not\subseteq O \Rightarrow \{o_i, o_j\} \not\subseteq P$.

Свойство 3. Ограничение *must-link* $ML(o_i, o_j)$ является дизъюнкцией монотонного и антимонотонного ограничений относительно $\subseteq$: $\{o_i, o_j\} \subseteq O$ монотонно, а $\{o_i, o_j\} \cap O = \emptyset$ антимонотонно:

$$\forall P \subseteq O\colon \{o_i, o_j\} \subseteq P \Rightarrow \{o_i, o_j\} \subseteq O,$$

$$\forall P \subseteq O\colon \{o_i, o_j\} \cap O = \emptyset \Rightarrow \{o_i, o_j\} \cap P = \emptyset.$$

Свойство 2 означает, что для выполнения ограничения $CL(o_i, o_j)$ необходимо искать кластеры подпространства, где объекты o_i и o_j никогда не присутствуют вместе. Поскольку k-мерный кластер никогда не может содержать больше объектов, чем любая из его $k-1$-мерных проекций, ограничение *cannot-link* является антимонотонным.

Свойство 3 гласит, что для выполнения ограничения $ML(o_i, o_j)$ необходимо искать кластеры подпространства, где объекты o_i и o_j либо оба присутствуют, либо оба отсутствуют. Если один из объектов присутствует, а другой отсутствует, кластер подпространства не имеет значения.

Эффективное использование этих ограничений может зависеть от процесса перечисления кластеров. Действительно, такие алгоритмы, как SUBCLU [11] или DUSC [40], напрямую обнаруживают кластеры в подпространстве, применяя алгоритм, подобный DBSCAN. Как следствие, ограничения на экземпляры объектов должны быть непосредственно введены в этот алгоритм, как, например, в C-DBSCAN [23].

Существующие методы «Снизу вверх» используют алгоритмы поиска замкнутых паттернов для перечисления возможных кластеров подпространства. Однако обычные алгоритмы поиска паттернов (например, Apriori [26]) не обрабатывают антимонотонные ограничения совместно с монотонными ограничениями, поскольку введение монотонных ограничений может привести к сокращению антимонотонного отсечения [52; 53].

В работе [51] расширяется алгоритм DMINER [54], который является алгоритмом интеллектуального анализа данных, до алгоритма SC-MINER, который обрабатывает ограничения на экземпляры объектов. Производится поиск только замкнутых наборов элементов, чтобы избежать избыточных подпространств, которые могут возникнуть при использовании предыдущих методов.

На первом этапе работы алгоритма SC-MINER выполняется *генерация кандидатов*. Основная техника, используемая SC-MINER для обработки ограничений на экземпляры объектов, основана на универсальном алгоритме «разделяй и властвуй», предложенном в [54; 55]. SC-MINER рекурсивно перечисляет в глубину сначала все кластеры подпространства (O, D), которые содержат элемент (объект или интервал) a, а затем все кластеры подпространства, которые не содержат a. В процессе перечисления элементы, которые уже были перечислены, отделяются от тех, которые еще предстоит перечислить. Кандидат в процессе перечисления описывается тройкой $\langle X, Y, N\rangle$, состоящей из трех пар кортежей:

пара $X = (O, D)$ — это набор объектов и набор интервалов, содержащихся в кандидате и его потомках (полученных рекурсивно). Эти элементы уже были перечислены как члены строящихся кластеров в подпространствах;

пара $Y = (O', D')$ содержит объекты и интервалы, которые еще предстоит перечислить;

пара $N = (O_N, D_N)$ используется для обеспечения ограничения близости, которое не является ни монотонным, ни антимонотонным. Эти элементы не принадлежат ни к одному подпространственному кластеру, находящемуся в стадии построения [51].

Кластеры подпространства состоят из максимальных наборов объектов и интервалов, которые находятся в отношении: каждый объект O должен принадлежать k-мерной ячейке, определенной в D, и каждый интервал в D должен содержать каждый объект O.

Следующим этапом является *отсечение.*

Алгоритм SC-MINER может накладывать монотонные и антимонотонные ограничения на набор объектов или на набор ячеек. Это интересное свойство основано на том факте, что для заданного кандидата $\langle X, Y, N\rangle$ кластеры подпространства (O_i, D_i), которые могут быть сгенерированы, удовлетворяют следующим выражениям: $O \subseteq O_i \subseteq O \cup O'$ и $D \subseteq D_i \subseteq D \cup D'$.

Как и во многих алгоритмах интеллектуального анализа данных, проверяется согласованность монотонных и антимонотонных ограничений, чтобы в конечном итоге остановить процесс рекурсии и, таким образом, обрезать пространство поиска [51].

Удовлетворение ограничения на плотность сводится к отсечению ячеек (O, D), имеющих плотность меньше τ, т. е. $|O|/|\mathcal{O}| < \tau$.

Алгоритм SC-MINER проверяет ограничение на близость, что гарантирует, что каждый элемент, который был отсеян в процессе генерации кандидатов, не может быть добавлен к текущему кандидату, т. е. не находится ли каждый такой элемент в N в связи хотя бы с одним элементом из $X \cup Y$. В противном случае кандидат и все его потомки не являются замкнутыми (поскольку данный элемент из N может быть добавлен для формирования большей ячейки, имеющей ту же поддержку) и их можно безопасно обрезать [51].

Этап распространения. На этом этапе SC-MINER проверяет, является ли (O, D) ячейкой: когда элемент a помещается в X, функция PROPAGATION удаляет все элементы Y, которые не связаны с a в наборе данных. Кроме того, PROPAGATION использует преимущества ограничений на экземпляры объектов следующим образом:

для ограничения *cannot-link* $CL(o_i, o_j)$, когда генерируется кандидат $\langle X \cup \{o_i\}, Y \setminus \{o_i\}, N\rangle$, функция PROPAGATION автоматически удаляет o_j из Y;

для ограничения *must-link* $ML(o_i, o_j)$, когда генерируется кандидат $\langle X \cup \{o_i\}, Y \setminus \{o_i\}, N\rangle$, функция PROPAGATION автоматически помещает o_j в X и удаляет элемент из D', который не связан с o_j. Когда генерируется кандидат $\langle X, Y \setminus \{o_i\}, N \cup \{o_i\}\rangle$, функция PROPAGATION автоматически удаляет o_j из Y [51].

Ограничения на формируемые подпространства

В кластеризации данных обычно не рассматриваются двоичные данные баз транзакций или дискретные данные в целом, а вместо этого чаще всего изучаются непрерывные векторные данные с действительными значениями, как правило, предполагается евклидово векторное пространство. В этом пространстве атрибуты могут быть зашумленными или даже нерелевантными для определенных кластеров. Если измерять сходство по всему пространству, т. е. по всем атрибутам, обнаружение таких кластеров в подпространствах становится все более трудным для большего числа нерелевантных измерений. С целью определения релевантных атрибутов и измерения сходства только по ним фундаментальная алгоритмическая идея алгоритма Apriori [26] была перенесена на кластеризацию в евклидовых пространствах, что привело к задаче «кластеризации в подпространствах», которая была определена как «нахождение всех кластеров во всех подпространствах» [17].

Этот перенос осуществлялся разными способами. Наиболее важными вариантами являются определение кластеров в подпространствах, которые, в свою очередь, квалифицируют некоторое подпространство как «частый паттерн», или определение *интересных* подпространств без прямой кластеризации, но как предпосылка для последующей кластеризации в этих подпространствах или как неотъемлемая часть некоторой процедуры кластеризации.

После идентификации подпространств для поиска кластеров внутри этих подпространств применяются традиционные алгоритмы кластеризации, выполняется адаптация меры расстояния к этим подпространствам в фактической процедуре кластеризации или итеративно уточняются кластеры и соответствующие подпространства. Таким образом, «интересные» подпространства здесь выявляются не по кластерам, которые они содержат (что специфично для конкретной модели кластеризации), «интересные» подпространства определяются более обобщенно, например, с точки зрения того, насколько сильно в них взаимодействуют атрибуты. В поиске подпространств, как и в кластеризации в подпространствах, концепции «элементов» и «наборов элементов» из анализа частых паттернов трансформируются в «измерение» и «подпространство» соответственно. Понятие «частого паттерна» в соответствии с некоторым пороговым значением частоты здесь становится «интересным подпространством» в соответствии с некоторой мерой «интересности» [14].

Поиск подпространств, основанный на концепции анализа частых паттернов, применяется как отдельно от конкретных алгоритмов кластеризации, так и в составе некоторого алгоритма кластеризации,

но независимо от модели кластера. В первом сценарии можно рассматривать поиск подпространств как глобальную идентификацию «интересных» подпространств, таких, в которых будут существовать кластеры, и, следовательно, как сужение пространства поиска. Во втором сценарии наблюдается локальная идентификация «интересных» подпространств. Типичным вариантом использования этих «локально интересных» подпространств является локальная адаптация мер расстояния, то есть для разных кластеров применяются разные меры сходства.

Например, в [20]:

(1) предлагается алгоритм ENCLUS, который позволяет определить новые значимые критерии высокой плотности и корреляции измерений для качества кластеризации в подпространствах;

(2) вводится использование энтропии и приводятся доказательства в поддержку ее использования;

(3) используются два свойства замыкания, основанные на энтропии, для эффективного отсечения неинтересных подпространств;

(4) предлагается механизм поиска не минимально коррелированных подпространств, которые представляют интерес из-за сильной кластеризации.

Для определения интересных подпространств используется энтропия Шеннона [56]. Энтропия измеряет неопределенность случайной величины, где высокое значение означает высокий уровень неопределенности. Равномерное распределение подразумевает наибольшую неопределенность, поэтому низкое значение энтропии (ниже некоторого порога) используется в качестве указания на кластеры подпространства.

«Интересными» подпространствами в этом смысле являются те, значение энтропии которых ниже (на некоторый порог), чем сумма энтропии каждого из его одномерных подпространств. Используя оба критерия, наиболее «интересные» подпространства для кластеризации подпространства согласно ENCLUS не расположены ни вверху, ни внизу пространства поиска подпространств, а находятся в некоторой средней размерности.

В работе [57] предложен алгоритм RIS (Ranking Interesting Subspaces), который позволяет представить этап предварительной обработки для традиционных алгоритмов кластеризации и обнаруживает все «интересные» подпространства высокоразмерных данных, содержащих кластеры. Для изучения всех таких подпространств определяется критерий качества «интересности» подпространства. Подпространства являются «интересными», если они имеют большое количество точек в окрестностях основных точек (т. е. точек с высокой локальной плотностью точек в соответствии с некоторыми пороговыми значениями), нормализованными по ожидаемому количеству точек, предполагающих равномерное распределение. Хотя этот критерий использует основанное на плотности понятие [58] «интересности», он не привязан к конкретному алгоритму кластеризации. Следовательно, ожидается, что эти подпространства окажутся «интересными» для различных алгоритмов кластеризации на основе плотности.

Заключение

Проведенный анализ методов кластеризации в подпространствах показал, что при их реализации находят широкое применение алгоритмы поиска частых паттернов. Известно, что, как правило, поиск всех частых паттернов бессмыслен, поскольку пользователю нужны только наиболее «информативные» («интересные») зависимости на данных. Аналогичная ситуация наблюдается и при поиске всех подпространств изначально заданного пространства признаков: анализ всех сгенерированных подпространств конечным пользователем практически неосуществим. Следует предоставлять пользователю только те подпространства, которые представляют определенный интерес, в частности, те подпространства, где содержится заданная доля объектов исходной обучающей выборки и где можно разделить ячейки на заданное число кластеров.

Значительную пользу для выявления «интересных» подпространств может принести учет дополнительных пользовательских ограничений. Однако в ходе исследований было выявлено, что анализ таких ограничений при развитии существующих методов, основанных на алгоритме Apriori, зачастую сопряжен со снижением их производительности.

Таким образом, проведенный анализ показал, что на настоящий момент актуальна проблема разработки эффективных методов кластеризации в подпространствах, которые учитывали бы дополнительные пользовательские ограничения к виду получаемого решения и использовали бы их для ускорения процесса поиска и повышения точности процесса кластеризации.

Список источников

1. Jain A. K., Dubes R. C. Algorithms for Clustering Data. Prentice Hall, 1988.
2. Kaufman L., Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, 1990.
3. Arabie P., Hubert L. J. An overview of combinatorial data analysis. Clustering and Classification, World Scientic Pub., New Jersey, 1996. P. 5–63.
4. Duda R. O., Hart P. E. Pattern Classification and Scene Analysis. John Wiley and Sons, 1973.
5. Fukunaga K. Introduction to Statistical Pattern Recognition. Academic Press, 1990.
6. Cheeseman P., Stutz J. Bayesian classification (autoclass): Theory and results. Advances in Knowledge. Discovery and Data Mining, AAAI/MIT Press, 1996. chapter 6. P. 153–180.
7. Michalski R. S., Stepp R. E. Learning from observation: Conceptual clustering. Machine Learning: An Artificial Intelligence Approach, Morgan Kaufmann. 1983. Vol. 1. P. 331–363.
8. Mittal H, Pandey A. C., Saraswat M. et al. A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets // Multim Tools App. 2022. Vol. 81. P. 1–26.
9. Xu R., Wunsch D. C. Clustering algorithms in biomedical research: a review // IEEE Rev Biomed Eng. 2010. Vol. 3. P. 120–154.
10. Kriegel H. P., Kroger P., Zimek A. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering // ACM Transactions on Knowledge Discovery from Data (TKDD). 2009. Vol. 3, Iss. 1. P. 1–58.
11. Kailing K., Kriegel H. P., Kroger P. Density-connected subspace clustering for high-dimensional data // Proceedings of the SDM. 2004. P. 246–257.
12. Hu J. Subspace clustering methods for understandable information organization // Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the School of Computing Science Faculty of Applied Sciences. Simon Fraser University. 2017.
13. Kaur A., Datta A. A novel algorithm for fast and scalable subspace clustering of high-dimensional data // Journal of Big Data. 2015. Vol. 2, Iss. 17, 24 p.
14. Charu C. Aggarwal, Han J. Editors. Frequent Pattern Mining. Springer Cham, 2014. 490 p.
15. Berchtold S., Bohm C., Keim D., Kriegel H.-P. A cost model for nearest neighbor search in high-dimensional data space // Proceedings of the 16th Symposium on Principles of Database Systems (PODS). 1997. P. 78–86.
16. Parsons L., Haque E., Liu H. Subspace clustering for high dimensional data: a review // ACM, Sigkdd explorations newsletter. 2004. Vol. 6, Iss. 1. P. 90–105.
17. Agrawal R., Gehrke J., Gunopulos D., Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications // Proceedings of the 1998 ACM SIGMOD Conference. 1998. P. 94–105.
18. Parsons L., Haque E., Liu H. Subspace clustering for high dimensional data: a review // SIGKDD Explor Newslett. 2004. № 6 (1). P. 90–105.
19. Assent I., Krieger R., Muller E., Seidl T. Dusc: Dimensionality unbiased subspace clustering // Proceedings of the IEEE ICDM. 2007. P. 409–414.
20. Cheng C.-H., Fu A. W., Zhang Y. Entropy-based subspace clustering for mining numerical data // Proceedings of the KDD. 1999. P. 84–93.
21. Nagesh H., Goil S., Choudhary A. Mafia: Efficient and scalable subspace clustering for very large data sets // Technical Report 9906-010, Northwestern University, 1999.
22. Sequeira K., Zaki M. J. Schism: A new approach for interesting subspace mining // Proceedings of the IEEE ICDM. 2004. P. 186–193.
23. Liu G., Li J., Sim K., Wong L. Distance based subspace clustering with flexible dimension partitioning // Proceedings of the IEEE ICDE. 2007. P. 1250–1254.
24. Kelkar B. A., Rodd S. F. Subspace Clustering—A Survey // Proceedings of ICDMAI. 2018. Vol. 1. P. 209–220.
25. Liu G., Sim K., Li J., Wong L. Efficient Mining of Distance-Based Subspace Clusters. Wiley Periodicals, Inc. 2009. Published online in Wiley InterScience [Электронный ресурс]. URL: www.interscience.wiley.com (дата обращения: 10.10.2025).

26. Aggarwal C. C., Procopiuc C. M., Wolf J. L., Yu P. S., Park J. S. Fast algorithms for projected clustering // Proceedings of the 1999 ACM SIGMOD Conference. Philadelphia, Pennsylvania, USA. 1999. P. 61–72.
27. Aggarwal C. C., Yu P. S. Finding generalized projected clusters in high dimensional spaces // Proceedings of the 2000 ACM SIGMOD Conference. Dallas, Texas, USA. 2000. P. 70–81.
28. Woo K.-G., Lee J.-H., Kim M.-H., Lee Y.-J. Findit: a fast and intelligent subspace clustering algorithm using dimension voting // Inf. Soft Tech. 2004. No. 46 (4). P. 255–271.
29. Kriegel H.-P., Kroger P., Zimek A. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. // ACM Trans Knowl Dis Data Min. 2009. № 3(1), P. 1–58.
30. Guimei L., Kelvin S., Jinyan L., Limsoon W. Distance Based Subspace Clustering with Flexible Dimension Partitioning // IEEE. 2007 [Электронный ресурс]. URL: https://www.researchgate.net/publication/4250954_Distance_Based_Subspace_Clustering_with_Flexible_Dimension_Partitioning (дата обращения: 10.10.2025).
31. Gan G. Subspace clustering for high dimensional categorical data // ACM SIGKDD Explorations Newsletter. 2003. Vol. 6, Iss. 2. P. 87–94.
32. Chang J.-W., Jin D.-S. A new cell-based clustering method for large, high-dimensional data in data mining applications // Proceedings of the 2002 ACM symposium on Applied computing. 2002. P. 503–507.
33. Liu B., Xia Y., Yu P. S. Clustering through decision tree construction // Proceedings of the 9th CIKM conference. 2000. P. 20–29.
34. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the 2nd ACM SIGKDD Conference. Portland, Oregon, USA. 1996. P. 226–231.
35. Procopiuc C. M., Jones M., Agarwal P. K., Murali T. M. A Monte Carlo algorithm for fast projective clustering // Proceedings of the ACM SIGMOD Intern. conf. on management of data. Madison, Wisconsin, USA, 2002. P. 418–427.
36. Zaki M. J. Scalable algorithms for association mining // IEEE Transactions on Knowledge and Data Engineering. 2000. No. 12 (3). P. 372–390.
37. Han J., Pei J., Yin Y. Mining frequent patterns without candidate generation // Proceedings of the ACM SIGMOD. 2000. P. 1–12.
38. Cheng C. H., Fu A. W.-C., Zhang Y. Entropy-based subspace clustering for mining numerical data // Proceedings of the 5th ACM SIGKDD Conference. 1999. P. 84–93.
39. Newman D. J., Hettich S., Blake C. L., Merz C. J. UCI repository of machine learning databases // Department of Information and Computer Science, University of California, Irvine, 1998 [Электронный ресурс]. URL: http//www.ics.uci.edu/~mlearn/MLRepository.html (дата обращения: 10.10.2025).
40. Assent I., Krieger R., Muller E., Seidl T. Dusc: dimensionality unbiased subspace clustering // Proceedings of the 7th ICDM Conference. Omaha, Nebraska, USA. 2007. P. 409–414.
41. Uno T., Kiyomi M., Arimura H. Lcm ver.3 // Proceedings of the ACM OSDM workshop. 2005. P. 77–86.
42. Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A. I. Fast Discovery of Association Rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press. 1996. chapter 12. P. 307–328.
43. Aho A., Hopcroft J., Ullman J. The Design and Analysis of Computer Algorithms. Addison-Welsle, 1974.
44. Wang H., Wang W., Yang J., Yu P. S. Clustering by pattern similarity in large data sets // Proceedings of the 2002 ACM SIGMOD Conference. 2002. P. 394–405.
45. Pasquier N., Bastide Y., Taouil R., Lakhal L. Discovering frequent closed itemsets for association rules // Proceedings of the 7th ICDT Conference. Jerusalem, Israel. 1999. P. 398–416.
46. Rymon R. Search through systematic set enumeration // Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, Cambridge, Massachusetts, USA, 1992.
47. Bilenko M., Basu S., Mooney R. J. Integrating constraints and metric learning in semi-supervised clustering // Proceedings of the ICML. 2004. P. 11.
48. Klein D., Kamvar S. D., Manning C. D. From instance level constraints to space-level constraints: Making the most of prior knowledge in data clustering // Proceedings of the ICML. 2002. P. 307–314.
49. Ruiz C., Spiliopoulou M., Ruiz E. M. C-dbscan: Density-based clustering with constraints // RSFDGrC. 2007. Vol. 4482, P. 216–223.
50. Struyf J., Dzeroski S. Clustering trees with instance level constraints // ECML. 2007. P. 359–370.
51. Fromont E., Prado A., Robardet C. Constraint-based Subspace Clustering [Электронный ресурс]. URL: https://www.researchgate.net/publication/29605468_Constraint-based_Subspace_Clustering (дата обращения: 10.10.2025).

52. Bonchi F., Giannotti F., Mazzanti A., Pedreschi D. Adaptive constraint pushing in frequent pattern mining // Proceedings of the PKDD. 2003. P. 47–58.
53. Jeudy B., Boulicaut J.-F. Optimization of association rule mining queries // Intell. Data Anal. 2002. No. 6 (4). P. 341–357.
54. Besson J., Robardet C., Boulicaut J.-F., Rome S. Constraint-based concept mining and its application to microarray data analysis // Intell. Data Anal. 2005. No. 9 (1). P. 59–82.
55. Gely A. A generic algorithm for generating closed sets of a binary relation // Proceedings of the ICFCA. 2005. P. 223–234.
56. Shannon C. E., Weaver W. The Mathematical Theory of Communication. University of Illinois Press, 1949.
57. Kailing K., Kriegel H.-P., Kröger P., Wanka S. Ranking interesting subspaces for clustering high dimensional data // Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Cavtat-Dubrovnik, Croatia, 2003. P. 241–252.
58. Kriegel H.-P., Kröger P., Sander J., Zimek A. Density-based clustering // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. No. 1 (3). P. 231–240.

References

1. Jain A. K., Dubes R. C. *Algorithms for Clustering Data*. Prentice Hall, 1988.
2. Kaufman L., Rousseeuw P. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, 1990.
3. Arabie P., Hubert L. J. An overview of combinatorial data analysis. *Clustering and Classification*, World Scientic Pub., New Jersey, 1996, pp. 5–63.
4. Duda R. O., Hart P. E. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.
5. Fukunaga K. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
6. Cheeseman P., Stutz J. Bayesian classification (autoclass): Theory and results. *Advances in Knowledge. Discovery and Data Mining*, AAAI/MIT Press, 1996, chapter 6, pp. 153–180.
7. Michalski R. S., Stepp R. E. Learning from observation: Conceptual clustering. *Machine Learning: An Artificial Intel ligence Approach*, Morgan Kaufmann, 1983, vol. 1, pp. 331–363.
8. Mittal H., Pandey A. C, Saraswat M., Kumar S., Pal R., Modwel G. A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multim Tools App*, 2022, vol. 81, pp. 1–26.
9. Xu R., Wunsch D. C. Clustering algorithms in biomedical research: a review. *IEEE Rev Biomed Eng*, 2010, vol. 3, pp. 120–154.
10. Kriegel H. P., Kroger P., Zimek A. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2009, vol. 3, Iss. 1, pp. 1–58.
11. Kailing K., Kriegel H. P., Kroger P. Density-connected subspace clustering for high-dimensional data. *Proceedings of the SDM*, 2004, pp. 246–257.
12. Hu J. Subspace clustering methods for understandable information organization. *Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the School of Computing Science Faculty of Applied Sciences*. Simon Fraser University, 2017.
13. Kaur A., Datta A. A novel algorithm for fast and scalable subspace clustering of high-dimensional data. *Journal of Big Data*, 2015, vol. 2, Iss. 17, 24 p.
14. Charu C. Aggarwal, Han J. Editors. *Frequent Pattern Mining*. Springer Cham, 2014, 490 p.
15. Berchtold S., Bohm C., Keim D., Kriegel H.-P. A cost model for nearest neighbor search in high-dimensional data space. *Proceedings of the 16th Symposium on Principles of Database Systems (PODS)*, 1997, pp. 78–86.
16. Parsons L., Haque E., Liu H. Subspace clustering for high dimensional data: a review. *ACM, Sigkdd explorations newsletter*, 2004, vol. 6, Iss. 1, pp. 90–105.
17. Agrawal R., Gehrke J., Gunopulos D., Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. *Proceedings of the 1998 ACM SIGMOD Conference*, 1998, pp. 94–105.
18. Parsons L., Haque E., Liu H. Subspace clustering for high dimensional data: a review. *SIGKDD Explor Newslett*, 2004, no. 6 (1), pp. 90–105.
19. Assent I., Krieger R., Muller E., Seidl T. Dusc: Dimensionality unbiased subspace clustering. *Proceedings of the IEEE ICDM*, 2007, pp. 409–414.
20. Cheng C.-H., Fu A. W., Zhang Y. Entropy-based subspace clustering for mining numerical data. *Proceedings of the KDD*, 1999, pp. 84–93.
21. Nagesh H., Goil S., Choudhary A. Mafia: Efficient and scalable subspace clustering for very large data sets. *Technical Report 9906-010, Northwestern University*, 1999.

22. Sequeira K., Zaki M. J. Schism: A new approach for interesting subspace mining. *Proceedings of the IEEE ICDM*, 2004, pp. 186–193.
23. Liu G., Li J., Sim K., Wong L. Distance based subspace clustering with flexible dimension partitioning. *Proceedings of the IEEE ICDE*, 2007, pp. 1250–1254.
24. Kelkar B. A., Rodd S. F. Subspace Clustering—A Survey. *Proceedings of ICDMAI*, 2018, vol. 1, pp. 209–220.
25. Liu G., Sim K., Li J., Wong L. Efficient Mining of Distance-Based Subspace Clusters. Wiley Periodicals, Inc. 2009. Published online in Wiley InterScience. Available at: www.interscience.wiley.com (accessed 10.10.2025).
26. Aggarwal C. C., Procopiuc C. M., Wolf J. L., Yu P. S., Park J. S. Fast algorithms for projected clustering. *Proceedings of the 1999 ACM SIGMOD Conference*. Philadelphia, Pennsylvania, USA, 1999, pp. 61–72.
27. Aggarwal C. C., Yu P. S. Finding generalized projected clusters in high dimensional spaces. *Proceedings of the 2000 ACM SIGMOD Conference.* Dallas, Texas, USA, 2000, pp. 70–81.
28. Woo K.-G., Lee J.-H., Kim M.-H., Lee Y.-J. Findit: a fast and intelligent subspace clustering algorithm using dimension voting. *Inf. Soft Tech.*, 2004, no 46 (4), pp. 255–271.
29. Kriegel H.-P., Kroger P., Zimek A. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans Knowl Dis Data Min*, 2009, no 3 (1), pp. 1–58.
30. Guimei L., Kelvin S., Jinyan L., Limsoon W. Distance Based Subspace Clustering with Flexible Dimension Partitioning. *IEEE.* 2007. Available at: https://www.researchgate.net/publication/4250954_Distance_Based_Subspace_Clustering_with_Flexible_Dimension_Partitioning (accessed 10.10.2025).
31. Gan G. Subspace clustering for high dimensional categorical data. *ACM SIGKDD Explorations Newsletter*, 2003, vol. 6, Iss. 2, pp. 87–94.
32. Chang J.-W., Jin D.-S. A new cell-based clustering method for large, high-dimensional data in data mining applications. *Proceedings of the 2002 ACM symposium on Applied computing*, 2002, pp. 503–507.
33. Liu B., Xia Y., Yu P. S. Clustering through decision tree construction. *Proceedings of the 9th CIKM conference*, 2000, pp. 20–29.
34. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd ACM SIGKDD Conference*. Portland, Oregon, USA, 1996, pp. 226–231.
35. Procopiuc C. M., Jones M., Agarwal P. K., Murali T. M. A Monte Carlo algorithm for fast projective clustering *Proceedings of the ACM SIGMOD Intern. conf. on management of data*. Madison, Wisconsin, USA, 2002, pp. 418–427.
36. Zaki M. J. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 2000, no 12 (3), pp. 372–390.
37. Han J., Pei J., Yin Y. Mining frequent patterns without candidate generation. *Proceedings of the ACM SIGMOD*, 2000, pp. 1–12.
38. Cheng C. H., Fu A. W.-C., Zhang Y. Entropy-based subspace clustering for mining numerical data. *Proceedings of the 5th ACM SIGKDD Conference,* 1999, pp. 84–93.
39. Newman D. J., Hettich S., Blake C. L., Merz C. J. UCI repository of machine learning databases. *Department of Information and Computer Science*, University of California, Irvine, 1998. Available at: http//www.ics.uci.edu/~mlearn/MLRepository.html (accessed 10.10.2025).
40. Assent I., Krieger R., Muller E., Seidl T. Dusc: dimensionality unbiased subspace clustering. *Proceedings of the 7th ICDM Conference*. Omaha, Nebraska, USA, 2007, pp. 409–414.
41. Uno T., Kiyomi M., Arimura H. Lcm ver.3 *Proceedings of the ACM OSDM workshop*, 2005, pp. 77–86.
42. Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A. I. Fast Discovery of Association Rules. *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996, chapter 12, pp. 307–328.
43. Aho A., Hopcroft J., Ullman J. *The Design and Analysis of Computer Algorithms*. Addison-Welsle, 1974.
44. Wang H., Wang W., Yang J., Yu P. S. Clustering by pattern similarity in large data sets. *Proceedings of the 2002 ACM SIGMOD Conference*, 2002, pp. 394–405.
45. Pasquier N., Bastide Y., Taouil R., Lakhal L. Discovering frequent closed itemsets for association rules. *Proceedings of the 7th ICDT Conference*. Jerusalem, Israel, 1999, pp. 398–416.
46. Rymon R. Search through systematic set enumeration. *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, Cambridge, Massachusetts, USA, 1992.
47. Bilenko M., Basu S., Mooney R. J. Integrating constraints and metric learning in semi-supervised clustering. *Proceedings of the ICML*, 2004, p. 11.
48. Klein D., Kamvar S. D., Manning C. D. From instance level constraints to space-level constraints: Making the most of prior knowledge in data clustering. *Proceedings of the ICML*, 2002, pp. 307–314.

49. Ruiz C., Spiliopoulou M., Ruiz E. M. C-dbscan: Density-based clustering with constraints. *RSFDGrC*, 2007, vol. 4482, pp. 216–223.
50. Struyf J., Dzeroski S. Clustering trees with instance level constraints. *ECML*, 2007, pp. 359–370.
51. Fromont E., Prado A., Robardet C. Constraint-based Subspace Clustering. Available at: https://www.researchgate.net/publication/29605468_Constraint-based_Subspace_Clustering (accessed 10.10.2025).
52. Bonchi F., Giannotti F., Mazzanti A., Pedreschi D. Adaptive constraint pushing in frequent pattern mining. *Proceedings of the PKDD*, 2003, pp. 47–58.
53. Jeudy B., Boulicaut J.-F. Optimization of association rule mining queries. *Intell. Data Anal*., 2002, no 6 (4), pp. 341–357.
54. Besson J., Robardet C., Boulicaut J.-F., Rome S. Constraint-based concept mining and its application to microarray data analysis. *Intell. Data Anal*, 2005, no 9 (1), pp. 59–82.
55. Gely A. A generic algorithm for generating closed sets of a binary relation. *Proceedings of the ICFCA*, 2005, pp. 223–234.
56. Shannon C. E., Weaver W. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.
57. Kailing K., Kriegel H.-P., Kröger P., Wanka S. Ranking interesting subspaces for clustering high dimensional data. *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Cavtat-Dubrovnik, Croatia, 2003, pp. 241–252.
58. Kriegel H.-P., Kröger P., Sander J., Zimek A. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2011, no 1 (3), pp. 231–240.

Информация об авторах

О. Н. Зуенко — младший научный сотрудник;
О. В. Фридман — кандидат технических наук, ведущий инженер.

Information about the authors

O. N. Zuenko — Junior Researcher;
O. V. Fridman — Candidate of Science (Tech.), Leading Engineer.

Статья поступила в редакцию 01.11.2025; одобрена после рецензирования 24.11.2025; принята к публикации 28.11.2025.
The article was submitted 01.11.2025; approved after reviewing 24.11.2025; accepted for publication 28.11.2025.

Научная статья
УДК 004.832; 004.853
doi:10.37614/2949-1215.2025.16.3.004

ПРИМЕНЕНИЕ RAG-ТЕХНОЛОГИИ ДЛЯ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ТЕСТОВ И ПРОВЕРКИ ЗНАНИЙ С ПОДДЕРЖКОЙ ДИАЛОГА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Алексей Владимирович Шестаков[1]✉, Александр Анатольевич Зуенко[2]
[1, 2]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[1]a.shestakov@ksc.ru✉, https://orcid.org/0000-0002-9052-25791
[2]a.zuenko@ksc.ru, https://orcid.org/0000-0002-7165-6651

Аннотация

Работа посвящена применению RAG-технологий (Retrieval-Augmented Generation) для автоматической генерации тестов и адаптивной проверки знаний с поддержкой диалога на естественном языке. Проведен анализ аналогичных систем поддержки исследовательской деятельности для генерации вопросов. В работе представлена общая архитектура системы, а также детально рассмотрены системы генерации вопросов и адаптивного тестирования, описаны особенности реализации модулей генерации тестовых вопросов и динамического подбора заданий на основе ответов пользователя. Приведены результаты тестирования и сделаны выводы о практической применимости системы.

Ключевые слова:

RAG-технологии, большие языковые модели, интеллектуальный поиск, вопросно-ответная система, обработка естественного языка, онтологии, адаптивное тестирование, научная деятельность

Благодарности:

работа выполнена в рамках научно-исследовательской работы «Методы и информационные технологии мониторинга и управления региональными критическими инфраструктурами Арктической зоны Российской Федерации» (FMEZ-2025-0054).

Для цитирования:

Шестаков А. В., Зуенко А. А., Применение RAG-технологии для автоматической генерации тестов и проверки знаний с поддержкой диалога на естественном языке // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 56–70. doi:10.37614/2949-1215.2025.16.3.004.

Original article

APPLICATION OF RAG TECHNOLOGY FOR AUTOMATED TEST GENERATION AND KNOWLEDGE ASSESSMENT WITH NATURAL LANGUAGE DIALOGUE SUPPORT

Aleksey V. Shestakov[1]✉, Alexander A. Zuenko[2]
[1, 2]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[1]a.shestakov@ksc.ru✉, https://orcid.org/0000-0002-9052-25791
[2]a.zuenko@ksc.ru, https://orcid.org/0000-0002-7165-6651

Abstract

This study explores the application of Retrieval-Augmented Generation (RAG) technology for the automated generation of tests and adaptive knowledge assessment, supported by natural language dialogue. An analysis of similar research support systems for question generation is conducted. The paper presents the overall system architecture and provides a detailed examination of its core components: the question generation subsystem and the adaptive testing engine. The implementation specifics of the test question generation module and the dynamic task selection mechanism based on user responses are described. The results of testing are presented, followed by conclusions regarding the system's practical applicability.

Keywords:

retrieval-augmented generation (rag), large language models (LLMs), intelligent information retrieval, question-answering systems, natural language processing (NLP), ontologies, adaptive testing, research activity

Acknowledgements:

This work was supported by the research project “Methods and Information Technologies for Monitoring and Managing Regional Critical Infrastructures in the Arctic Zone of the Russian Federation” (Project No. FMEZ-2025-0054).

Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 56–70.
Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2025. Vol. 16, No. 3. P. 56–70.

For citation:

Shestakov A. V., Zuenko A. A. Application of RAG technology for automated test generation and knowledge assessment with natural language dialogue support. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 56–70. doi:10.37614/2949-1215.2025.16.3.004.

Введение

Научные сотрудники постоянно имеют дело с поиском и анализом информационных источников. Однако, несмотря на доступность информации, ее объемы быстро растут. Ежедневно выпускается огромное количество публикуемых работ: научных статей, монографий, патентов и технических отчетов. Традиционные методы поиска, такие как поиск по содержимому (полнотекстовый поиск) и по ключевым словам, демонстрируют ограниченность, так как они не учитывают контекст и семантику данных, вследствие чего не способны в значительной мере ускорить и упростить поиск информации. Таким образом, возрастает потребность в контекстно-ориентированных интеллектуальных системах поиска и анализа информации, которые способны понимать семантику информации при интеллектуальном анализе.

Отличительной чертой научных работ является специфика научной информации. При разработке интеллектуальных систем поддержки исследовательской деятельности особое внимание уделяется принципиальной разнородности форм представления информации в научных работах, которая может иметь форму текста или быть сгруппированной в таблицы, описываться математическими формулами, схемами, графиками, изображениями. Кроме этого, научная информация может храниться в разных форматах, таких как PDF, DOCX, XLSX. В процессе разработки интеллектуальной системы поддержки исследовательской деятельности для каждого из форматов необходимо применять свои специальные алгоритмы обработки, чтобы сохранить структуру и смысловые связи текста. При написании научных текстов обычно соблюдаются определенные традиции изложения результатов исследований, тексты характеризуются наличием определенной структуры и стилем изложения материала. Помимо изложенного выше, значение терминов может различаться в зависимости от предметной области, что осложняет разработку интеллектуальных поисковых систем.

К таким интеллектуальным системам поддержки исследовательской деятельности предъявляются строгие требования:

1. Система должна обеспечивать высокую степень релевантности получаемого ответа, учитывая контекст: конкретную предметную область, историю диалога с пользователем и его профиль в системе (студент, аспирант, ведущий научный сотрудник и т. п.).

2. Генерируемые системой ответы должны содержательно соответствовать исходным источникам, а также принятым нормам изложения на русском языке, что подразумевает использование устоявшейся терминологии, точность формулировок и соблюдение логики научного стиля.

3. Так как исследовательская деятельность не ограничивается поиском информации, система должна предоставлять широкий спектр дополнительных возможностей. К ним относится проверка усвоения информации. Для этого предлагается применять адаптивное тестирование пользователей, которое позволяет автоматически составлять тесты на основе загруженных материалов для проверки понимания материала и выявления пробелов в знаниях. Это может быть полезно при решении таких задач, как построение индивидуальной траектории обучения, формирование команды под проект.

4. Система должна обеспечивать поддержку диалога на естественном языке как при формулировке запросов (вопросов), так и при анализе ответов пользователя.

Анализ существующих платформ показывает, что имеющиеся интеллектуальные системы поиска не удовлетворяют всем имеющимся требованиям, особенно в части автоматической генерации тестов для проверки знаний.

В качестве подхода к решению задач интеллектуального поиска и автоматизации проверки знаний можно применять подход на основе больших языковых моделей (БЯМ) [1]. Такие модели способны осуществлять поиск и генерировать ответы на естественном языке, а также могут служить основой для генерации вопросов и проведения диалога. Однако существенным недостатком БЯМ является генерация недостоверных сведений — «галлюцинаций» (т. е. неточных или вводящих в заблуждение

результатов), что является серьезным ограничением при применении БЯМ. Для преодоления этого недостатка предлагается использовать гибридные модели с использованием RAG-технологии (Retrieval Augmented Generation) [2]. Совместное применение этих моделей потенциально способно повысить эффективность исследовательского процесса и уменьшить число рутинных задач. Методы RAG-технологии дополняют ответы БЯМ релевантными фрагментами из базы данных и поддерживают семантический поиск с учетом контекста.

Целью настоящего исследования является разработка системы автоматической генерации тестов и проверки знаний с поддержкой диалога на естественном языке. Система должна осуществлять генерацию наборов вопросов на основе загруженных научных материалов, адаптивное тестирование пользователя на основе сгенерированных тестов. В процессе ответов на вопросы теста система должна обеспечивать автоматическую адаптацию сложности вопросов, анализ ответов пользователя и формирование индивидуальной траектории обучения.

Обзор существующих решений

Интеллектуальные системы поддержки исследовательской деятельности являются быстро развивающейся областью. Далее производится анализ существующих решений для задач семантического поиска и анализа научных текстов.

Международные поисковые системы

Наиболее известными интеллектуальными системами поддержки исследовательской деятельности на международном рынке являются Elicit [3], Semantic Scholar [4], IBM Watson Discovery [5]. Эти системы применяют для работы с научными текстами современные методы обработки естественного языка. Тем не менее при их применении для работы с русскоязычными научными текстами возникают существенные проблемы. Во-первых, рассматриваемые системы имеют низкую эффективность при обработке морфологии русского языка и специфической научной терминологии. Во-вторых, их алгоритмы настроены на структуру и стиль зарубежных публикаций, что приводит к существенным стилистическим ошибкам при анализе русскоязычных публикаций. Также для организаций могут возникать сложности при использовании зарубежных систем из-за требований к конфиденциальности. При этом стоит отметить, что ни одна из перечисленных систем не предлагает полноценных возможностей по автоматической генерации тестов и проведению адаптивного тестирования на основе документов пользователя.

На настоящий момент наиболее перспективным подходом к созданию контекстно-ориентированных систем поиска является применение RAG-технологии. Например, в работе [6] предлагается фреймворк для поиска по научной литературе с применением методов семантической фрагментации и стратегии «abstract-first» для быстрой фильтрации статей. Авторы делают основной акцент на тонкой настройке моделей эмбеддингов (фрагментов текста в векторном представлении) под конкретную предметную область для повышения точности результатов поиска. Другой пример — система LitLLM [7] для автоматизации написания литературных обзоров. Основное преимущество предлагаемой системы заключается в интеграции с внешними API (Semantic Scholar [4], OpenAlex [8]) для поиска публикаций и использовании БЯМ для сортировки результатов и генерации текстов. Однако данная система не поддерживает возможность ведения диалога на естественном языке.

Системы автоматической генерации вопросов

Среди систем автоматической генерации вопросов выделяют несколько подходов. Один из подходов к генерации вопросов представлен в работе [9]. В его основе лежит оптимизированная под задачи генерации вопросов нейросетевая модель, поддерживающая механизм внимания. Для анализа контекста и повышения качества понимания семантических связей авторы работы применяют двунаправленные LSTM-сети (Long short-term memory) [10]. Для формирования вопросов доступно два режима работы: a) с учетом только текущего предложения; b) с привлечением более широкого контекста абзаца.

Для обучения модели авторы работы применяют датасеты SQuAD [11] и MS MARCO [12] на парах «предложение — вопрос» и предобученные эмбеддинги (GloVe) [13] для улучшения качества генерации. Для ситуации, когда модель сталкивается с неизвестными словами, в работе [9] предлагается применять технологию замены токенов UNK (от “unknown” — неизвестный). Таким образом, если система сталкивается с незнакомым модели словом, происходит его замена на термин из исходного текста с наибольшим весом в механизме внимания. Тем не менее при практическом применении система не всегда в полной мере способна обеспечить качество генерации и семантической корректности вопросов, как показано на рис. 1.

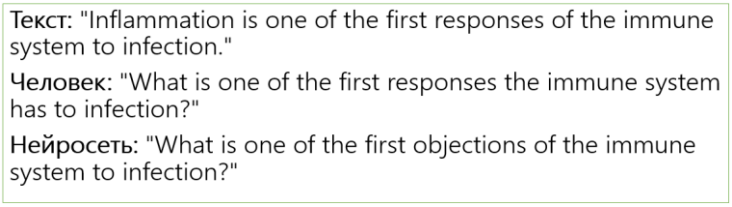


Рис. 1. Пример составления вопроса

В данном примере показано, что система в некоторых случаях генерирует «неправильные» слова, то есть слово “responses” на “objections”.

Таким образом, рассмотренный подход, несмотря на высокую эффективность, требует больших объемов данных и переобучения модели при изменении предметной области, а также иногда система генерирует семантически некорректные слова, а модель с контекстом абзаца не всегда улучшает результаты.

Еще один подход, направленный на применение БЯМ для генерации вопросов, представлен в работе [14]. Авторы работы представляют систему, которая использует дообученные на датасете Turkish-Quiz-Instruct модели GPT-3.5-Turbo и Llama-2 для турецкого языка. Подход ориентирован на генерацию вопросов на основе учебных материалов с предоставлением в качестве ответов набора альтернатив, а также позволяет анализировать краткие ответы на естественном языке. Процесс генерации вопросов включает несколько этапов. На первом этапе происходит сбор и «очистка» учебных материалов, далее выполняется создание промптов для генерации вопросов. После этого происходит использование БЯМ для формирования тестовых вопросов. На последнем этапе система производит оценку качества вопросов с помощью экспертной проверки и метрик ROUGE [15], то есть метрик, которые учитывают, прежде всего, лексическое перекрытие, то есть формальное сходство текстов на уровне слов, а не их смысловую эквивалентность.

В работе предлагается два способа оценки качества вопросов: количественно с применением ROUGE-1, ROUGE-2, ROUGE-L; качественно с использованием экспертных оценок по пятибалльной шкале. По метрикам ROUGE, модель БЯМ GPT-3.5-Turbo обеспечивает более высокую точность генерации, а по результатам экспертной оценки лучше оказывается языковая модель Llama-2-13b-chat-hf.

Таким образом, к преимуществам рассматриваемого подхода относятся: поддержка нескольких форматов вопросов, уменьшение времени на внедрение и требуемой аппаратной производительности по сравнению с подходами, основанными на обучении нейросети за счет применения БЯМ. Использование метрик ROUGE не всегда способно отразить реальное качество вопросов путем оценки поверхностного сравнения сходства без глубокого семантического анализа.

Кратко сформулируем результаты проведенного анализа подходов к генерации тестов с возможностью получения ответов в произвольной форме на естественном языке. Созданные под задачи генерации вопросов нейросетевые модели обеспечивают стабильное качество в пределах своей предметной области, но требуют больших объемов данных и переобучения в случае перехода к другой предметной области. Применение дообученных БЯМ делают систему более адаптивной, однако такие модели зависимы от качества оригинальной модели и «проработки» управляющих промптов.

Существенным недостатком второго подхода является отсутствие интеграции систем генерации вопросов с механизмами тестирования и проверки знаний, а также ориентация на метрики типа ROUGE, которые направленны на оценку поверхностного сходства. Также одной их проблем является

слабая поддержка структурированных данных, представленных в виде таблиц, что критически важно для систем, работающих с научными текстами.

Методология и архитектура системы

Для решения задач интеллектуальной поддержки научной деятельности, которые включают семантический поиск и автоматизированную проверку знаний, предлагается использовать гибридную архитектуру, основанную на применении БЯМ и RAG-технологии.

Методология

Для реализации функций разрабатываемой интеллектуальной системы были выбраны несколько методов: гибридный поиск на основе RAG-технологии; генерация вопросов на основе промпт-инжиниринга; динамическое онтологическое моделирование.

Для реализации базового функционала в рамках настоящего исследования предлагается использовать RAG-технологию, что позволяет применять БЯМ в сочетании с семантическим поиском по базе документов и минимизировать количество «галлюцинаций» у генеративных моделей без необходимости переобучения модели. Применение подобного подхода является важным при работе с научной документацией, поскольку «чистые» языковые модели неспособны предоставить необходимое качество ответов на запросы. Используемый подход способен реализовать генерацию ответов на естественном языке, которые соответствуют стилю речи научного русского языка.

Перейдем к рассмотрению процесса генерации вопросов определенных типов и сложности на основе контекста. БЯМ функционирует с использованием специализированных промптов (специализированных шаблонов для управления поведением языковой модели). Такой подход способен обеспечивать генерацию вопросов приемлемого качества с учетом загруженного контекста. В настоящей статье предлагается алгоритм, который динамически регулирует последовательность и сложность вопросов, которые будут задаваться пользователю. Таким образом, применяемые методы дают системе возможность анализировать ответ пользователя, принимать синонимичные формулировки, а не только сравнивать ответ с заранее заданным строгим шаблоном.

Система не требует переобучения базовой языковой модели, так как и оценка качества ответов, и генерация тестов реализуются преимущественно на основе применения RAG-архитектуры и методов промпт-инжиниринга.

Архитектура системы

Архитектура разработанной системы представлена на рис. 2. Детальное описание компонентов системы и их взаимодействия приведено в работе [16]. Краткое описание основных модулей системы представлено ниже.

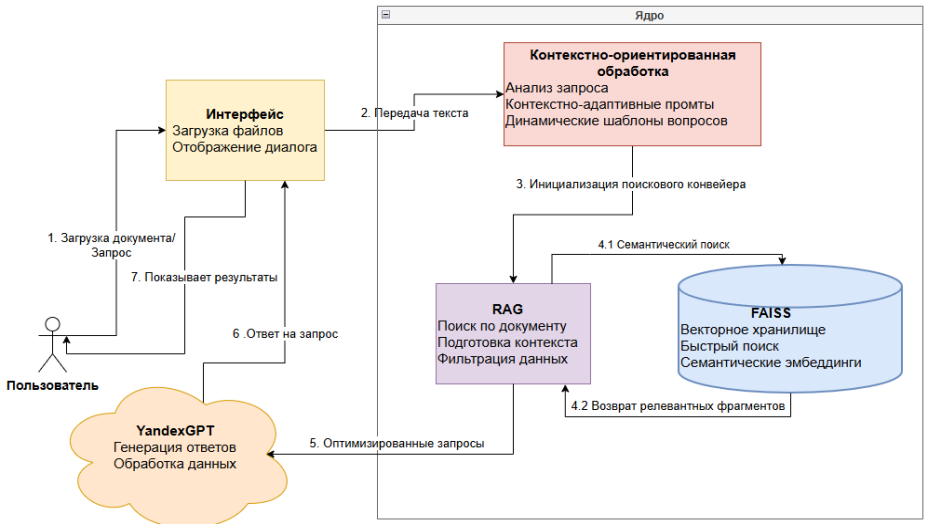


Рис. 2. Архитектура интеллектуальной системы поддержки исследовательской деятельности

Модуль загрузки и обработки документов

В модуле загрузки и обработки документов происходит решение задач, связанных с обработкой различных типов данных и форматов их представления, что влияет на качество всего последующего анализа. Модуль реализует следующий функционал.

Поддержка форматов. Для реализации эффективной обработки и парсинга документов применяются специальные библиотеки и авторские методы. Система поддерживает работу с наиболее распространенными форматами (PDF, DOCX и XLSX), сохраняя структуру и семантические связи исходного текста при обработке документов.

Семантическое фрагментация (чанкирование). При работе с большими объемами частой проблемой являются жесткие ограничения БЯМ по количеству токенов на размер контекста. Стратегия адаптивного семантического чанкирования (деление информации большого объема на небольшие фрагменты) позволяет обойти проблемы, связанные с размером контекста. В отличие от обычного разбиения текста на символы, в процессе фрагментации происходит выделение в тексте логических блоков. В процессе исследований выявлено, что для наиболее эффективного сохранения смысловой целостности информации размер фрагментов должен находиться в диапазоне от 300 до 2000 токенов.

Обработка таблиц. Одним из наиболее часто встречающихся видов представления результатов исследований в научных работах являются таблицы. Для работы с такими структурированными данными авторами реализован механизм обработки таблиц. Система идентифицируют структуру таблицы (заголовки, типы данных, связи между ячейками), а затем производит процесс векторизации, учитывая структурно-семантические связи внутри таблицы. Таким образом, система способна проводить точный семантический поиск по данным внутри таблиц.

Векторное хранилище и RAG-ядро

RAG-технология и векторное хранилище отвечают за семантический поиск и обеспечение приемлемого качества результата.

Для хранения данных используется векторная база данных (ВБД) FAISS (Facebook AI Similarity Search) [17]. После обработки данных на предыдущем этапе получившиеся фрагменты преобразуются в числовые векторные представления (эмбеддинги), которые затем индексируются и помещаются в ВБД. Такой подход позволяет осуществлять поиск фрагментов исходного документа, семантически наиболее схожих с запросом пользователя, в том числе в случаях неполного лексического совпадения.

RAG-технология является центральным звеном при генерации достоверных ответов. В процессе работы технологии можно выделить несколько ключевых шагов: 1) получение запроса пользователя и преобразование в векторное представление; 2) семантический поиск в ВБД наиболее релевантных фрагментов текста (чанков); 3) объединение найденных фрагментов с пользовательским запросом и подходящим к данному типу запроса промптом, а также подача на вход БЯМ; 4) генерация БЯМ связанного ответа на основе представленного контекста. Так как языковая модель опирается исключительно на заданный источник, а не на свои «внутренние знания», этот подход значительно повышает качество ответа за счет снижения «галлюцинаций» модели.

Модуль интеграции с БЯМ (YandexGPT/GigaChat)

В рамках предложенной архитектуры для обеспечения возможности быстрого и простого способа переключения между различными языковыми моделями, как локальными, так и облачными, например российскими YandexGPT [18] и GigaChat [19], используется фреймворк LangChain [20]. Такой подход обладает следующими достоинствами. Во-первых, обеспечивается унифицированный интерфейс для взаимодействия с различными языковыми моделями. Во-вторых, за счет применения методов *промпт-инжениринга* появляется возможность управлять поведением БЯМ, в частности поддерживать стиль ответов, их структуру. Промпт представляет собой специальные четкие инструкции для языковой модели, которые предписывают, как БЯМ должна вести себя при генерации ответов. Наконец, поддерживается возможность хранения истории диалога и учета его при генерации ответов, что дает возможность пользователю задавать уточняющие вопросы, а системе — поддерживать диалог на естественном языке, основываясь на предыдущем взаимодействии.

Блок контекстно-ориентированной обработки информации

Центральным элементом всей разрабатываемой системы является блок контекстно-ориентированной обработки информации (взаимодействия с пользователем). Основная задача блока — классификация запроса пользователя относительно контекста загруженных документов, контекста диалога и профиля пользователя, включая направление запроса в поисковый или тестовый модуль. Блок включает следующие элементы.

Классификатор типа запроса. Выполняет анализ и классификацию входящего запроса пользователя на основании специальных логических правил, определяет его назначение: является ли запросом на поиск информации, командой генерации теста, командой для начала тестирования, ответом на тестовое задание. В случае типовых поисковых запросов система может предоставить заготовленный ответ из базы знаний без применения БЯМ.

Модуль поиска, основанный на RAG-технологии. Осуществляет управление контекстом диалога, которое включает как анализ предыдущей последовательности вопросов-ответов, так и метаинформацию: знания предметной области, контекст документа, профиль и уровень подготовки пользователя. Получившийся из перечисленных элементов расширенный контекст (итоговый промпт) передается на обработку БЯМ.

Модуль адаптивного тестирования. Активируется для типовых запросов пользователя, связанных с операциями по адаптивному тестированию. Его работа подробно описана в следующем подразделе.

Таким образом, блок контекстно-ориентированного поиска является «мозгом» разрабатываемой системы и обеспечивает ведение осмысленного, персонализированного диалога в рамках конкретной предметной области.

Предлагаемый метод генерации вопросов и проверки качества ответов

Генерация тестов

Процесс генерации тестов можно представить в виде UML-диаграммы последовательности на рис. 3.

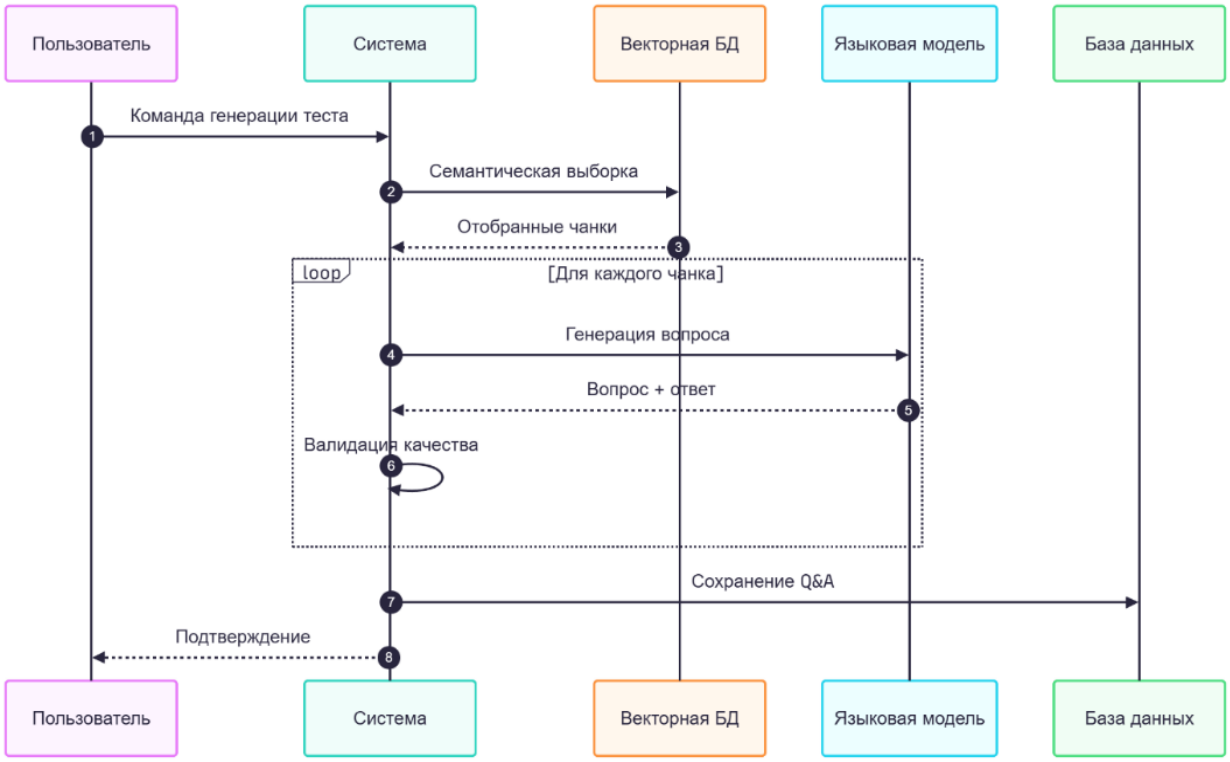


Рис. 3. UML-диаграмма последовательности: генерации тестовой базы

Инициализация. Пользователь дает команду «генерация теста» (или синонимичную команду), после чего система запускает процесс генерации тестовых заданий в соответствии с указанными параметрами (количество вопросов, уровень сложности, тематическая направленность). Происходит обращение к ВБД и семантическая выборка из загруженных материалов фрагментов текста, которые затрагивают разные темы. Алгоритм семантической выборки включает:

1) выделение тем в базе документа, выбранного как контекст для теста;

2) отбор фрагментов таким образом, чтобы обеспечить пропорциональное покрытие тематических разделов;

3) группирование вопросов по сложности;

4) фильтрация фрагментов: исключение вводных фраз и общих рассуждений, а также проверка на наличие релевантных определений и формулировок и т. д.

Критерии отбора семантических фрагментов для генерации по ним вопросов включают: 1) семантическое сходство с ключевой темой (определяется через косинусное сходство между векторными представлениями и должно превышать порог 0,7); 2) степень разнообразия вопросов; 3) семантическая независимость смысловых блоков (например, абзацев).

Создание пары «Вопрос — ответ» (Q&A). Система генерирует заданное число вопросов разных типов и сложности на основе отобранных семантических фрагментов и специализированных промптов. В итоге получается структурированный список «тема — вопрос — эталонный ответ», который сохраняется в БД. Стоит отметить, что при генерации эталонных ответов система основывается исключительно на загруженном контексте.

Процесс формирования вопросов осуществляется с использованием специализированных промптов, которые обеспечивают соответствие заданному стилю и типу вопроса, а также требованиям к сложности вопроса. Пример базового промпта для генерации вопросов показан ниже.

Листинг 1. Пример базового промпта для генерации вопросов

```
Сгенерируй {N} вопросов по следующему контексту, соблюдая требования:
1. **Структура**: раздели вопросы по темам, в каждой теме не менее 3 вопросов
2. **Тип вопросов**: только открытые вопросы, требующие анализа текста
3. **Сложность**: Включи вопросы трех уровней:
   - Базовые (фактологические, 40%)
   - Средние (понимание взаимосвязей, 40%)
   - Сложные (анализ и синтез, 20%)
4. **Ответы**: Полные предложения, содержащиеся в тексте дословно или в близкой формулировке
5. **Формат**: Четкое структурирование по темам с нумерацией
Пример выполнения:
Тема: [Название темы]
1. Вопрос: [формулировка] | Уровень: [базовый/средний/сложный]
   Ответ: [полный ответ]
```

Для градации вопросов по сложности применяются следующие модификаторы промпта:

1) базовый уровень: «Сформулируй вопрос, проверяющий знание определений и основных понятий»;

2) средний уровень: «Создай вопрос, требующий объяснения причинно-следственных связей»;

3) сложный уровень: «Разработай вопрос, требующий проведение анализа преимуществ/недостатков или сравнение концепций».

Процесс валидации качества пары «Вопрос — ответ» включает проверку на соответствие следующим требованиям: 1) наличие точного ответа в исходном тексте; 2) соответствие уровня сложности заявленному; 3) отсутствие дублирования формулировок; 4) соблюдение требуемого формата вывода.

Адаптивное тестирование

На UML-диаграмме последовательности, представленной на рис. 4, показан процесс адаптивного тестирования пользователя.

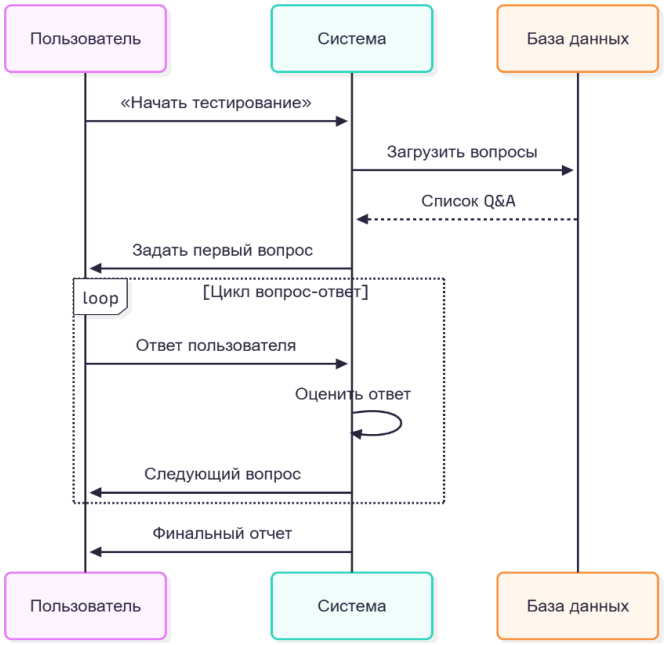


Рис. 4. UML-диаграмма последовательности: процесс тестирования пользователя

Инициализация. При получении запроса «Начать тестирование» система запускает алгоритм тестирования, который включает: 1) загрузку структурированного списка вопросов из тестовой базы данных; 2) инициализацию параметров адаптации тестирования с указанием начального уровня сложности (задается в начале тестирования), целевых показателей обучения и ограничения по количеству вопросов.

Адаптивная логика выбора вопроса. Система задает пользователю первый вопрос, после чего анализирует ответ на корректность и семантическую схожесть с эталонным ответом. Сравнение с эталонным ответом основано на базовом промпте:

Листинг 2. Пример промпта для оценки текущего ответа на вопрос

```
СИСТЕМА: Ты — модуль семантической валидации. Оцени степень соответствия ответа
пользователя эталону.
КОНТЕКСТ:
- ЭТАЛОН: "{правильный_ответ}"
- ОТВЕТ: "{ответ_пользователя}"
ШКАЛА ОЦЕНКИ:
TRUE — полное семантическое соответствие, все ключевые элементы присутствуют
PARTIAL — частичное соответствие, основные понятия сохранены, но есть неточности
FALSE — существенные расхождения или отсутствие ключевых элементов
ПРАВИЛА:
- Допускаются синонимы и перефразирования
- Учитывай контекстную уместность
- Игнорируй стилистические различия
ВЫВОД: Только одно слово — TRUE/PARTIAL/FALSE
```

Динамический подбор следующего вопроса. Далее применяется специальный алгоритм подбора вопросов, с помощью которого выполняется построение индивидуальной траектории тестирования. Общая схема алгоритма динамического подбора вопроса показана на рис 5. Для его работы применяются специализированные правила. Если пользователь дал правильный ответ, то система осуществляет переход к следующей теме тестовой базы, выбранной случайно. Если пользователь дал частично правильный ответ, то ему предлагается дополнительный вопрос аналогичного уровня

сложности, для закрепления материала. В случае, если пользователь дал неправильный ответ, то система подбирает более простой вопрос по этой тематике и предлагает его пользователю.

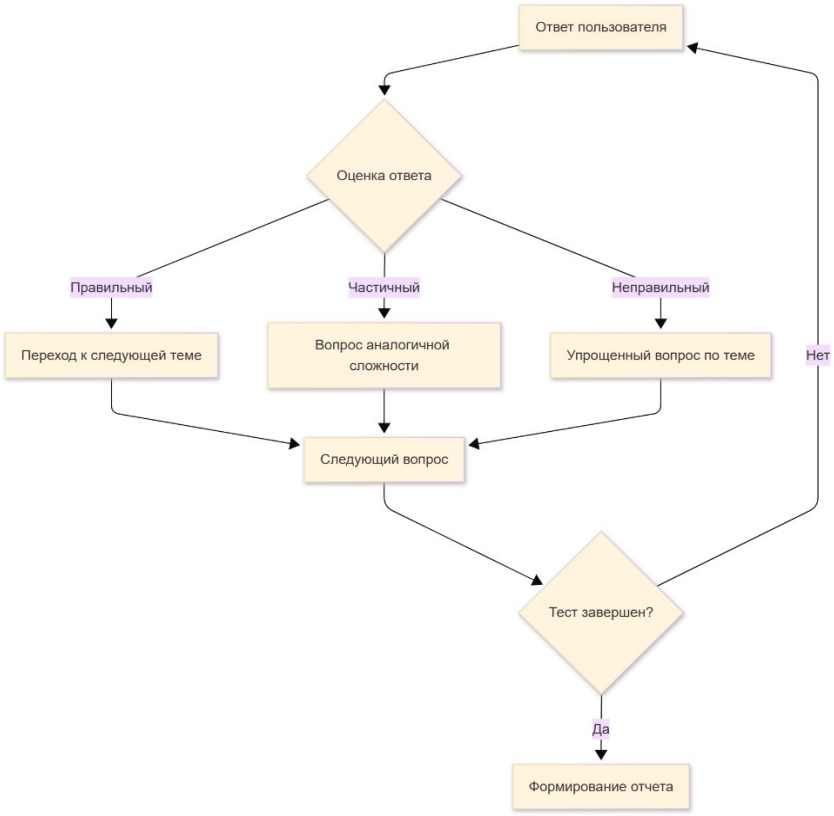


Рис. 5. Схема подбора следующего вопроса

Формирование отчета. После проведения тестирования система проводит анализ статистики по всем ответам пользователя и генерирует персонифицированные рекомендации по «проблемным» темам. Процесс анализа статистики происходит на основе базового промпта:

Листинг 3. Пример базового промпта для формирования итогового отчета

```
Проанализируй результаты тестирования и сформируй итоговый отчет:
КОНТЕКСТ:
- Вопросы и эталонные ответы: {question_answer_pairs}
- Ответы пользователя: {user_answers}
ИНСТРУКЦИИ:
1. Рассчитай общий балл (1 балл за каждый правильный ответ)
2. Определи темы с ошибками пользователя
3. Сформулируй конкретные рекомендации по повторению материала
ФОРМАТ ОТЧЕТА:
ИТОГОВЫЙ БАЛЛ: [X] из [Y] ([Z]%)
АНАЛИЗ ОШИБОК:
[Номер вопроса]. [Тема вопроса]
- Ваш ответ: [ответ пользователя]
- Правильный ответ: [эталонный ответ]
РЕКОМЕНДАЦИИ:
Для улучшения результатов рекомендуется повторить:
- [Тема 1] (ошибки в вопросах: [номера вопросов])
- [Тема 2] (ошибки в вопросах: [номера вопросов])
```

Во время анализа результатов (сравнения ответов пользователя с эталонными формулировками) система производит подсчет общего количества верных/неверных ответов, их процентного соотношения, а также группирует ошибки по тематическим разделам. Затем на основе анализа частоты ошибок формируются рекомендации путем составления перечня тем для повторного изучения, а также составляются рекомендации в привязке к конкретным разделам учебного материала. Пример отчета представлен на рис. 6.

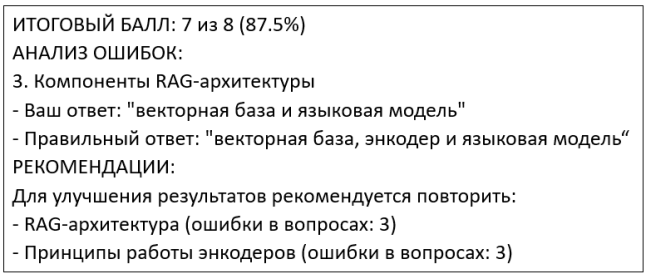


Рис. 6. Пример итогового отчета по прохождению теста

Эксперименты и обсуждение результатов

В ходе работы была реализована интеллектуальная система поддержки исследовательской деятельности. Для оценки эффективности разработанной системы был проведен ряд экспериментов, направленных на проверку функционала по автоматической генерации тестов и адаптивному тестированию.

Оценка модуля автоматической генерации тестов и адаптивного тестирования

Для проверки качества системы по автоматической генерации тестовых вопросов и адаптивному тестированию были проведены тесты на основе учебных материалов по тематике «Задачи удовлетворения ограничений» (CSP) [21].

Качество работы системы оценивалось по следующим аспектам: 1) качество генерации: релевантность, грамматическая корректность и фактическая точность сгенерированных вопросов; 2) корректность работы адаптивного алгоритма: способность системы динамически менять последовательность вопросов на основе проверки ответов.

Для оценки модуля был сгенерирован список вопросов, часть которого представлена в таблице.

На рисунке 7 представлен пример, при котором ответ пользователя система засчитывает как полностью верный. В этом случае происходит переход к другой теме вопросов.

На рисунке 8 продемонстрирован случай, когда пользователь дал частично правильный ответ. Ответ содержит назначение эвристики, но не содержит описание преимуществ от ее применения, заключающееся в минимизации количества исключаемых вариантов для соседних переменных. Система предлагает еще вопрос из той же темы. Если пользователь дал на него правильный ответ, то происходит переход к другой теме тестирования.

Если пользователь дает кардинально неверный ответ, то система подбирает вопрос из той же темы, но более низкой сложности. Далее, если ответ на вопрос правильный, то происходит переход к следующей теме с ранее установленной сложностью. Такой случай показан на рис. 9.

Демонстрация генерации отчета показана на рис. 10. На нем рассмотрен результат прохождения теста по теме «Задачи удовлетворения ограничений», состоящего из 8 вопросов.

В процессе исследований выявлено, что, что модули генерации и подбора вопросов демонстрируют высокую эффективность при проведении тестирования системы. Система корректно анализирует ответы и изменяет последовательность вопросов в соответствии с заданной логикой адаптации. Тем не менее в ряде случаев (7–9 %) у системы наблюдаются ошибки при семантической оценке ответов пользователя и арифметические ошибки при подсчете итогового балла за тест, что говорит о необходимости модернизации механизма подсчета баллов.

Фрагмент из списка сгенерированных вопросов

Тема: Основные понятия CSP		
Вопрос	Уровень	Ответ
Как определяются переменные и ограничения в задаче удовлетворения ограничений?	Базовый	Задача удовлетворения ограничений определена множеством переменных $x_1, x_2, \ldots, x_n$ и множеством ограничений $c_1, c_2, \ldots, c_n$. Каждая переменная x_i имеет непустую область определения D_i возможных значений. Каждое ограничение c_i включает некоторое подмножество переменных и задает допустимые комбинации значений для этого подмножества
Что такое «полное присваивание» и чем оно отличается от «решения» в CSP?	Средний	Полным называется такое присваивание, в котором участвует каждая переменная, а решением задачи CSP является полное присваивание, которое удовлетворяет всем ограничениям
…	…	…
Тема: Алгоритмы поиска для CSP		
Какие два основных подхода к поиску решений в CSP описаны в тексте?	Базовый	Для решения задач CSP широко применяются алгоритмы поиска в глубину (поиск с возвратами), а также могут использоваться методы локального поиска на основе формулировки с полным состоянием
…	…	…
Почему алгоритм с минимальными конфликтами эффективен для решения задачи с N ферзями?	Сложный	Алгоритм с минимальными конфликтами показал себя чрезвычайно эффективным при решении многих задач CSP, особенно при наличии подходящего начального состояния. Больше всего достойно удивления то, что при решении задачи с n ферзями время прогона алгоритма с минимальными конфликтами остается почти независимым от размера задачи. Данный алгоритм решает в среднем за 50 этапов (после начального присваивания) даже задачу с миллионом ферзей

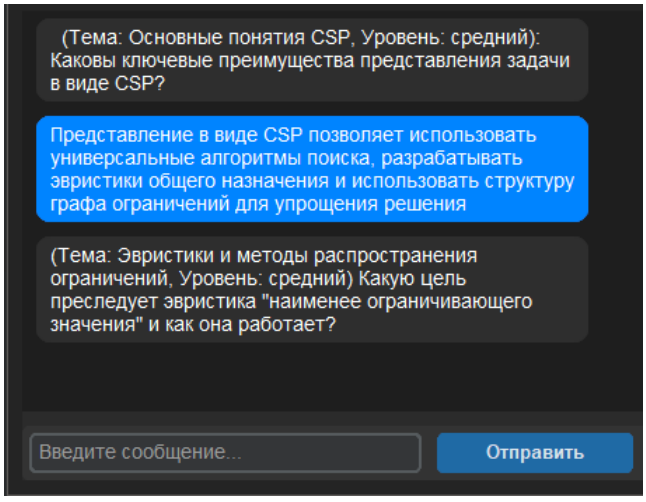


Рис. 7. Пример работы системы при правильном ответе пользователя

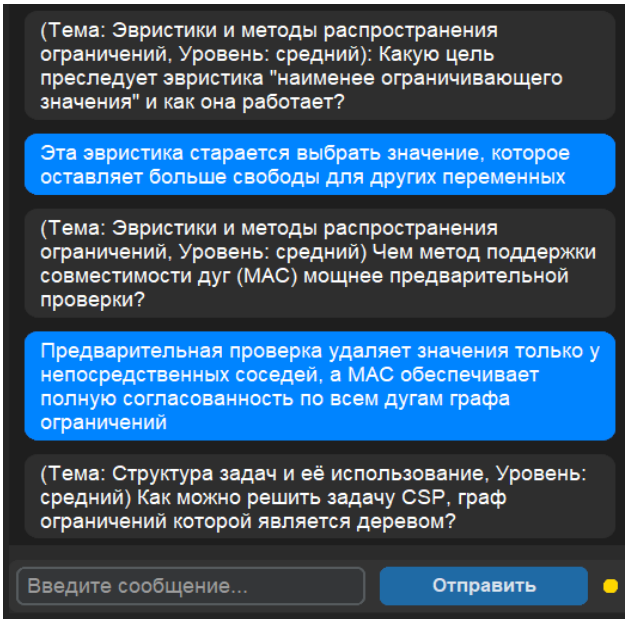


Рис. 8. Пример работы системы при частично правильном ответе пользователя

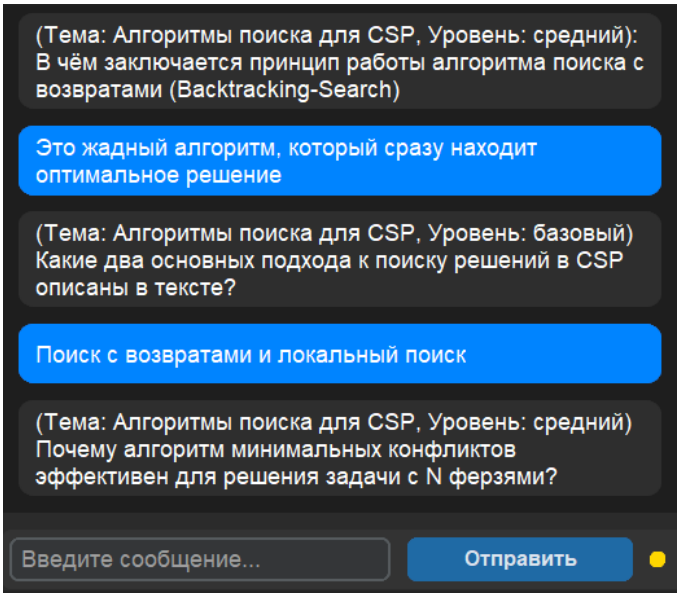


Рис. 9. Пример работы системы при неверном ответе пользователя

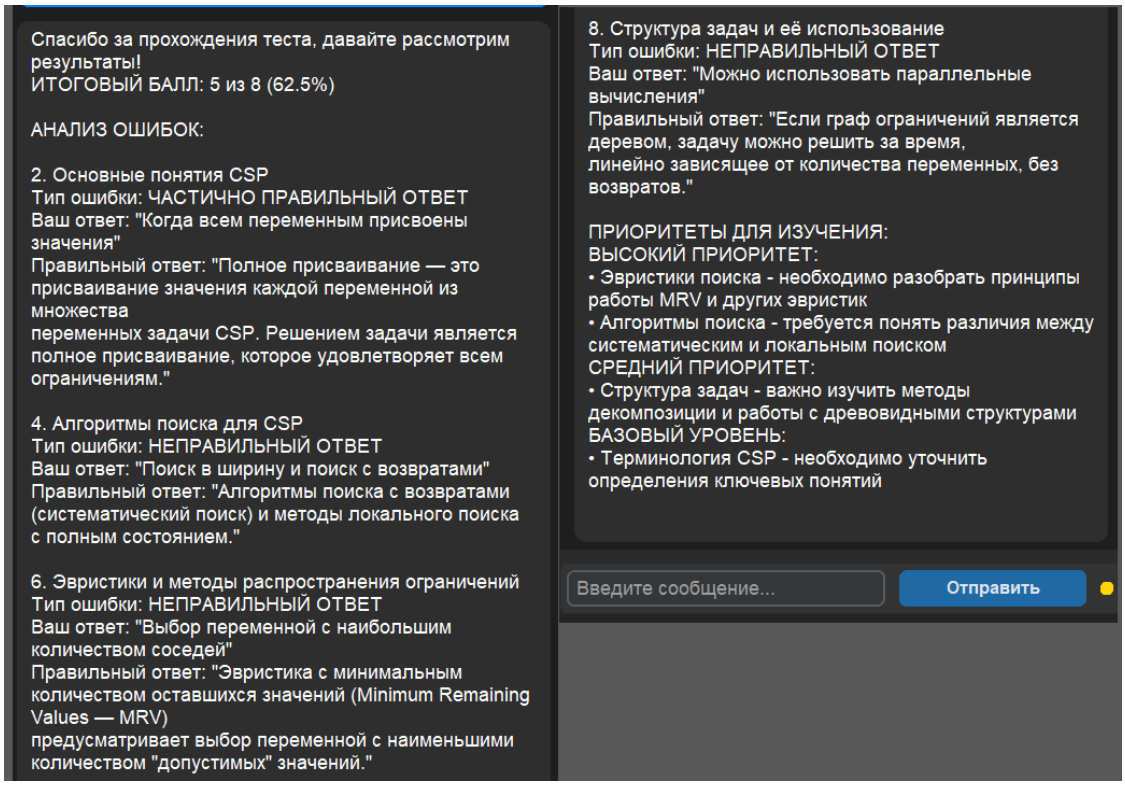


Рис. 10. Пример отчета по результатам прохождения теста

Заключение

В работе представлена система для решения задач автоматизированной генерации тестов на основе указанных пользователем документов и адаптивного тестирования. Подробно рассмотрена ее архитектура, основным компонентом которой является блок контекстно-ориентированной обработки информации. Разработанная программная система зарегистрирована в Роспатенте [22].

В работе приводятся базовые промпты для генерации тестовых вопросов, эталонных ответов, а также для проверки результатов тестирования и формирования отчетов на основании этих результатов. Исследования показали, что уже на текущем этапе система вполне удовлетворительно способна решать большинство стоящих перед ней задач, в частности регулировать сложность вопросов относительно ответов пользователя при формировании индивидуальной траектории обучения и определять уровень квалификации пользователей, что востребовано в исследовательской деятельности. Однако более интеллектуальные механизмы генерации вопросов, а также их адаптивного подбора во время тестирования, по всей видимости, требует привлечения дополнительных инструментов и технологий. На текущий момент учет и анализ контекста выполняются преимущественно на основе RAG-технологии. В дальнейшем для повышения точности обработки информации планируется применять онтологии предметных областей. В качестве еще одного направления исследований можно отметить целесообразность доработки блока арифметического подсчета результатов, улучшения модуля динамического выбора вопросов, а также реализации алгоритмов для работы с формулами.

Список источников

1. Большая языковая модель [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B0%D1%8F_%D1%8F%D0%B7%D1%8B%D0%BA%D0%BE%D0%B2%D0%B0%D1%8F_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8C (дата обращения: 01.10.2025).
2. Gao Y. et al. Retrieval-Augmented Generation for Large Language Models: A Survey // arXiv:2312.10997 [cs]. 2023.
3. Elicit: The AI Research Assistant [Электронный ресурс]. URL: https://support.elicit.com/en/categories/146369 (дата обращения: 01.04.2025).
4. Semantic Scholar API Tutorial [Электронный ресурс]. URL: https://www.semanticscholar.org/product/api/tutorial (дата обращения: 12.09.2025).
5. IBM Watson Discovery [Электронный ресурс]. URL: https://www.ibm.com/products/watson-discovery (дата обращения: 03.09.2025).
6. A Retrieval-Augmented Generation Framework for Academic Literature Navigation in Data Science // arXiv:2412.15404 [cs]. 2024.
7. LitLLM: A Toolkit for Scientific Literature Review // arXiv:2402.01788 [cs]. 2024.
8. OpenAlex: Technical Documentation [Электронный ресурс]. URL: https://docs.openalex.org/ (дата обращения: 10.01.2025).
9. Learning to Ask: Neural Question Generation for Reading Comprehension // arXiv:1705.00106 [cs]. 2017.
10. LSTM-сети долгой краткосрочной памяти [Электронный ресурс]. URL: https://habr.com/ru/companies/wunderfund/articles/331310/ (дата обращения: 10.01.2025).
11. Rajpurkar P. et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2016. P. 2383–2392.
12. Nguyen T. et al. MS MARCO: A Human Generated Machine Reading Comprehension Dataset // arXiv:1611.09268 [cs]. 2016.
13. Pennington J. et al. GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. P. 1532–1543.
14. Mysore S. et al. AClimb: Annotating and Analyzing Clarification in Conversational Search // arXiv:2406.03397 [cs]. 2024.
15. ROUGE (metric) [Электронный ресурс]. URL: https://en.wikipedia.org/wiki/ROUGE_(metric) (дата обращения: 10.01.2025).
16. Олейник А. Г. и др. Об использовании RAG-технологии для исследовательского поиска в справочных и нормативных текстах // Труды Кольского научного центра РАН. Серия: Технические науки. 2024. Т. 15, № 3. С. 5–26.
17. Faiss: A Library for Efficient Similarity Search [Электронный ресурс]. URL: https://faiss.ai/ (дата обращения: 01.11.2024).
18. Yandex Foundation Models [Электронный ресурс]. URL: https://yandex.cloud/ru/docs/foundation-models/ (дата обращения: 20.03.2025).
19. GigaChat API [Электронный ресурс]. URL: https://developers.sber.ru/docs/ru/gigachat/api/overview (дата обращения: 20.07.2025).
20. Langchain Introduction [Электронный ресурс]. URL: https://python.langchain.com/docs/introduction/ (дата обращения: 01.11.2024).
21. Рассел С., Норвиг П. Искусственный интеллект: современный подход. М.: Вильямс, 2021. 1416 с.

22. Свидетельство о государственной регистрации программы для ЭВМ № 2025682172 Российская Федерация. Программа автоматической генерации тестов и проверки знаний в научно-исследовательской деятельности / А. А. Зуенко, А. В. Шестаков; заявл. 31.07.2025; опубл. 21.08.2025.

References

1. Bol'shaya yazykovaya model' [Large language model]. (In Russ.). Available at: https://ru.wikipedia.org/wiki/%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B0%D1%8F_%D1%8F%D0%B7%D1%8B%D0%BA%D0%BE%D0%B2%D0%B0%D1%8F_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8C (accessed 01.10.2025).
2. Gao Y. et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs], 2023.
3. Elicit: The AI Research Assistant. Available at: https://support.elicit.com/en/categories/146369 (accessed 01.04.2025).
4. Semantic Scholar API Tutorial. Available at: https://www.semanticscholar.org/product/api/tutorial (accessed 12.09.2025).
5. IBM Watson Discovery. Available at: https://www.ibm.com/products/watson-discovery (accessed 03.09.2025).
6. A Retrieval-Augmented Generation Framework for Academic Literature Navigation in Data Science. arXiv:2412.15404 [cs], 2024.
7. LitLLM: A Toolkit for Scientific Literature Review. arXiv:2402.01788 [cs], 2024.
8. OpenAlex: Technical Documentation. Available at: https://docs.openalex.org/ (accessed 10.01.2025).
9. Learning to Ask: Neural Question Generation for Reading Comprehension. arXiv:1705.00106 [cs], 2017.
10. LSTM-seti dolgoy kratkosrochnoy pamyati [LSTM-Long Short-Term Memory networks]. (In Russ.). Available at: https://habr.com/ru/companies/wunderfund/articles/331310/ (accessed 10.01.2025).
11. Rajpurkar P. et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 2383–2392.
12. Nguyen T. et al. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. arXiv:1611.09268 [cs], 2016.
13. Pennington J. et al. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
14. Mysore S. et al. AClimb: Annotating and Analyzing Clarification in Conversational Search. arXiv:2406.03397 [cs], 2024.
15. ROUGE (metric). Available at: https://en.wikipedia.org/wiki/ROUGE_(metric) (accessed 10.01.2025).
16. Oleynik A. G. et al. Ob ispol'zovanii RAG-tekhnologii dlya issledovatel'skogo poiska v spravochnyh i normativnyh tekstah [On the use of RAG technology for research search in reference and regulatory texts]. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tehnicheskie nauki* [*Proceedings of the Kola Science Centre of the Russian Academy of Sciences. Series: Engineering Sciences*], 2024, vol. 15, no 3, pp. 5–26. (In Russ.).
17. Faiss: A Library for Efficient Similarity Search. Available at: https://faiss.ai/ (accessed 01.11.2024).
18. Yandex Foundation Models. Available at: https://yandex.cloud/ru/docs/foundation-models/ (accessed 20.03.2025).
19. GigaChat API. Available at: https://developers.sber.ru/docs/ru/gigachat/api/overview (accessed 20.07.2025).
20. Langchain Introduction. Available at: https://python.langchain.com/docs/introduction/ (accessed 01.11.2024).
21. Russell S., Norvig P. *Iskusstvennyy intellekt: sovremennyy podkhod* [Artificial Intelligence: A Modern Approach]. Moscow, Williams Publ., 2021, 1416 p. (In Russ.).
22. Certificate of state registration of the computer program No. 2025682172 Russian Federation. *Programma avtomaticheskoy generatsii testov i proverki znaniy v nauchno-issledovatel'skoy deyatel'nosti* [Program for automatic test generation and knowledge testing in research activities]. A. A. Zuenko, A. V. Shestakov, declared 31.07.2025, published 21.08.2025.

Информация об авторах

А. В. Шестаков — аспирант, стажер-исследователь;
А. А. Зуенко — кандидат технических наук, ведущий научный сотрудник.

Information about the authors

A. V. Shestakov — Graduate Student, Intern Researcher;
A. A. Zuenko — Candidate of Science (Tech.), Leading Scientific Officer.

Статья поступила в редакцию 01.11.2025; одобрена после рецензирования 24.11.2025; принята к публикации 28.11.2025.
The article was submitted 01.11.2025; approved after reviewing 24.11.2025; accepted for publication 28.11.2025.

Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 71–79.
Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2025. Vol. 16, No. 3. P. 71–79.

Научная статья
УДК 004.853
doi:10.37614/2949-1215.2025.16.3.005

ПРЕДИКТИВНОЕ МОДЕЛИРОВАНИЕ СОЦИАЛЬНЫХ РЕАКЦИЙ ДЛЯ РЕГИОНАЛЬНОГО УПРАВЛЕНИЯ НА БАЗЕ МЕТОДОВ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Владимир Витальевич Диковицкий ✉
Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия, dikovitsky@gmail.com✉*, https://orcid.org/0000-0003-0329-9979*

Аннотация

Данная работа представляет модульную архитектуру, использующую современные методы искусственного интеллекта. Подход базируется на четырех компонентах: каузальных агентных моделях для имитации динамики общественных реакций и понимания механизмов формирования мнений; RAG-системах, привязывающих рассуждения к историческим фактам; динамическом графе знаний, связывающем социальные сигналы и события и деонтической логике для обеспечения объяснения решений. Представленная архитектура преодолевает ограничения ретроспективных методов, предоставляя инструмент для понимания причин возникновения трендов, учета нормативной среды и генерации конкретных сценариев действий в реальном времени, что способствует прозрачному использованию методов объяснимого искусственного интеллекта в управлении.

Ключевые слова:

объяснимый искусственный интеллект, социальная сеть, граф знаний

Благодарности:

исследование выполнено в рамках государственного задания ИИММ КНЦ РАН Министерства науки и высшего образования РФ, темы научно-исследовательской работы «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределённости и риска» (шифр темы FMEZ-2025-0053).

Для цитирования:

Диковицкий В. В. Предиктивное моделирование социальных реакций для регионального управления на базе методов объяснимого искусственного интеллекта // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 71–79. doi:10.37614/2949-1215.2025.16.3.005.

Original article

PREDICTIVE MODELING OF SOCIAL REACTIONS FOR REGIONAL MANAGEMENT BASED ON EXPLAINABLE ARTIFICIAL INTELLIGENCE METHODS

Vladimir V. Dikovitsky[1] ✉
Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia, v.dikovitsky@ksc.ru✉*, https://orcid.org/0000-0003-0329-9979*

Abstract

This paper presents a modular architecture utilizing modern artificial intelligence methods. The approach is based on four components: causal agent-based models for simulating the dynamics of public reactions and understanding the mechanisms of opinion formation; RAG systems that anchor reasoning in historical facts; a dynamic knowledge graph linking social signals and events; and deontic logic to provide rationale for decisions. The presented architecture overcomes the limitations of retrospective methods by providing a tool for understanding the causes of trends, taking into account the regulatory environment, and generating specific action scenarios in real time, thereby facilitating the transparent use of explainable artificial intelligence methods in governance.

Keywords:

explainable artificial intelligence, social network, knowledge graph

Acknowledgments:

The study was carried out within the framework of the state assignment of the Institute of Mechanics and Mathematics of the Kola Science Center of the Russian Academy of Sciences of the Ministry of Science and Higher Education of the Russian Federation, research topic “Methods and technologies for creating intelligent information systems to support the development of complex dynamic systems with regional specificity in conditions of uncertainty and risk” (topic code FMEZ-2025-0053).

For citation:
Dikovitsky V. V. Predictive modeling of social reactions for regional management based on explainable artificial intelligence methods. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 71–79. doi:10.37614/2949-1215.2025.16.3.005.

Введение

В условиях динамичной цифровизации общества и повсеместного внедрения искусственного интеллекта (ИИ) в информационные системы, традиционные подходы к мониторингу и прогнозированию общественного мнения, основанные на опросах и статическом анализе данных, сталкиваются с серьезными ограничениями, такими как снижение репрезентативности и невозможность улавливать быстро меняющуюся динамику онлайн-дискурса [1]. Алгоритмы, предназначенные для предсказания социальных реакций на основе анализа социальных сетей, часто вызывают опасения как инструменты контроля, а не как средства для развития гражданского диалога [2]. Некорректное применение методов социального инжиниринга может привести к результатам, противоположным ожидаемым. Системы, стремящиеся к управлению общественными настроениями, рискуют подорвать доверие к институтам [3].

Современные парадигмы ИИ претерпевают значительные изменения, смещая акцент с чистого прогнозирования на поддержку и сопровождение рассуждений [4]. Целью становится не создание систем, которые «думают за человека», а разработка нормативно-ориентированных агентов, которые помогают человеку принимать решения с помощью объяснимого ИИ [5]. Этот переход отражает стремление к более прозрачному, подотчетному и этичному использованию ИИ в управлении, где системы не только предсказывают, но и объясняют свои выводы [6].

Настоящая статья представляет архитектуру Human-in-the-Loop [7], предназначенную для совместного управления, в которой ИИ функционирует не как автономный субъект решения, а как совместный советник. Такой подход позволяет интегрировать человеческое суждение и ценности в процесс принятия решений, используя ИИ для повышения эффективности и этичности [8]. Эта архитектура базируется на четырех взаимосвязанных компонентах:

1) каузальные агентные модели, имитирующие динамику общественных реакций. Эти модели, основанные на взаимодействии автономных агентов, позволяют исследовать возникновение таких явлений, как консенсус и поляризация, путем моделирования причинно-следственных связей [9]. Использование таких моделей позволяет не только прогнозировать, но и понимать механизмы формирования общественного мнения, а также оценивать влияние различных факторов и интервенций [10];

2) RAG-системы (Retrieval-Augmented Generation), привязывающие рассуждения к историческим фактам [11];

3) динамический граф знаний, связывающий социальные сигналы и события [12];

4) деонтическая логика, предоставляющая формальный аппарат для выражения нормативных концепций, что позволяет ИИ-системам не только соблюдать этические нормы, но и объяснять свои решения в терминах дозволенного, обязательного и запрещенного [13]. Это способствует распределению моральной ответственности между человеком и ИИ [14]. Применение деонтической логики в государственном управлении позволяет повысить прозрачность, подотчетность и доверие к автоматизированным процессам и обеспечивает соответствие действий ИИ установленным нормам и ценностям [15].

Представленный подход отличается от предыдущих работ, которые преимущественно сосредоточены на выявлении сложившихся тенденций путем анализа статистики публикаций и различных вариаций тематического моделирования. Эти ретроспективные методы, несмотря на свою техническую состоятельность, позволяют обнаруживать уже произошедшие события, сложившиеся темы и доминирующие сообщения [16]. Они не приближают к пониманию причин возникновения трендов, норм и ценностей, которые их провоцировали, групп, инициировавших дискуссию, и потенциальных внешних вмешательств, способных изменить ее направление. Кроме того, модели ранжирования контента, вовлеченности или тематического моделирования не эффективны для анализа причин и не функционируют в реальном времени. Они не взаимодействуют с нормативной средой, не учитывают юридические ограничения, а также часто игнорируют семантику и контекст, что препятствует генерации конкретных сценариев и действий [17].

Предыдущая работа

В прошлых работах были предложены методы построения предиктивных моделей для определения популярности публикаций и определения нормативных высказываний. В работе [18] проведена оценка возможности современных LLM (Large Language Model) для моделирования трендов в социальных сетях. Исследование показало, что LLM способны эффективно анализировать контент социальных медиа для прогнозирования популярности публикаций, что является шагом к пониманию распространения информации и формированию общественного мнения. В работе [19] исследуется возможность автоматического извлечения компонентов деонтических высказываний. Методология, предложенная в этой работе, обеспечивает преобразование слабоструктурированных правовых документов в формализованные деонтические выражения, что является ключевым для интеграции нормативных знаний в ИИ-системы.

Обзор существующих подходов к моделированию социальных реакций

В условиях постоянно меняющейся цифровой среды, социальные сети трансформировались из простых коммуникационных платформ в общественные арены, оказывающие значительное влияние на коллективные настроения и поведение граждан [20]. В сфере регионального управления это создает как новые возможности для прямого взаимодействия с населением, так и серьезные вызовы, в первую очередь связанные с необходимостью оперативного мониторинга и прогнозирования социальных настроений [18]. Традиционные модели предиктивного анализа социальных сетей, ориентированные на раннее выявление зарождающихся очагов напряжения, оценку эффективности информационных кампаний и предотвращение кризисов [21], часто демонстрируют ограниченный охват, ретроспективный характер анализа и уязвимость результатов к искажениям.

В таких условиях мониторинг не соответствует требованиям управления, которое нуждается в инструментах, способных не только предсказывать всплески негативных высказываний, призывы к действию или массовую поддержку инициатив, но и формировать нормативно обоснованный контекст для вмешательства, а также проводить анализ возможных сценариев. Фрагментарность и несистемность существующих систем мониторинга подчеркивают необходимость перехода к методам прогнозирования и управления, основанным на больших данных и алгоритмических подходах. Ключевым требованием становится не только способность к предсказанию, но и прозрачная объяснимость: система должна аргументировать, почему ожидается та или иная реакция и какие последствия могут иметь различные управленческие решения [22].

Государственные и муниципальные органы активно используют мониторинг социальных сетей для решения широкого круга задач — от обеспечения общественной безопасности и управления кризисами до анализа общественного мнения, выявления угроз и повышения эффективности коммуникации [23]. Например, раннее обнаружение слухов о загрязнении воды через анализ пользовательских сообщений позволяет органам власти оперативно реагировать и предотвращать панику [24]. В период пандемии COVID-19 мониторинг тональности обсуждений помог правительствам учитывать общественную реакцию на локдауны и кампании вакцинации, адаптируя стратегии донесения сообщений [18]. В правоохранительной практике методы предиктивной аналитики применяются для выявления закономерностей преступной активности и более рационального распределения ресурсов, хотя прямые причинно-следственные эффекты от таких вмешательств еще недостаточно обоснованы [25]. В совокупности социальные медиа выступают как ценный источник обратной связи, предоставляющий возможность корректировать политику и повышать качество взаимодействия органов власти с населением в режиме реального времени. Тем не менее для перехода от мониторинга к обоснованному управлению, способному включать причинно-следственные рассуждения и обеспечивать объяснимость решений, необходима интеграция передовых методов ИИ и алгоритмов, которые не просто выявляют тренды, но и интерпретируют их в рамках правовых и этических норм, а также моделируют последствия различных управленческих сценариев.

Большие языковые модели являются современным фундаментом для анализа текстовых сигналов и прогнозирования социальных процессов благодаря своей способности обрабатывать обширные объемы неструктурированного текста и выявлять скрытые взаимосвязи [26]. В основе работы LLM лежит концепция векторизации: слова, фразы, предложения и документы преобразуются

в векторы в многомерном пространстве, где семантически близкие элементы располагаются ближе друг к другу. Ранние прообразы таких подходов, например модели GloVe, продемонстрировали, как контекстуальные представления слов могут отражать семантические связи [27]. С появлением архитектуры трансформеров векторизация приобрела контекстуальный характер: значение слова определяется не его статическим представлением, а зависит от всего предложения или документа, что позволяет учитывать полисемию и тончайшие смысловые оттенки. Размерность таких эмбеддингов существенно возросла и достигла в продвинутых системах тысяч измерений (например, до 12288 в некоторых конфигурациях GPT) [28]. Более того, современные LLM способны векторизовать не только текст, но и код, векторную графику (SVG), числовые последовательности и иные типы данных, расширяя свою сферу применения за пределы классической обработки естественного языка.

Для повышения достоверности и адаптивности моделей широко применяется архитектура Retrieval-Augmented Generation (RAG), в которой генеративный модуль LLM дополняется внешним хранилищем фрагментов знаний в виде векторной базы данных [29]. Это хранилище содержит эмбеддинги релевантных текстов: исторические данные, нормативные документы, фрагменты графов знаний. При поступлении запроса модель сначала осуществляет поиск релевантных фрагментов, затем интегрирует их в контекст и генерирует ответ с учетом как исходного запроса, так и дополнительной информации. Такой подход снижает необходимость переобучения модели и уменьшает риск «галлюцинаций», повышая достоверность прогнозов. Однако в динамичных предметных областях, таких как анализ социальных реакций, этот метод требует адаптации, поскольку его прямое применение может привести к одновременному добавлению противоречивых и/или утративших актуальность фрагментов данных.

Система предиктивного моделирования социальных реакций Crowdsearch

Система представляет собой модульный фреймворк, объединяющий обработку больших данных, современные методы представления текста и графов, каузальное агентное моделирование и формализацию нормативных требований. Система ориентирована на поддержку органов регионального и муниципального управления путем предоставления инструментов для проактивного понимания общественных настроений, проведения контрфактического анализа последствий управленческих решений и оптимизации коммуникационных стратегий. Для достижения этих целей архитектура системы должна обеспечить надежное, масштабируемое и объяснимое представление как содержательных, так и структурных аспектов коммуникативного пространства. Это достигается сочетанием контекстуальных текстовых эмбеддингов, графовых эмбеддингов, динамического графа знаний и каузального модуля (рис. 1).

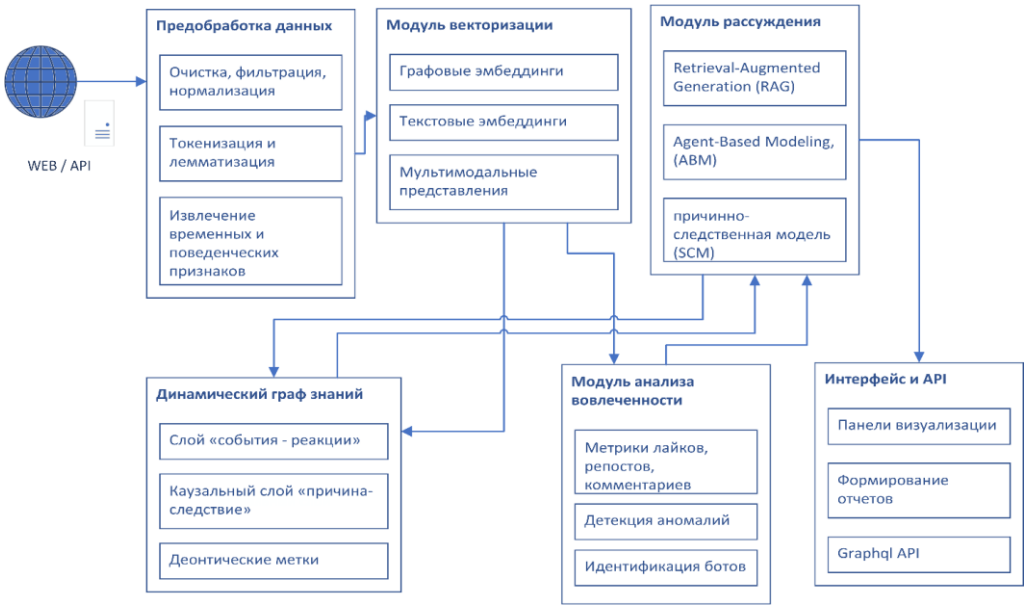


Рис. 1. Компоненты системы предиктивного моделирования социальных реакций Crowdsearch

Для учета нормативного контекста применяется технология обнаружения нормативных высказываний как компонентов деонтической логики, описанная в работе [19]. Деонтическая логика — раздел модальной логики, формализующий такие понятия, как обязательство, разрешение и запрет, служит инструментом формализации нормативных ожиданий и ограничений [13]. Ее операторы $O(\phi)$ («обязательно, чтобы ϕ»), $P(\phi)$ («разрешено, что ϕ») и $F(\phi)$ («запрещено, что ϕ») позволяют задавать нормы вроде «обязательно соблюдать правила платформы». Деонтическая логика применяется для анализа политических высказываний, оценки их нормативной нагрузки и возможного воздействия на аудиторию [30]. В юридической логике она используется для проверки корректности аргументации и соответствия вывода нормативной базе. Однако реальная среда, в которой действуют пользователи, сочетает как формализованные, так и неформализованные законы, правила платформ, общественные ожидания и т. п. Поэтому применена гибридная модель: деонтическая логика интегрирована с графом знаний, что позволяет формально моделировать конфликт норм и контекстуальные исключения. Технически это реализуется путем добавления в граф знаний слоя с деонтическими метками. Начальной базой для извлечения норм и правил служат правила платформы. Система может проверять соответствие прогнозируемого сообщения существующим нормам, генерировать рекомендации с учетом нормативных ограничений и обосновывать свои решения. Для анализа динамики норм и их внедрения в общественное пространство применяется алгоритмическая модель распространения: агенты (пользователи) в сети получают бинарные состояния $s_u \in \{0,1\}$, где 1 — соблюдает норму, 0 — нет. Распространение регулируется пороговыми механизмами: агент принимает норму, если доля соседей, ее соблюдающих, превышает индивидуальный порог θ_u. Также допускается затухание нормы с вероятностью p_{drop}. При каждом временном шаге измеряются проникновение нормы $P_{total}(t)$, скорость ее распространения и образование кластеров сопротивления. Модель учитывает неоднородность порогов, влияние лидеров мнений и структуру сети. Алгоритмическое моделирование позволяет моделировать динамику норм, выявлять точки сопротивления или узлы-распространители, что важно при проектировании коммуникации.

Модуль векторизации и семантического представления выполняет три параллельные функции: генерация эмбеддингов для постов и комментариев, построение графовых эмбеддингов для сетевых структур и формирование мультимодальных представлений при наличии изображений или других медиа. Для текстовой части применяется предобученная трансформер-модель RuBERT, адаптированная под русский язык [31]. Для векторизации графа знаний используется метод представлений узлов GraphSAGE — метод агрегации соседних признаков, позволяющий строить эмбеддинги для новых (ранее невидимых) узлов в динамически растущих графах. GraphSAGE позволяет обучать генератор эмбеддингов (функцию агрегации), что принципиально для социально генерируемого контента с постоянно добавляющимися узлами и источниками [32]. Практическая реализация этих компонентов эффективно поддерживается библиотекой PyTorch Geometric, предоставляющей инструменты для обработки больших графов и механизмы масштабирования на мульти-GPU серверах [33].

Динамический граф знаний с каузальным слоем является ядром архитектуры и содержит события, представленные в виде кортежей:

$$E = (P, T_{pub}, K, G_{rep}, \Phi),$$

где P — текст поста-триггера; T_{pub} — время публикации; K – набор тем; G_{rep} — множество групп-репостеров (с порогом $| G_{rep} | \geq 2$ для сигнала межгруппового охвата); Ф — деонтическая метка. Реакция A формализуется через вектор метрик активности $M(P, \Delta t) = (L, R, C)$ и сравнение с базовой статистикой аналогичных постов: фиксируется всплеск, если для какого-либо типа метрик выполняется $M_{type} > \bar{M}_{100,type} + \alpha\sigma_{100,type}$ (порог α настраиваем). Тональность реакции оценивается агрегированием классификаций сентимента отдельных комментариев $S(c_j)$ и последующей интерпретацией суммарной оценки $S_{overall}$. Одно из ключевых отличий архитектуры — интеграция графовых эмбеддингов и каузального слоя с Retrieval-Augmented Generation (RAG) для обеспечения объяснимости и опоры на источники. В рабочих сценариях LLM не генерирует рекомендации «из головы»: перед генерацией RAG-модуль извлекает из векторного индекса релевантные фрагменты (исторические кейсы, части графа знаний), которые затем становятся контекстом для генеративного шага. Такой подход улучшает

фактическую точность и обеспечивает возможность указывать на источники, снижая риск «галлюцинаций». Однако использование RAG без адаптации индекса влечет проблему множественности контекстов, когда в векторном хранилище содержатся релевантные, но противоречивые и/или утратившие актуальность фрагменты данных. Для решения этой проблемы применяется процедура отбора и ранжирования фрагментов данных с учетом версионирования индекса [34].

Для причинно-следственного вывода использованы методы из теории каузальности: модельные подходы и графы причинно-следственных связей, позволившие формализовать интервенции [35]. Практически это реализуется через сочетание причинно-следственных графов с агентно-ориентированными моделями. Граф причинно-следственных связей осуществляет идентифицируемое представление события и его зависимостей, агентно-ориентированная модель позволяет симулировать механизмы распространения и эмпирически проверять гипотезы о поведении агентов в сетевом окружении.

Компонент обнаружения аномалий сочетает статистическую детекцию всплесков, признаки сетевой активности и набор моделей классификации на основе анализа сетевых метрик и контента. Для оценки достоверности предложенных интервенций последние дополняются разделом «обоснование» в виде ссылок на извлеченные фрагменты RAG-хранилища и «дерево причин» каузального графа.

Источниками данных являются публичные группы и страницы социальной сети «ВКонтакте» за длительный период (2020 — н. в.), что позволяет учитывать как фоновую активность, так и кризисные всплески. Для оперативной аналитики система использует инкрементальное извлечение метрик с временными метками и шагом обновления каждые 20 минут; хранится история метрик для каждого поста, что служит базой для детекции аномалий и расчета нормализованных порогов всплесков.

Заключение

Предложенный архитектурный фреймворк сочетает проверенные алгоритмические блоки (GraphSAGE, RAG, LLM, каузальный аппарат и агентно-ориентированные модели) и опирается на современные библиотечные экосистемы (PyTorch Geometric, Milvus для векторных индексов). Такой набор обеспечивает практическую реализуемость, масштабируемость и объяснимость решений, необходимых для нормативно-ориентированного рассуждения и контрфактического анализа в региональном управлении. Ведутся работы по экспериментальной проверке эффективности отдельных предложенных модулей. Для задач дообучения LLM, построения динамических графов и верификации деонтических и каузальных предсказаний значительная часть текстов (публикации и комментарии) аннотируются в полуавтоматическом режиме. Процесс аннотации включает: категоризацию социальных реакций (негативные комментарии, призывы к действию, поддержку инициатив, репосты), оценку тональности реакций, аннотацию деонтических характеристик. Для создания обучающего набора каузальных связей и управленческих воздействий промаркированы пары «вмешательство — реакция» и «событие — реакция» путем учета наличия официальной реакции, например «всплеск негативных комментариев — разъяснение уполномоченного представителя». Эти данные являются основой (обучающей выборкой) для каузального слоя графа знаний и для валидации контрфактических сценариев.

Список источников

1. Kim R. M., Veselovsky V., Anderson A. Capturing dynamics in online public discourse: A case study of universal basic income discussions on reddit // Proceedings of the International AAAI Conference on Web and Social Media. 2025. Vol. 19. P. 1021–1037.
2. Bennett C. J., Lyon D. Data-driven elections: implications and challenges for democratic societies // Internet Policy Review. 2019. 8 (4). doi: 10.14763/2019.4.1433.
3. Карпова И. В., Воронкова И. Е., Бирюкова Е. А. PR-технологии как средство манипулирования политической активностью молодёжи // Реклама и PR в координатах социума, бизнеса и медиапространства. 2020. С. 18–73.
4. Revel M., Pénigaud T. AI-Enhanced Deliberative Democracy and the Future of the Collective Will. 2025.
5. Argyle L. P. et al. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale // Proceedings of the National Academy of Sciences. 2023. Vol. 120, No. 41. P. e2311627120.
6. Evkoski B., Pollak S. XAI in Computational Linguistics: Understanding Political Leanings in the Slovenian Parliament //arXiv preprint arXiv:2305.04631. 2023.

7. Mosqueira-Rey E. et al. Human-in-the-loop machine learning: a state of the art // Artificial Intelligence Review. 2023. Vol. 56, No. 4. P. 3005–3054.
8. Chen X. Ethical Governance of AI: An Integrated Approach via Human-in-the-Loop Machine Learning // Comput. Sci. Math. Forum. 2023. 8. 29. https://doi.org/10.3390/cmsf2023008029.
9. Antosz P., Szczepanska T., Bouman L., Gareth Polhill J., Jager W. Sensemaking of causality in agent-based models // International Journal of Social Research Methodology. 2022. 25:4. P. 557–567. doi: 10.1080/13645579.2022.2049510.
10. Arnold K. F. et al. DAG-informed regression modelling, agent-based modelling and microsimulation modelling: a critical comparison of methods for causal inference // International journal of epidemiology. 2019. Vol. 48, No. 1. P. 243–253.
11. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks // Advances in neural information processing systems. 2020.Vol. 33. P. 9459–9474.
12. Yang R., Salim F. D., Xue H. Sstkg: Simple spatio-temporal knowledge graph for intepretable and versatile dynamic information embedding // Proceedings of the ACM Web Conference 2024. 2024. P. 551–559.
13. Benzmüller C., Parent X., van der Torre L. Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework // Methodology, and Tool Support. 2020.
14. Rao S. et al. Deontic Temporal Logic for Formal Verification of AI Ethics //arXiv preprint arXiv:2501.05765. 2025.
15. Priya T. V., Rao S. Deontic Temporal Logic for Formal Verification of AI Ethics //arXiv preprint arXiv:2501.05765. 2025
16. Tong, Yangfan, Sun, Wei. Multimedia Network Public Opinion Supervision Prediction Algorithm Based on Big Data // Complexity. 2020. 6623108, 11 p. https://doi.org/10.1155/2020/6623108.
17. Yukhno A. Digital transformation: Exploring big data governance in public administration // Public Organization Review. 2024. Vol. 24, No. 1. P. 335–349.
18. Shishaev M., Dikovitsky V. Predicting the Popularity of Social Network Publications Based on Content Analysis Using the Transformer Language Model // International Scientific and Practical Conference Digital and Information Technologies in Economics and Management. Cham: Springer Nature Switzerland, 2023. P. 180–191.
19. Dikovitsky V. V., Shishaev M. G. Automated extraction of deontological statements through a multilevel analysis of legal acts // Proceedings of the Computational Methods in Systems and Software. Cham: Springer International Publishing, 2018. P. 102–110.
20. Yang G. R. et al. A method of predicting and managing public opinion on social media: An agent-based simulation // Information Sciences. 2024. Vol. 674. P. 120722.
21. Li H. et al. Detecting early-warning signals for social emergencies by temporal network sociomarkers // Information Sciences. 2023. Vol. 627. P. 189–204.
22. Wang X. et al. Impact of social participation types on depression in the elderly in China: an analysis based on counterfactual causal inference // Frontiers in Public Health. 2022. Vol. 10. P. 792765.
23. Akopova T. S., Tikhonova A. V. Citizens' activity in social networks: factors of influence. 2020. Креативная экономика, 18 (12), 3805-3826. doi: 10.18334/ce.18.12.122155
24. Xiong J., Hswen Y., Naslund J. A. Digital surveillance for monitoring environmental health threats: a case study capturing public opinion from Twitter about the 2019 Chennai water crisis // International journal of environmental research and public health. 2020. Vol. 17, No. 14. P. 5077.
25. Lee Y., Bradford B., Posch K. The effectiveness of big data-driven predictive policing: Systematic review // Justice Evaluation Journal. 2024. Vol. 7, No. 2. P. 127–160.
26. Rusnachenko N. L. Language Models Application in Sentiment Attitude Extraction Task // Trudy ISP RAN/Proc. ISP RAS. 2021. Vol. 33, issue 3. P. 199–222. doi: 10.15514/ISPRAS-2021-33(3)-14.
27. Shishaev M., Dikovitsky V., Pimeshkov V., Kuprikov N., Kuprikov M., Shkodyrev V. Extracting relations from texts using vector language models and a neural network classifier // PeerJ Computer Science. 2023. 9. e1636.
28. Wang Z. et al. History, development, and principles of large language models: an introductory survey // AI and Ethics. 2025. Vol. 5, No. 3. P. 1955–1971.
29. Huwiler D., Stockinger K., Fürst J. VersionRAG: Version-Aware Retrieval-Augmented Generation for Evolving Documents //arXiv preprint arXiv:2510.08109. 2025.
30. Сытько А. В. Говорящий как субъект деонтики в политической речи (на материале немецкого и русского языков) // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2019. Т. 10, № 4. С. 1003–1020. doi: 10.22363/2313-2299-2019-10-4-1003-1020.
31. RAS F.ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian // Proceedings of the International Conference “Dialogue 2025”. 2025.
32. Huang K., Chen C. Subgraph generation applied in GraphSAGE deal with imbalanced node classification // Soft Computing. 2024. Vol. 28, No. 17. P. 10727–10740.
33. Fey M., Lenssen J. E. Fast graph representation learning with PyTorch Geometric //arXiv preprint arXiv:1903.02428. 2019.

34. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks // Advances in neural information processing systems. 2020. Vol. 33. P. 9459–9474.
35. Pearl J. Causality. 2nd ed. Cambridge University Press, 2009.

References

1. Kim R. M., Veselovsky V., Anderson A. Capturing dynamics in online public discourse: A case study of universal basic income discussions on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 2025, Vol. 19, pp. 1021–1037.
2. Bennett C. J., Lyon D. Data-driven elections: implications and challenges for democratic societies. *Internet Policy Review*, 2019, 8 (4). doi: 10.14763/2019.4.1433.
3. Karpova I. V., Voronkova I. E., Biryukova E. A. PR technologies as a means of manipulating the political activity of young people. *Advertising and PR in the coordinates of society, business, and media space*, 2020, pp. 18–73. (In Russ.).
4. Revel M., Pénigaud T. *AI-Enhanced Deliberative Democracy and the Future of the Collective Will*, 2025.
5. Argyle, Lisa P et al. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences of the United States of America*, 2023, vol. 120, 41, p. e2311627120. doi:10.1073/pnas.2311627120.
6. Evkoski B., Pollak S. XAI in computational linguistics: understanding political leanings in the Slovenian Parliament. 2023. Available at: https://arxiv.org/pdf/2305.04631 (accessed 20.11.2025).
7. Mosqueira-Rey E. et al. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 2023, 56 (4), pp. 3005–3054.
8. Chen X. Ethical governance of AI: an integrated approach via human-in-the-loop machine learning. *Comput. Sci. Math. Forum*, 2023, 8, 29. https://doi.org/10.3390/cmsf2023008029.
9. Antosz P., Szczepanska T., Bouman L., Polhill J. G., Jager W. Sensemaking of causality in agent-based models. *International Journal of Social Research Methodology*, 2022, 25 (4), pp. 557–567.
10. Arnold K. F. et al. DAG-informed regression modeling, agent-based modeling and microsimulation modeling: a critical comparison of methods for causal inference. *International Journal of Epidemiology*, 2019, 48 (1), pp. 243–253.
11. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 2020, Vol. 33, pp. 9459–9474.
12. Yang R., Salim F.D., Xue H. Sstkg: simple spatial-temporal knowledge graph for interpretable and versatile dynamic information embedding. *Proceedings of the ACM Web Conference 2024*, 2024, pp. 551–559.
13. Benzmüller C., Parent X., van der Torre L. Designing normative theories for ethical and legal reasoning: LogiKEy framework. *Methodology and Tool Support*, 2020.
14. Rao S. et al. Deontic Temporal Logic for Formal Verification of AI Ethics. 2025.
15. Priya T. V., Rao S. Deontic Temporal Logic for Formal Verification of AI Ethics. 2025. Available at: https://arxiv.org/pdf/2501.05765 (accessed 19.11.2025).
16. Tong Y., Sun W. Multimedia network public opinion supervision prediction algorithm based on big data. *Complexity*, 2020, 6623108. doi: 10.1155/2020/6623108.
17. Yukhno A. Digital transformation: exploring big data governance in public administration. *Public Organization Review*, 2024, 24 (1), pp. 335–349.
18. Shishaev M., Dikovitsky V. Predicting the popularity of social network publications based on content analysis using the transformer language model. *Digital and Information Technologies in Economics and Management*. Cham, Springer, 2023, pp. 180–191.
19. Dikovitsky V. V., Shishaev M. G. Automated extraction of deontological statements through multilevel analysis of legal acts. *Computational Methods in Systems and Software*. Cham, Springer, 2018, pp. 102–110. (In Russ.)
20. Yang G.R. et al. A method of predicting and managing public opinion on social media: an agent-based simulation. *Information Sciences*, 2024, 674. doi: 10.1016/j.ins.2024.120722
21. Li H. et al. Detecting early-warning signals for social emergencies by temporal network sociomarkers. *Information Sciences*, 2023, 627, pp. 189–204.
22. Wang X. et al. Impact of social participation types on depression in the elderly in China: an analysis based on counterfactual causal inference. *Frontiers in Public Health*, 2022, 10, 792765. doi:10.3389/fpubh.2022.792765.
23. Akopova T. S., Tikhonova A. V. Citizens' activity in social networks: influence factors, 2020. Creative Economy, 18 (12), 3805–3826. (In Russ.). doi: 10.18334/ce.18.12.122155.

24. Xiong J., Hswen Y., Naslund J. A. Digital surveillance for monitoring environmental health threats: a case study capturing public opinion from Twitter about the 2019 Chennai water crisis. *International Journal of Environmental Research and Public Health*, 2020, 17 (14), 5077. doi: 10.3390/ijerph17145077.
25. Lee Y., Bradford B., Posch K. The effectiveness of big data-driven predictive policing: systematic review. *Justice Evaluation Journal*, 2024, 7 (2), pp. 127–160.
26. Rusnachenko N. L. Language models application in sentiment attitude extraction task. *Proc. ISP RAS*, 2021, 33 (3), pp. 199–222.
27. Shishaev M., Dikovitsky V., Pimeshkov V., Kuprikov N., Kuprikov M., Shkodyrev V. Extracting relations from texts using vector language models and a neural network classifier. *PeerJ Computer Science*, 2023, 9, e1636.
28. Wang Z. et al. History, development, and principles of large language models: an introductory survey. *AI and Ethics*, 2025, 5 (3), pp. 1955–1971.
29. Huwiler D., Stockinger K., Fürst J. VersionRAG: Version-aware retrieval-augmented generation for evolving documents.2025. Available at: https://arxiv.org/pdf/2510.08109 (accessed 25.10.2025).
30. Sytko A. V. The speaker as a subject of deontics in political speech (based on the German and Russian languages). *RUDN Journal of Language Theory. Semiotics. Semantics,* 2019, 10 (4), pp. 1003–1020.
31. RAS F. ruSciFact: open benchmark for verifying scientific facts in Russian. *Proceedings of the International Conference "Dialogue"*, 2025.
32. Huang K., Chen C. Subgraph generation applied in GraphSAGE to deal with imbalanced node classification. *Soft Computing*, 2024, 28 (17), pp. 10727–10740.
33. Fey M., Lenssen J.E. Fast graph representation learning with PyTorch Geometric. Available at: https://arxiv.org/pdf/1903.02428 (accessed 20.10.2025).
34. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 2020, Vol. 33, pp. 9459–9474.
35. Pearl J. *Causality*. Cambridge, Cambridge University Press, 2009.

Информация об авторе

В. В. Диковицкий — кандидат технических наук, старший научный сотрудник.

Information about the author

V. V. Dikovitsky — Candidate of Science (Tech.), Senior Research Fellow.

Статья поступила в редакцию 23.10.2025; одобрена после рецензирования 30.10.2025; принята к публикации 03.11.2025.
The article was submitted 23.10.2025; approved after reviewing 30.10.2025; accepted for publication 03.11.2025.

Научная статья
УДК 004.853
doi:10.37614/2949-1215.2025.16.3.006

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ИЗВЛЕЧЕНИЯ ДАННЫХ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Роман Александрович Горбунов[1], Александр Владимирович Вицентий[2]✉
[1, 2]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[2]Филиал Мурманского арктического университета в г. Апатиты, Апатиты, Россия
[1]gorbunov-roma@inbox.ru, https://orcid.org/0009-0004-4627-504X
[2]alx_2003@mail.ru✉, https://orcid.org/0000-0003-1331-4749

Аннотация

Данная работа посвящена исследованию возможностей больших языковых моделей (LLM) для решения задач извлечения структурированных данных в формате RDF-троек из неструктурированных разнородных текстов на естественном языке. Рассматривается проблема эффективности извлечения данных, которая актуальна для автоматического построения семантических сетей, служащих основой для представления геопространственных знаний. Представлена сравнительная оценка различных типов промптинга, являющихся ключевым инструментом взаимодействия с LLM.

Ключевые слова:

большая языковая модель (LLM), граф знаний, DeepSeek, RDF-тройки, промптинг, извлечение структурированных данных, семантическая сеть

Благодарности:

исследование выполнено в рамках государственного задания Института информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук от Министерства науки и высшего образования Российской Федерации, тема научно-исследовательской работы «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределенности и риска» (регистрационный номер 1023032300374-0-2.2.1).

Для цитирования:

Горбунов Р. А., Вицентий А. В. Исследование возможностей больших языковых моделей для извлечения данных из текстов на естественном языке // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 80–105. doi:10.37614/2949-1215.2025.16.3.006.

Original article

RESEARCH OF THE CAPABILITIES OF LARGE LANGUAGE MODELS FOR EXTRACTING DATA FROM NATURAL LANGUAGE TEXTS

Roman A. Gorbunov[1], Alexander V. Vicentiy[2]✉
[1, 2]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[2]Apatity branch of Murmansk Arctic State University, Apatity, Russia
[1]gorbunov-roma@inbox.ru, https://orcid.org/0009-0004-4627-504X
[2]alx_2003@mail.ru✉, https://orcid.org/0000-0003-1331-4749

Abstract

This paper is devoted to the study of the capabilities of large language models (LLM) for solving the problems of extracting structured data in the RDF-triples format from unstructured heterogeneous texts in natural language. The problem of data extraction efficiency is considered, which is relevant for the automatic construction of semantic networks that serve as the basis for the representation of geospatial knowledge. A comparative assessment of various types of prompting, which are a key tool for interacting with LLM, is presented.

Keywords:

large language model (LLM), knowledge graph, DeepSeek, RDF-triples, prompting, structured data extraction, semantic network

Acknowledgments:

The study was carried out within the framework of the Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences state assignment of the Ministry of Science and Higher Education of the Russian Federation, research topic “Methods and technologies for creating intelligent information systems to support the development of complex dynamic systems with regional specifics in conditions of uncertainty and risk” (registration number of the research topic 1023032300374-0-2.2.1).

For citation:

Gorbunov R. A., Vicentiy A. V. Research of the capabilities of large language models for extracting data from natural language texts. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 80–105. doi:10.37614/2949-1215.2025.16.3.006.

Введение

Извлечение данных из неструктурированных текстов на естественном языке является одной из ключевых задач в области их обработки и одним из важных этапов для решения прикладных проблем, связанных с управлением информацией [1]. В контексте разработки технологии синтеза адаптивных геосемантических изображений на основе геопространственных знаний эта задача приобретает особую значимость [2]. Извлечение сущностей и отношений между ними актуально для автоматического построения и обогащения онтологий — спецификации репрезентативного словаря для общей области дискурса, включающей определения классов, отношений, функций и других объектов [3], которые служат формальной основой для представления геопространственных знаний. Особый интерес представляют нехудожественные тексты (новостные сводки, официальные отчеты, сообщения из обсуждений в социальных сетях). Подобные тексты, как правило, содержат в качестве основы описание некоторого события или ситуации через множество связанных между собой фактов. Например, рассматриваемые в рамках данного исследования тексты о чрезвычайных ситуациях техногенного характера включают в себя информацию о географических объектах и связях между ними в контексте произошедших событий. Получаемые при извлечении структурированные данные, такие как RDF-тройки <субъект, отношение, объект> (Resource Description Framework, RDF) [4], представляют собой элементарную семантическую единицу, формализованное высказывание для построения графовых моделей данных, где субъект и объект определены в качестве узлов графа, а отношение — в качестве связующего элемента. Подобные структуры данных позволяют разрабатываемым программным решениям «понимать» семантические связи, имеющие место между геопространственными объектами в рамках конкретной предметной области, что является фундаментом для последующего семантического моделирования и синтеза геоизображений.

Автоматическое извлечение отношений из текстов на естественном языке также рассматривается в качестве одной из возможных задач в области обработки естественного языка (Natural Language Processing, NLP). Практическая значимость подобной задачи обусловлена ее ролью в решении множества прикладных проблем. Так, в медицине извлечение временных связей из клинических текстов является инструментом доступа к обширному источнику медицинской информации, который описывает статус пациента, включая клинические заключения, процедуры и информацию о проводимом лечении [5]. В финансовом секторе извлечение связей из финансовых отчетов используется для предоставления инвесторам информации о компании, на основе которой они могут принимать соответствующие решения об инвестировании [6]. В сфере безопасности извлечение связей из текстовых записей о чрезвычайных ситуациях в работе городского железнодорожного транспорта обеспечивает эффективную справочную информацию для принятия решений [7]. В качестве основных типовых задач извлечения данных можно обозначить несколько примеров. Распознавание именованных сущностей (Named Entity Recognition, NER) — идентификация и классификация определенных категорий слов в тексте, например имен людей, названий организаций, даты и времени, числовых и денежных выражений, где одной из специализированных вариаций является извлечение топонимов (наименований географических объектов). Извлечение отношений (Relation Extraction, RE) — определение и извлечение семантических связей между выделенными сущностями. Определение семантических ролей (Semantic Role Labeling, SRL) — процесс определения роли и отношения каждого слова в контексте предложения. Исторически

для решения подобных задач применяется широкий спектр методов, эволюция которых демонстрирует переход от жестких правил к нейросетевому обучению. К таким методам можно отнести, например, методы на основе словарей и правил, которые, несмотря на высокую точность, требуют большого объема ручного труда экспертов и плохо масштабируются; статистические методы, где используются размеченные корпуса текстов для автоматического сбора данных, но которые часто создают проблемы с точностью из-за зашумленности данных; нейросетевые модели, в том числе предобученные, которые стали современным стандартом, демонстрируют впечатляющие результаты по извлечению данных за счет способности учитывать контекст. Такой современный подход часто используется для решения задач извлечения сущностей и отношений [8; 9], поэтому он может отлично подойти для решения задачи извлечения множества RDF-троек из текстов на естественном языке, представленной в данном исследовании.

Высокоперспективным направлением в решении задач извлечения данных из текстов на естественном языке является применение больших языковых моделей (Large Language Model, LLM). Они представляют собой предобученные на значительных объемах информации нейросетевые инструменты для обработки запросов с возможностью выявления зависимостей между словами в последовательности. Предварительное обучение позволяет им не только генерировать связанный текст, но и решать описанные ранее сложные семантические задачи. Эволюция LLM — от ранних концептов до современных высокоразвитых систем — привела к появлению множества мощных моделей, доступных через API (Application Programming Interface, API) или чат-интерфейсы. Среди популярных примеров можно выделить ChatGPT (OpenAI) [10], Claude (Anthropic) [11], Gemini (Google) [12]. Активно развиваются и отечественные разработки, включая GigaChat (Сбер) [13] и YandexGPT (Яндекс) [14]. В подобных моделях ключевым элементом взаимодействия является промпт (англ. prompt — подсказка), а под промптингом (promting) понимается методология составления текстовых инструкций (промптов), направляющих модель при решении конкретных задач.

Поскольку LLM по своей сути являются универсальными предсказателями слов в последовательности, то именно промпт фокусирует их вычислительные возможности на требуемой операции, такой как извлечение RDF-троек. При этом эффективность извлечения данных напрямую зависит от выбранной стратегии промптинга, которая помогает экспертам ответить на вопрос «Какой способ подачи информации модели является наиболее эффективным для решения поставленной задачи?», и типа промптинга для ответа на вопрос «В какой форме представить модели инструкцию для эффективного решения поставленной задачи?». Таким образом, стратегия определяет методику подачи информации, а тип — структуру инструкции.

В работе рассматривается проблема оценки влияния выбора стратегии промптинга на эффективность извлечения структурированных данных из текстов на естественном языке с помощью применения больших языковых моделей. При этом ключевой гипотезой исследования является предположение о значительном влиянии выбранных стратегий и типов промптинга на эффективность извлечения структурированных данных из нехудожественных разнородных текстов. В настоящее время LLM демонстрируют революционные возможности для решения задач обработки данных, выступая мощным инструментом для их извлечения. Их применение для обработки разнородных текстовых данных является не только практическим инструментом, но и объектом исследования, поскольку эффективность работы LLM сильно зависит от методики взаимодействия пользователя с конкретной реализацией такой модели при использовании структурированных промптов. Существует множество различных классификаций и подходов к конструированию промптов, предложенных различными научными группами [15–17], однако, несмотря на это, нет единого подхода к стандартизации типов промптов, а также общей точки зрения на их эффективность в задачах извлечения данных из текстов на естественном языке.

В научной сфере представлены различные примеры того, где LLM используются для извлечения данных, включая извлечение данных для получения информации о вредителях в сельском хозяйстве [18], извлечение структурированных табличных данных из текстовых медицинских отчетов [19], а также извлечение структурированных данных для проведения химических исследований [20]. Существующие решения на основе LLM показывают, что, несмотря на их высокий потенциал, результаты извлечения данных могут значительно варьироваться в зависимости

от формулировки промпта, используемой модели, формата и вида текста, а также специфики предметной области. На основе представленных сведений можно сделать вывод о том, что существующие типы промптинга требуют систематизированного анализа для последующего эффективного и точного извлечения RDF-троек для решения задач обработки разнородных текстов на естественном языке, что является основным стимулом для проведения настоящего исследования. Основной целью данной работы является исследование и систематизация имеющихся стратегий и типов промптинга для понимания их эффективности при извлечении RDF-троек из разнородных текстовых данных в рамках глобальной темы по формированию геопространственных знаний.

Материалы и методы

В рамках данной работы проводится исследование типов промптов, относящихся к различным стратегиям (техникам) промптинга для проверки их эффективности при извлечении RDF-троек из разнородных текстов на естественном языке. Методика включает в себя следующие этапы: отбор исходного текстового экземпляра, который является основой для выполняемого языковой моделью задания; проектирование структуры задания по преобразованию текста в набор структурированных RDF-троек; систематизация техник и типов промптинга и последующее составление универсальных шаблонов для взаимодействия с языковой моделью; сборка и отправка заданий языковой модели по преобразованию текста в набор RDF-троек по заготовленным шаблонам; выборка наиболее эффективных стратегий промптинга.

Для обеспечения репрезентативности результатов выбран текст о чрезвычайной ситуации техногенного характера, повествующий о железнодорожной аварии на станции Княжая, содержащий описание некоторого множества географических объектов. Текст обладает характерными для подобного вида повествования признаками, такими как хронологическая последовательность изложения, наличие конкретных географических ориентиров, упоминание участников события и описание динамических взаимодействий между объектами. Хронология событий четко выражена от первоначальной остановки грузового состава на уклоне до последующего самопроизвольного движения и столкновения с пассажирским поездом. Географические ориентиры представлены многоуровневой системой локализации от станции Княжая до Мурманской области, что демонстрирует иерархическую организованность географического пространства. Участники события включают как физических лиц (машинист и помощник), так и технические объекты (грузовой состав, пассажирский поезд), между которыми описываются различные виды взаимодействий через систему глагольных конструкций: «укатился с перегона», «врезался в хвост», «повредив третий вагон». Текст также содержит разнообразные типы отношений между различными категориями сущностей, таких как пространственные («на станции Княжая», «недалеко от станции»), временные («18 декабря 2024 года», «спустя полтора часа»), причинно-следственные («вместо 33 башмаков — лишь пять» — «воздух ушёл» — «состав начал двигаться» — «столкновение»).

Текст, описывающий событие железнодорожного происшествия, который был использован в качестве экземпляра для анализа эффективности извлечения RDF-троек различными типами промптов, представлен ниже.

Использованный текст: *«Столкновение поездов на станции Княжая — железнодорожная авария, произошедшая 18 декабря 2024 года на станции Княжая в посёлке Зеленоборский Кандалакшского района Мурманской области России. Грузовой состав № 2013 задним ходом самопроизвольно укатился с перегона, в результате чего произошло боковое столкновение с хвостом стоящего на станции пассажирского поезда № 11 Мурманск — Санкт-Петербург. Погибли 2 человека, пострадал 31. По предварительным данным, грузовой поезд недалеко от станции Княжая остановился на 17-тысячном уклоне, но вместо положенных 33 тормозных башмаков его зафиксировали лишь пятью. В результате спустя полтора часа воздух из тормозной системы ушел, и состав начал самопроизвольно двигаться вниз, набирая скорость. Попытки остановить его оказались безуспешными, и грузовой поезд врезался в хвост пассажирского состава пассажирского поезда № 11 «Санкт Петербург — Мурманск», повредив третий с конца вагон. В результате столкновения погибли два человека, более тридцати пострадали (из них 5 детей; тяжёлые ранения получили четверо). Следствие задержало двух сотрудников — 45-летнего машиниста*

Алексея Зычкова и его 23-летнего помощника Рамиля Садыгова. Во время допроса они признались, что закрепили состав минимальным количеством башмаков и ожидали помощи, не учитывая риск, связанный с уклоном пути. По факту трагедии возбуждено уголовное дело по статье 263 УК РФ — «Нарушение правил безопасности движения и эксплуатации железнодорожного транспорта, повлекшее гибель людей». Пока единственным подозреваемым остаётся машинист грузового состава. До устранения последствий аварии ожидались задержки в движении трёх пассажирских поездов.» [21].

Одним из важных шагов исследования является проектирование структуры унифицированного задания по преобразованию неструктурированного текста в набор RDF-троек. На этом этапе разработана четкая, краткая формулировка задания без детализированных правил извлечения, обеспечивающая стандартизацию взаимодействия с языковой моделью. Разработанная формулировка задания выглядит следующим образом: *«Проведите разбиение заданного текста на триплеты. После разбиения проведите подсчёт количества полученных триплетов. Ответ на данное задание предоставьте в следующем формате: «Триплеты: [субъект, отношение, объект], …; Количество триплетов: ».*

Такой подход выбран по нескольким причинам. Во-первых, это позволило оценить способность языковой модели к самостоятельному выявлению семантических связей без внешних подсказок о том, какие именно типы отношений считаются релевантными. Во-вторых, такой подход способствовал существенному уменьшению количества используемых токенов при отправке запросов и получении ответа от языковой модели, что повысило эффективность использования ресурсов для взаимодействия с ней. Необходимо отметить тот факт, что краткость формулировки позволила более детально исследовать те виды промптов, где подразумевается личное рассуждение самой языковой модели, поскольку некоторые инструкции не навязывали готовых шаблонов анализа и оставляли пространство для проявления ее собственных аналитических возможностей. Данный подход выявил готовность языковой модели к работе в условиях неполной спецификации задачи и подтвердил ее способность восполнять недостающую информацию за счет собственных языковых знаний. Стандарт построения RDF-троек <субъект, отношение, объект> является семантически прозрачным, но при этом достаточно абстрактным, что создает необходимые условия для креативности языковой модели в интерпретации текста при сохранении базовой структурной согласованности. Требование подсчета количества RDF-троек послужило дополнительным критерием контроля выполнения заданий и позволило количественно оценить эффективность различных стратегий промптинга.

Для проведения экспериментальной части исследования была выбрана большая языковая модель DeepSeek-V3.2 [22]. Выбор был обусловлен комплексом практических и методологических соображений. Ключевыми факторами стали доступность и экономическая целесообразность использования. Модель DeepSeek предоставляет доступ к чату с достаточно широкими лимитами использования, что делает ее доступной для научных исследований без необходимости привлечения дополнительного финансирования. Это особенно важно при тестировании типов промптов, для которых требуется выполнение нескольких последовательных запросов к модели, требующих повышенного количества токенов на ответ. С технической точки зрения модель основана на современной трансформерной архитектуре и демонстрирует конкурентные показатели в различных бенчмарках [23–25], что обеспечивает релевантность получаемых результатов при решении задач по извлечению структурированных данных.

Предварительные тесты подтвердили, что данная большая языковая модель может эффективно справляться с задачами по структурированию данных и извлечению RDF-троек из текстов на русском языке, проявляя адекватную производительность при обработке нехудожественных разнородных текстов с включением геопространственной семантики. Важно подчеркнуть, что в современных условиях существует множество альтернативных языковых моделей, а также отечественных разработок, о которых упоминалось ранее. Такие модели потенциально также могли бы быть использованы в данном исследовании, однако принципиальный выбор конкретной реализации модели не является критически важным для достижения основной цели работы, поскольку фокус исследования направлен в сторону методологии промптинга, а не сравнительного анализа архитектурных особенностей или возможностей различных моделей. В контексте данного исследования LLM рассматривается прежде всего как инструмент-посредник для реализации различных стратегий и типов промптинга. Такой подход позволяет абстрагироваться от специфических особенностей конкретных моделей и сфокусироваться на универсальных принципах конструирования промптов.

Процесс взаимодействия с языковой моделью DeepSeek был организован через веб-интерфейс чата, доступного на официальном сайте разработчика. Для каждого экспериментального прогона использовался новый отдельный сеанс чата, что исключало влияние предыдущего контекста на результаты работы модели. Важно отметить, что дополнительные опции платформы, такие как «глубокое мышление», которое позволяет нейросети решать задачи поэтапно с подробным описанием мыслительного процесса, и «поиск в интернете», который позволяет нейросети исследовать актуальные интернет-ресурсы, не применялись при проведении экспериментальных прогонов для обеспечения чистоты исследования и исключения факторов, которые могли бы исказить оценку эффективности типов промптинга. Каждый экспериментальный прогон выполнялся в идентичных условиях, что обеспечило сопоставимость данных, полученных при тестировании различных типов промптинга.

Структура отправляемого запроса языковой модели (рис. 1) представляет собой целостный объект, состоящий из нескольких блоков. Поставленная задача включает в себя текст задания и анализируемый моделью экземпляр текста о происшествии на станции Княжая. Запрос, кроме задачи, включает в себя также модификатор, образованный от типа промптинга. Сформированный экземпляр запроса вводится в чат модели единым сообщением, где инструкции модификатора могут располагаться в начале или в конце запроса или (в редких случаях) могут быть разделены на две части, где одна часть располагается в начале, а другая в конце запроса.

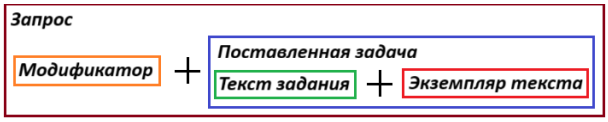


Рис. 1. Общая структура отправляемого запроса языковой модели

Далее представлен пример типичного процесса взаимодействия с большой языковой моделью, наглядно демонстрирующий полный цикл обработки запроса. На рис. 2 показан пример пользовательского промпта, введенного в интерфейсе чата DeepSeek, где был применен модификатор типа Zero-Shot Chain-of-Thought, о чем явно свидетельствует инструкция модели решить задачу шаг за шагом. На рис. 3 зафиксирован процесс генерации ответа, где интерфейс чата отображает последовательную цепочку логических рассуждений модели, выполняемую в соответствии с полученной инструкцией. На рис. 4 представлен финальный этап, где DeepSeek дает окончательный ответ, который был сгенерирован в строгом соответствии с заданным форматом. В дальнейшем сгенерированный ответ подвергался экспертному анализу.

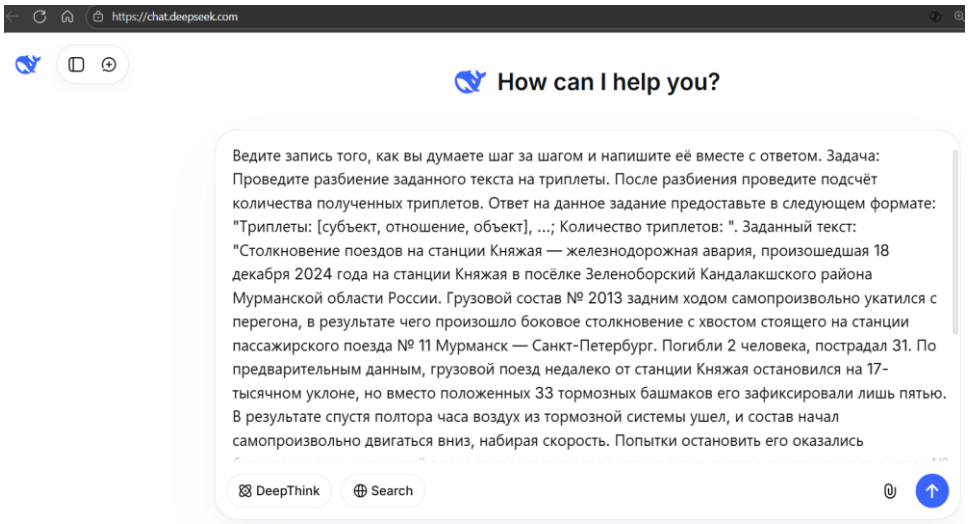


Рис. 2. Пример промпта в интерфейсе чата DeepSeek с запросом Zero-Shot Chain-of-Thought

Хорошо, давайте разберём задачу шаг за шагом.

Шаг 1: Понимание задачи

Нужно извлечь из текста триплеты в формате [субъект, отношение, объект], а затем посчитать их количество.

Триплет — это факт, где есть субъект (кто/что), отношение (что делает/как связано) и объект (с кем/с чем/куда).

Шаг 2: Выделение основных фактов из текста

1. Столкновение поездов — железнодорожная авария.
2. Авария произошла 18 декабря 2024 года.
3. Место: станция Княжая (посёлок Зеленоборский, Кандалакшский район, Мурманская область, Россия).
4. Грузовой состав № 2013 укатился задним ходом.
5. Столкновение боковое с пассажирским поездом № 11 (Мурманск — Санкт-Петербург).
6. Погибли 2 человека.
7. Пострадал 31 человек.
8. Грузовой поезд остановился на 17-тысячном уклоне.
9. Вместо 33 тормозных башмаков использовали 5.
10. Через 1,5 часа воздух из тормозной системы ушёл.
11. Состав начал двигаться вниз, набирая скорость.
12. Попытки остановить — безуспешны.
13. Грузовой поезд врезался в хвост пассажирского.
14. Повреждён третий с конца вагон.
15. Погибли два человека.
16. Более 30 пострадали (5 детей, 4 тяжёлых).
17. Следствие задержало двух сотрудников: машиниста Алексея Зычкова (45 лет) и помощника Рамиля Садыгова (23 года).
18. Они признались, что закрепили состав минимальным количеством башмаков и ожидали помощи, не учитывая риск.
19. Возбуждено уголовное дело по статье 263 УК РФ.
20. Подозреваемый — машинист грузового состава.
21. До устранения последствий — задержки движения трёх пассажирских поездов.

Шаг 3: Форматирование в триплеты

[Столкновение поездов на станции Княжая, является, железнодорожная авария]
[Столкновение поездов на станции Княжая, произошло, 18 декабря 2024 года]
[Столкновение поездов на станции Княжая, место, станция Княжая]
[Станция Княжая, находится в, посёлок Зеленоборский]
[Посёлок Зеленоборский, находится в, Кандалакшский район]
[Кандалакшский район, находится в, Мурманская область]
[Мурманская область, находится в, Россия]
[Грузовой состав № 2013, укатился задним ходом, с перегона]
[Грузовой состав № 2013, столкнулся с, пассажирский поезд № 11]
[Пассажирский поезд № 11, маршрут, Мурманск — Санкт-Петербург]
[Авария, погибшие, 2 человека]
[Авария, пострадавшие, 31 человек]
[Грузовой поезд, остановился на, 17-тысячный уклон]
[Грузовой поезд, должен был быть зафиксирован, 33 тормозных башмака]
[Грузовой поезд, зафиксирован, 5 тормозных башмаков]
[Через 1,5 часа, воздух из тормозной системы, ушёл]
[Грузовой состав, начал двигаться, вниз]
[Грузовой состав, набирал, скорость]
[Попытки остановить грузовой состав, результат, безуспешны]
[Грузовой поезд, врезался в, хвост пассажирского поезда]
[Грузовой поезд, повредил, третий с конца вагон]
[Следствие, задержало, машинист Алексей Зычков]
[Следствие, задержало, помощник Рамиль Садыгов]
[Алексей Зычков, возраст, 45 лет]
[Рамиль Садыгов, возраст, 23 года]
[Они, признались, закрепили состав минимальным количеством башмаков]
[Они, признались, ожидали помощи]
[Они, не учитывали, риск уклона пути]
[По факту трагедии, возбуждено, уголовное дело]
[Уголовное дело, статья, 263 УК РФ]
[Статья 263 УК РФ, название, Нарушение правил безопасности движения и эксплуатации железнодорожного транспорта, повлекшее гибель людей]
[Подозреваемый, машинист грузового состава]
[До устранения последствий, задержки в движении, трёх пассажирских поездов]

Шаг 4: Подсчёт триплетов

Выше получилось 32 триплета.

Рис. 3. Интерфейс чата DeepSeek с рассуждением модели во время ответа на запрос

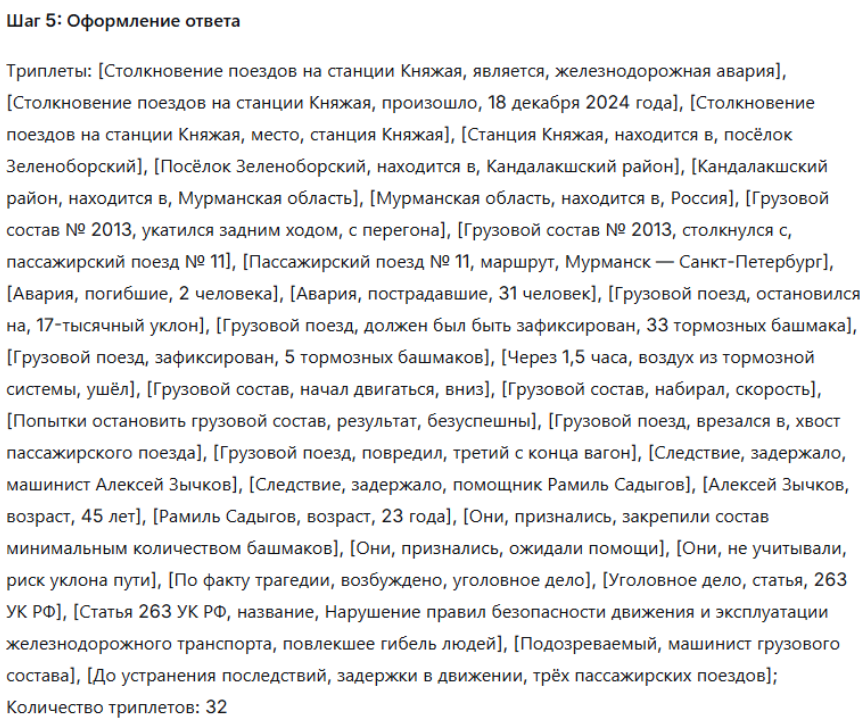


Рис. 4. Интерфейс чата DeepSeek с итоговым ответом модели на запрос

Для комплексной оценки результатов извлечения структурированных данных из текста на естественном языке применялись количественные и качественные показатели эффективности извлечения RDF-троек. Оценка указанных показателей проводилась с помощью привлечения экспертов, что обеспечивает надежность результатов. Количественная оценка включала подсчет общего количества извлеченных из текста RDF-троек. Данная информация была напрямую получена от модели и уже была включена в ее ответ. Особое внимание уделялось подсчету количества встречающихся в ответе модели уникальных RDF-троек, которое позволяет исключить дублирование семантических связей, что повысило точность количественной оценки. Качественная оценка включала определение информационной ценности извлеченных RDF-троек. Основными показателями являлись общее соответствие извлеченных связей фактическому содержанию исходного текста, а также извлечение наиболее значимых, а не поверхностных или второстепенных связей.

Представленная методика обеспечивает надежную основу для получения достоверных результатов исследовательской части работы и может быть адаптирована для решения схожих задач в области обработки и извлечения данных из текстовых источников.

Результаты

В рамках данного исследования проведена оценка эффективности 59 типов промптинга, относящихся к 6 основным стратегиям (техникам) промптинга, для решения задач извлечения RDF-троек из разнородных текстов на естественном языке. Все исследованные типы промптинга классифицированы в соответствии с представленным набором стратегий: 1) Zero-Shot Prompting (типы 1–6); 2) Few-Shot Prompting (типы 7–13); 3) Thought Generation Prompting (типы 14–30); 4) Ensembling Prompting (типы 31–40); 5) Self-Criticism Prompting (типы 41–47); 6) Decomposition Prompting (типы 48–59).

Визуальное представление разработанной классификации представлено на рис. 5.

Для каждого типа промптинга составлен и апробирован специальный модификатор — дополнительный текстовый блок, интегрируемый в экземпляр запроса. Результаты применения каждого модификатора, выраженные в количестве извлеченных RDF-троек, а также пометки

о применимости модификатора систематизированы в сводной таблице, в которой аббревиатура НП, означает, что данный тип неприменим к поставленной задаче, ТПР указывает на то, что для использования этого модификатора необходима программная реализация алгоритма или формальной логики, составляющей основу метода для получения результатов извлечения RDF-троек в данном типе промптинга, а ТЭП — на требование предоставления экспертных примеров.

Типы промптинга, обозначенные в таблице как неприменимые к поставленной в исследовании задачи извлечения RDF-троек, были исключены из данного исследования. Например, типы, предназначенные для машинного перевода, анализа диалогов или работы с визуальными данными, принципиально не могут быть адаптированы для решения поставленной задачи без полного искажения их исходной концепции.

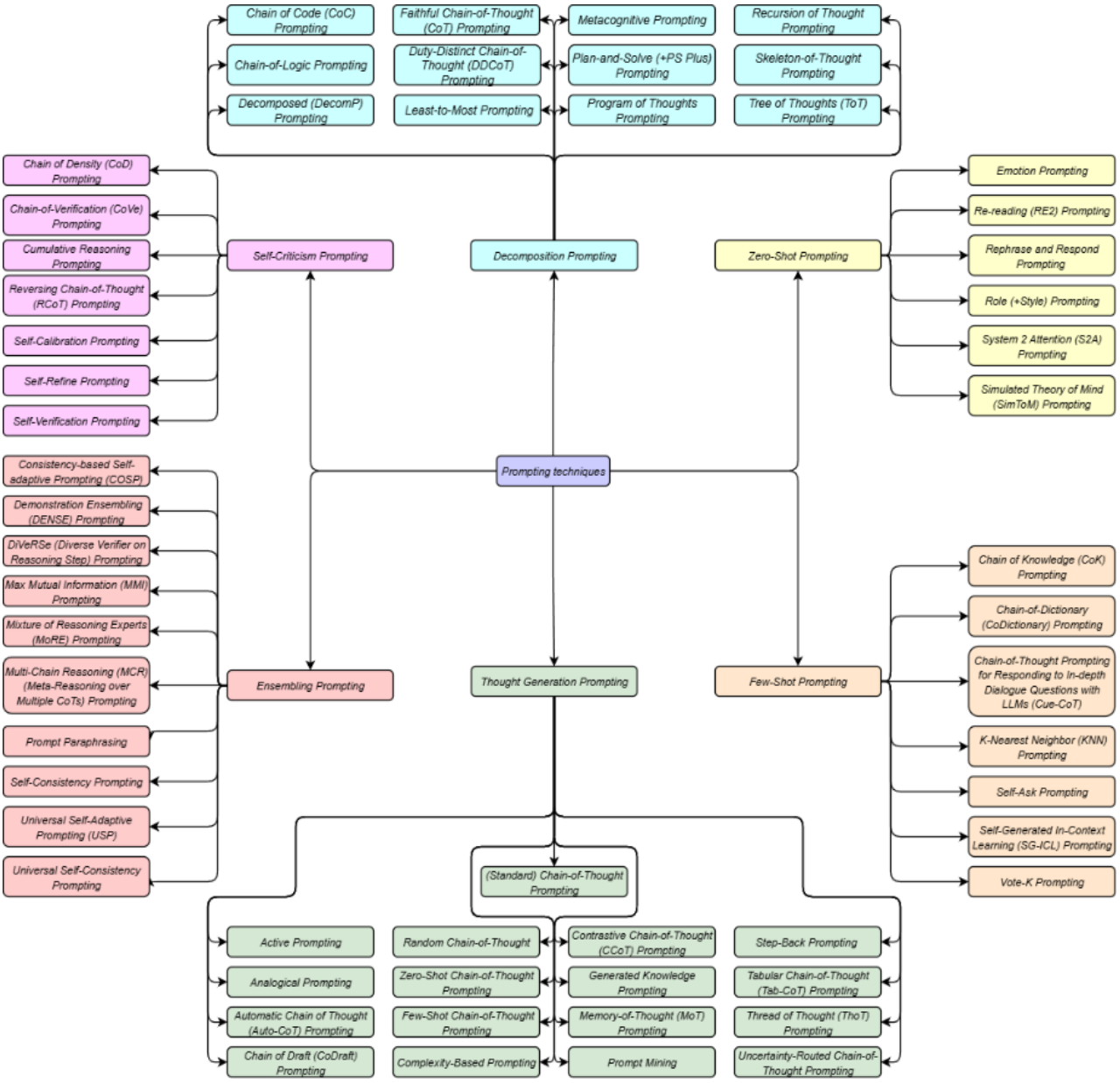


Рис. 5. Классификация стратегий и типов промптинга

Сводная таблица стратегий и типов промптинга

Техника промптинга	№	Тип промптинга	Модификатор промптинга	Количество извлеченных RDF-троек
1	2	3	4	5
Zero-Shot Prompting	*1*	*Emotion Prompting*	"Прошу помогите мне! От успешности выполнения данного задания зависит положение в моей карьере! Задача: [Поставленная задача]"	**45**
	2	*Re-reading (RE2) Prompting*	"Задача: [Поставленная задача]; Прочитайте задачу ещё раз: [Поставленная задача]"	**46**
	3	*Rephrase and Respond (RaR) Prompting*	"Задача: [Поставленная задача]; Перефразируйте и разверните задачу, а затем напишите ответ."	**44**
	4	*Role (+Style) Prompting*	"Вы являетесь [Непрямое указание роли]; Задача: [Поставленная задача]"; {Пример роли: "Вы решили стать одним из самых успешных учёных России в области ИТ, и это возможно, если вы, в меру своих научных познаний, зная, что вы являетесь старшим научным сотрудником, справитесь со следующей задачей"}	**37**
	5	*Simulated Theory of Mind (SimToM) Prompting*	"Ниже приведена последовательность событий: [Последовательность событий поставленной задачи]; О каких событиях знает [имя/название основного персонажа/элемента]?", "Составьте рассказ с точки зрения [имя/название основного персонажа/элемента]; Ответьте на следующий вопрос: [Главный вопрос поставленной задачи]"	**НП**
	6	*System 2 Attention (S2A) Prompting*	"Получив текст от пользователя, извлеките ту часть, которая является непредвзятой, а не его мнение, чтобы использование только этого текста было хорошим контекстом для предоставления непредвзятого ответа на вопросительную часть текста. Укажите фактический вопрос или запрос пользователя. Разделите его на две категории, помеченные как "Непредвзятый текстовый контекст (включает весь контент, кроме предвзятости пользователя):" и "Вопрос/Запрос (не включает предвзятость/предпочтения пользователя):". После, дайте ответ на полученный Вопрос/Запрос. Текст пользователя: [Поставленная задача]"	**48**

Продолжение таблицы

1	2	3	4	5
Few-Shot Prompting	*7*	*Chain of Knowledge (CoK) Prompting*	"Задача: [Пример задачи]; Доказательство: 1. [субъект, отношение, объект], 2. [субъект, отношение, объект], 3. [субъект, отношение, объект] …; Объяснение: [Обоснование логических связей между доказательствами]; Ответ: [Заключение]; Задача: [Поставленная задача]"	**НП**
	8	*Chain-of-Dictionary (CoDictionary) Prompting*	"Переведите следующий текст с [исходный язык] на [целевой язык]: [исходное предложение]; Подсказка для перевода: Переведите следующий текст с [исходный язык] на [целевой язык]: [исходное предложение]; [слово X на языке оригинала] означает [слово X на языке перевода] означает [слово X на вспомогательном языке 1] означает [слово X на вспомогательном языке 2]"	**НП**
	9	*Chain-of-Thought Prompting for Responding to In-depth Dialogue Questions with LLMs (Cue-CoT)*	{O-Cue CoT (One-step Cue Chain-of-Thought):} "Здесь происходит диалог между пользователем и системой. [Контекст диалога]; Сначала выведите одну строку, содержащую статус пользователя, такой как черты характера пользователя, его психологические и эмоциональные состояния, проявленные в разговоре. В следующей строке, пожалуйста, сыграйте роль системы и сгенерируйте ответ на основе статуса пользователя и контекста диалога; Задача: [Поставленная задача]", {M-Cue CoT (Multi-Step Cue Chain-of-Thought):} "[Копия O-Cue CoT без поставленной задачи]; Здесь происходит диалог между пользователем и системой. [Контекст диалога]; Вот статус пользователя. [Извлечённый моделью статус пользователя]; Пожалуйста, играйте роль системы и генерируйте ответ на основе статуса пользователя и контекста диалога. Задача: [Поставленная задача]"	**НП**
	10	*K-Nearest Neighbor (KNN) Prompting*	{Банк примеров: Запрос (Q) - Ответ (A)}, {Значение K метода KNN}; "[Пример N1(k): Q - A, N2(k): Q - A, …]; Задача: [Поставленная задача]"	**ТПР**
	11	*Self-Ask Prompting*	"Задача: [Пример задачи]; Нужны ли здесь дополнительные вопросы: Да; Продолжение: [Подзадача 1] Промежуточный ответ: [Верный ответ на подзадачу 1]; Последующие действия: [Подзадача 2]; Промежуточный ответ: [Верный ответ на подзадачу 2]; Таким образом, окончательный ответ: [Верный ответ на пример задачи]; Задача: [Поставленная задача]; Здесь необходимы дополнительные вопросы: "	**ТЭП**

Продолжение таблицы

1	2	3	4	5
Few-Shot Prompting	*12*	*Self-Generated In-Context Learning (SG-ICL) Prompting*	"Задача: [Пример задачи 1]; Ответ: [Верный ответ на пример задачи 1]; Задача: [Пример задачи схожего с 1]: ", "Задача: [Пример задачи 2]; Ответ: [Верный ответ на пример задачи 2]; Задача: [Пример задачи схожего с 2]: ", "Задача: [Пример задачи схожего с 1]; Ответ: [Ответ модели на пример задачи схожего с 1]; Задача: [Пример задачи схожего с 2]; Ответ: [Ответ модели на пример задачи схожего с 2]; Задача: [Поставленная задача]"	**ТЭП**
	13	*Vote-K Prompting*	{Банк примеров: Запрос (Q) - Ответ (A)}, {Значение K метода Vote-K}; "[Пример N1(k): Q - A, N2(k): Q - A, …] Задача: [Поставленная задача]"	**ТПР**
Thought Generation Prompting	*14*	*Active Prompting*	{Банк примеров задач (Q)} {Uncertainty Estimation: (Uncertainty = H/K, H - количество уникальных ответов)} "[Провести K раз Standard CoT или Zero-Shot CoT или Few-Shot CoT над каждым примером задачи из банка задач]", {Selection: (U↑ => 1 (Сложная задача - берём на дороботку), U↓ => 0 (Лёгкая задача - отбрасываем))}, {Annotation: (Создаём вручную рассуждение над сложной задачей через Chain-of-Thought}, {Inference:} "[Пример(ы) сложной(ых) задач(и) с рассуждением]; З адача: [Поставленная задача]"	**ТПР**
	15	*Analogical Prompting*	"Задача: [Поставленная задача]; Инструкция: 1) Релевантные задачи: Вспомните примеры задач, которые имеют отношение к исходной задаче. Ваши задачи должны отличаться друг от друга и от исходной задачи (например, с разными числами и именами). 2) После написания "Q: ", опишите задачу, после "A: ", поясните решение. Решите первоначальную задачу: "	**33**
	16	*Automatic Chain of Thought (Auto-CoT) Prompting*	{Банк примеров задач (Q)} {Clustering Questions: (Реализация кластеризации при помощи Sentence-BERT)}, {Generating Reasoning Chains:} "[Провести N раз Zero-Shot CoT над каждым примером репрезентативной задачи из каждого кластера банка задач]; [Пример(ы) сложной(ых) задач(и) с рассуждением]; Задача: [Поставленная задача]"	**ТПР**
	17	*Chain of Draft (CoDraft) Prompting*	"Ведите запись того, как вы думаете шаг за шагом и напишите её вместе с ответом. Составляйте только минимальные записи для каждого шага мышления отдельно от ответа, максимум 5 слов; Задача: [Поставленная задача]"	**17**

Продолжение таблицы

1	2	3	4	5
Thought Generation Prompting	*18*	*(Standard) Chain-of-Thought Prompting*	"Задача: [Пример задачи]; Рассуждение: 1. …, 2. …, 3. …; Окончательный ответ: [Верный ответ на пример задачи]; Задача: [Поставленная задача]"	**ТЭП**
	19	*Zero-Shot Chain-of-Thought Prompting*	"Ведите запись того, как вы думаете шаг за шагом и напишите её вместе с ответом. Задача: [Поставленная задача]"	**42**
	20	*Few-Shot Chain-of-Thought Prompting*	"Задача: [Пример задачи 1]; Рассуждение: 1. …, 2. …, 3. …; Окончательный ответ: [Верный ответ на пример задачи 1]; Задача: [Пример задачи 2]; Рассуждение: 1. …, 2. …, 3. …; Окончательный ответ: [Верный ответ на пример задачи 2]; Задача: [Поставленная задача]"	**ТЭП**
	21	*Complexity-Based Prompting*	{Банк примеров: Запрос (Q) - Рассуждение (R) - Ответ (A)} {Select Complex Prompts: (Выборка примеров с большим количеством шагов рассуждения)}, {Generate Multiple Outputs: (Выборка рассуждений на поставленную задачу с большим количеством шагов)} "[Провести N раз Standard CoT над поставленной задачей в моделе обученной на выбранных сложных примерах]", {Vote Among Complex Chains: (Выборка окончательного ответа из отобранных с большим количеством шагов по более частым ответам)}	**ТПР**
	22	*Contrastive Chain-of-Thought (CCoT) Prompting*	"Задача: [Пример задачи]; Верное рассуждение: 1. …, 2. …, 3. …; Верный окончательный ответ: [Верный ответ на пример задачи]; Неверное рассуждение: 1. …, 2. …, 3. …; Неверный окончательный ответ: [Неверный ответ на пример задачи]; Задача: [Поставленная задача]"	**ТЭП**
	23	*Generated Knowledge Prompting*	{Single prompt approach:} "Соберите и выпишите факты по теме поставленной задачи, решите её, воспользовавшись ими; Задача: [Поставленная задача]" или {Dual prompt approach:} "Собери факты по теме поставленной задачи; Задача: [Поставленная задача с полученными от модели фактами]"	**36**
	24	*Memory-of-Thought (MoT) Prompting*	{Банк примеров: Запрос (Q) - Ответ (A)}, {Pre-thinking: (Модель занимается построением нескольких цепочек мыслей к каждому примеру и отбирает самые точные)}, "Решите поставленную задачу, опираясь на прошлые размышления по отобранному набору примеров. Задача: [Поставленная задача]", {Recalling: (Модель извлекает релевантные мысли из памяти, основываясь на сходстве между текущим вопросом и сохраненными)}	**ТПР**

Продолжение таблицы

Thought Generation Prompting	*25*	*Prompt Mining*	{Prompt Generation: (Создание набора шаблонов [субъект, отношение, объект] по изначальной задачи из внешнего источника данных)}, {Prompt Selection: (Отбор лучших шаблонов по определённым метрикам)}, "Задача: [Оптимизированная по отобранным шаблонам поставленная задача]"	**НП**
	26	*Random Chain-of-Thought*	"Задача: [Пример задачи из внешнего источника данных]; Рассуждение: [Цепочки рассуждений к примеру задачи из внешнего источника данных]; Окончательный ответ: [Верный ответ на пример задачи из внешнего источника данных]; Задача: [Поставленная задача]"	**ТЭП**
	27	*Step-Back Prompting*	{Step-back question - производный вопрос от исходного на более высоком уровне абстракции} "[Задаём модели step-back question по поставленной задаче]; Задача: [Поставленная задача, дополненная ответом на step-back question]" {Пример step-back question: "Как представить содержание текста в виде структурированных семантических троек?"}	**53**
	28	*Tabular Chain-of-Thought (Tab-CoT) Prompting*	"Представьте процесс решения поставленной задачи в виде таблицы: "|шаг|подзадача|процедура|результат|"; После, предоставьте ответ, извлёкши результат из заполненной таблицы; Задача: [Поставленная задача]"	**32**
	29	*Thread of Thought (ThoT) Prompting*	"[«Хаотичный» контекст]; Задача: [Поставленная задача]; Проведите меня через этот контекст по управляемым частям шаг за шагом, подводя итоги и анализируя по ходу дела.", "[«Хаотичный» контекст]; Задача: [Поставленная задача]; Проведите меня через этот контекст по управляемым частям шаг за шагом, подводя итоги и анализируя по ходу дела; [Ответ модели на предыдущий запрос]; Таким образом, ответ таков: "	**НП**
	30	*Uncertainty-Routed Chain-of-Thought Prompting*	"Ведите N независимых записей того, как вы думаете шаг за шагом, напишите их вместе с ответами и оцените свою уверенность от 1 до 10 для каждых шагов рассуждений и окончательных ответов. Сравните полученные ответы и оценки уверенности. Если все цепочки пришли к одному ответу и средняя оценка уверенности превышает порог в 7 баллов, то этот ответ считается надежным, иначе это указывает на высокую неопределенность. На основе мажоритарного голосования и анализа уверенности дайте окончательный ответ. Задача: [Поставленная задача]"	**13**

Продолжение таблицы

1	2	3	4	5
Ensembling Prompting	*31*	*Consistency-based Self-adaptive Prompting (COSP)*	{Банк примеров: Запрос (Q) - Ответ (A)}, {Generating Responses: (Модель занимается построением нескольких цепочек мыслей Zero-Shot CoT к каждому примеру, после все пути оцениваются по надежности и согласованности)}, {Selecting Demonstrations: (Лучшие ответы отбираются в качестве демонстраций на основе таких критериев, как последовательность (повторяется ли один и тот же ответ), минимальное количество повторений и разнообразие путей рассуждения)}; "Решите поставленную задачу несколькими путями, опираясь на отобранные демонстрации, оцените каждый ответ от 1 до 10 на предмет их надежности, после мажоритарного голосования дайте окончательный ответ. Задача: [Поставленная задача]"	**ТПР**
	32	*Demonstration Ensembling (DENSE) Prompting*	{Повторить N раз:} "Проанализируйте следующие примеры - [Пример задачи 1], [Пример задачи 2], [Пример задачи 3] и решите следующую задачу: [Поставленная задача]", "Сравните все полученные ответы обоснованных и надежных вариантов в итоговый ответ"	**ТЭП**
	33	*DiVeRSe (Diverse Verifier on Reasoning Step) Prompting*	"Ведите N независимых записей того, как вы думаете шаг за шагом, напишите их вместе с ответами. Задача: [Поставленная задача]", {Score Reasoning Paths: (Оценка вероятности правильности для каждых шагов рассуждений в процентах при помощи верификатора голосования)}, {Step-Aware Verification: (Применение пошаговой верификации для проверки правильности отдельных шагов рассуждения)}, "Используйте взвешенное голосование, чтобы прийти к окончательному ответу, выбрав наиболее вероятный правильный ответ на основе проверенных путей рассуждения"	**ТПР**
	34	*Max Mutual Information (MMI) Prompting*	"Сгенерируйте набор шаблонов промптов для поставленной задачи. Задача: [Поставленная задача]", {Calculate Mutual Information Scores: (Подключите каждый шаблон к алгоритму взаимной информации, чтобы получить оценку для каждого шаблона)}, "Выберите и примените шаблон, получивший наивысшую оценку взаимной информации, на поставленной задаче, и решите её. Задача: [Поставленная задача]"	**НП**
	35	*Mixture of Reasoning Experts (MoRE) Prompting*	"Задача: [Поставленная задача]", {Expert Predictions: (Каждая специализированная модель генерирует ответ на основе своего опыта в рассуждениях)}, {Answer Selection: (Селектор ответов выбирает наиболее надежный ответ, основанный на согласии между экспертами и достоверности прогноза, если такого ответа нет, система воздержится)}	**ТПР**

Продолжение таблицы

1	2	3	4	5
Ensembling Prompting	*36*	*Multi-Chain Reasoning (MCR) (Meta-Reasoning over Multiple CoTs) Prompting*	"Задача: [Пример задачи]; Нужны ли здесь дополнительные вопросы: Да; Продолжение: [Подзадача 1] Промежуточный ответ: [Верный ответ на подзадачу 1]; Последующие действия: [Подзадача 2]; Промежуточный ответ: [Верный ответ на подзадачу 2]; Таким образом, окончательный ответ: [Верный ответ на пример задачи]; Задача: [Поставленная задача]; Ведите запись того, как вы думаете шаг за шагом при решении промежуточных подзадач и напишите её вместе с окончательным ответом. Здесь необходимы дополнительные вопросы: ", "На основе ваших цепочек рассуждений, представленных при решении промежуточных подзадач, сгенерируйте окончательный ответ"	**ТЭП**
	37	*Prompt Paraphrasing*	"Переведите условие исходного запроса на другой язык, не трогая предоставленные данные, а затем обратно, не решая его. Задача: [Поставленная задача]", "Задача: [Перефразированная переводом поставленная задача]"	**НП**
	38	*Self-Consistency Prompting*	{Повторить N раз, вводя после первого запроса фразу "Попробуйте решить задачу ещё раз":} "Задача: [Поставленная задача]", "На основе ваших предыдущих ответов предоставьте окончательный"	**45**
	39	*Universal Self-Adaptive Prompting (USP)*	{Банк примеров: Запрос (Q) - Ответ (A)}, {Task-Type Categorization: (Определение задач к одному из трех типов: CLS, SFG или LFG)}, {Pseudo-Demonstrations: (Генерация демонстрационных примеров для этой категории задач. Модель создает несколько ответов для каждого запроса)}, {Scoring and Selection: (Для отбора наилучших псевдодемонстраций используется целевая функция уверенности. Для CLS - энтропия логов, для SFG - самосогласованность ответов, для LFG - метрики схожести)}, "Используйте отобранные псевдодемонстрации в качестве примеров для решения поставленной задачи. Задача: [Поставленная задача]"	**ТПР**
	40	*Universal Self-Consistency Prompting*	"Сгенерируйте N вариантов решения следующей задачи через цепочки рассуждений с ответами. Оцените все ответы. Выберите наиболее последовательный ответ на основе консенсуса большинства. Начните свой ответ со слов «Наиболее последовательным ответом является ответ X» (без кавычек). Задача: [Поставленная задача]"	**20**

Продолжение таблицы

1	2	3	4	5
Self-Criticism Prompting	*41*	*Chain of Density (CoD) Prompting*	"Статья: [Информационная статья]; Вы будете создавать все более краткие, насыщенные сущностями резюме вышеупомянутой статьи. Повторите следующие 2 шага 5 раз. Шаг 1. Определите 1-3 информативных сущности («;» с разделителями) из статьи, которые отсутствуют в ранее сгенерированном резюме. Шаг 2. Напишите новое, более плотное резюме той же длины, которое охватывает все сущности и детали из предыдущего резюме плюс отсутствующие сущности. Отсутствующая сущность – это: 1) Актуально: к основному сюжету; 2) Конкретный: описательный, но лаконичный (5 слов или меньше); 3) Роман: нет в предыдущем кратком изложении; 4) Верные: присутствуют в Статье; 5) В любом месте: находится в любом месте статьи. Руководящие принципы: 1) Первое резюме должно быть длинным (4-5 предложений, ~80 слов), но крайне неконкретным, содержащим мало информации, за исключением объектов, помеченных как отсутствующие. Используйте слишком многословный язык и заполнители (например, «в этой статье обсуждается»), чтобы достичь ~80 слов; 2) Сделайте каждое слово важным: перепишите предыдущее резюме, чтобы улучшить поток и освободить место для дополнительных сущностей; 3) Освободите пространство с помощью слияния, сжатия и удаления неинформативных фраз типа «обсуждается в статье»; 4) Резюме должно быть очень плотным и кратким, но в то же время самодостаточным, т.е. легко воспринимаемым без статьи. 5) Отсутствующие сущности могут отображаться в любом месте новой сводки; 6) Никогда не удаляйте сущности из предыдущей сводки. Если пространство освободить не удается, добавьте меньше новых объектов; Помните, что для каждого резюме используется одно и то же количество слов. Ответ в формате JSON. JSON должен представлять собой список (длина 5) словарей, ключами к которым являются "Missing_Entities" и "Denser_Summary""	**НП**
	42	*Chain-of-Verification (CoVe) Prompting*	"Задача: [Поставленная задача]", "[Задаём модели возвращающий к поставленной задаче вопрос]", "Воспользуйтесь своим уточнением, чтобы дать окончательный вариант ответа" {Пример возвращающего вопроса: "Верно ли извлечены все триплеты из заданного текста, все ли триплеты указаны?"}	**25**

Продолжение таблицы

1	2	3	4	5
Self-Criticism Prompting	*43*	*Cumulative Reasoning Prompting*	"Для решения поставленной задачи вы будете последовательно исполнять три роли: Proposer, Verifier и Reporter. Proposer: Инициирует рассуждение, предлагая возможные следующие логические шаги или гипотезы на основе текущего контекста и предыдущих установленных фактов. Verifier: Критически оценивает предложения предлагающего. Проверяет их на логическую обоснованность, соответствие предыдущим шагам и отсутствие противоречий. Отвергает неверные предложения. Reporter: Фиксирует верифицированные шаги, обновляя общую цепочку рассуждений. Решает, когда процесс рассуждения может быть завершен для формирования окончательного ответа, или дает команду предлагающему продолжить, исходя из накопленного контекста. Задача: [Поставленная задача]"	**16**
	44	*Reversing Chain-of-Thought (RCoT) Prompting*	"Задача: [Поставленная задача]", "Укажите конкретную задачу, которая может привести к такому ответу. Задача должна содержать всю основную и необходимую информацию и соответствовать ответу. В задаче может быть задан только один результат. Ответ: [Ответ модели на поставленную задачу].", "Перечислите условия исходной задачи и реконструированной задачи. Условий может быть несколько. Исходная задача: [Поставленная задача]; Реконструированная задача: [Реконструированная моделью задача]", "Найдите галлюцинированные и пропущенные условия из списка, если такие имеются: [Список условий исходной и реконструированной задачи]", "Как вы считаете, эти задачи, в конечном счете, задают один и тот же вопрос? Распишите свою причину и ответьте "да" или "нет". Задача: [Поставленная задача]; Реконструированная задача: [Реконструированная моделью задача]", "Соберите все фактические несоответствия из ваших рассуждений и пересмотрите своё решение, если это необходимо. Предоставьте окончательный ответ на поставленную задачу. Задача: [Поставленная задача]"	**48**
	45	*Self-Calibration Prompting*	"Задача: [Поставленная задача]", "Предложенный вами ответ является истинным (верным) или ложным (неверным)?", {Ответ ложный (неверный) или частично истинный (верный):} "Проведите повторное решение поставленной задачи"	**35**
	46	*Self-Refine Prompting*	"Задача: [Поставленная задача]", "Что вы думаете о данном ответе на поставленную задачу?", "Воспользуйтесь своим уточнением, чтобы дать окончательный вариант ответа"	**37**

Продолжение таблицы

1	2	3	4	5
Self-Criticism Prompting	*47*	*Self-Verification Prompting*	{Повторить N раз:} "Задача: [Пример задачи]; Рассуждение: 1. …, 2. …, 3. …; Окончательный ответ: [Верный ответ на пример задачи]; Задача: [Поставленная задача]", "Переформулируйте ответ в утверждение, включающее все условия задачи", {Отсекаем ответ от полученного утверждения:} "[Утверждение без ответа]; Назовите ответ, подразумевающийся в данном утверждении.", "Выберите лучший ответ, учитывая логическую согласованность и отсутствие противоречий"	**ТЭП**
Decomposition Prompting	*48*	*Chain of Code (CoC) Prompting*	{Code Generation:} "Задача: [Поставленная задача]; # Код на Python, возвращающий ответ", {Code Execution with an LMulator: LMulator использует рассуждения на основе языка для моделирования того, каким должен быть вывод на основе контекста семантических частей кода и задачи}	**ТПР**
	49	*Chain-of-Logic Prompting*	"Правило: [Пример правила]; Факты: [Факты, относящиеся к примеру правила]; Задача: [Пример задачи, которую требуется решить]; Я проанализирую эти данные, используя логическую цепочку: 1. Структура входных данных: Правило: [Переформулированное правило]; Факты: [Факты, относящиеся к переформулированному правилу]; Проблема: [Переформулированная задача]; 2. Элементы правила: A: [Первый ключевой элемент правила]; B: [Второй ключевой элемент]; C: [Третий ключевой элемент, если применимо]; 3. Логическое выражение: [Показать, как элементы сочетаются с И/ИЛИ]; 4. Анализ элементов: {Повторить для каждого элемента:} Для элемента A: Q: [Перефразируйте элемент как вопрос] - A: [Анализ, основанный на фактах + Истина/Ложь]; 5. Логический синтез: [Показать выражение с истинными/ложными значениями]; 6. Окончательное решение: [Решение логического выражения]; Теперь, проанализируйте этот случай: Правило: [Правило по поставленной задаче]; Факты: [Факты, относящиеся к исходному правилу]; Задача: [Поставленная задача]"	**НП**
	50	*Decomposed (DecomP) Prompting*	{QC: [Поставленная задача]}, {Каждая подзадача задаётся разным моделям или в разных чатах} "Q1: [Первая подзадача]", "Q2: [Вторая подзадача]", "Q3: [Третья подзадача]", "Задача: [Поставленная задача]; Выведи окончательный ответ на основе уже полученных ответов на подзадачи; A1: [Ответ на первую подзадача]; A2:[Ответ на вторую подзадачу]; A3:[Ответ на третью подзадачу]"	**40**

Продолжение таблицы

1	2	3	4	5
Decomposition Prompting	*51*	*Duty-Distinct Chain-of-Thought (DDCoT) Prompting*	"Система: Вы полезный, высокоинтеллектуальный управляемый помощник. Вы делаете все возможное, чтобы помочь людям выбрать правильный ответ на вопрос. Обратите внимание, что недостаточная информация для ответа на вопросы является обычным явлением, потому что у вас может не быть информации о картине. Окончательный ответ должен быть одним из вариантов; Пользователь: Учитывая контекст, вопросы и варианты, продумайте шаг за шагом предварительные знания для ответа на вопрос, разберите вопрос как можно полнее, вплоть до необходимых подвопросов на основе контекста, вопросов и вариантов. Затем, чтобы помочь людям ответить на исходный вопрос, попытайтесь ответить на подвопросы. Ожидаемая форма ответа выглядит следующим образом: Подвопросы: [Подвопрос 1]; [Подвопрос 2] ...; Дополнительные ответы: [Подответ 1] или "Неопределенно"; [Подответ 2] или "Неопределенно" ...; Ответ: [Один из вариантов] или "Неопределенно"; Что касается вопроса, если у вас нет никакой информации о картине, то всё равно попытайтесь ответить на подвопросы, расставьте приоритеты, могут ли ваши общие знания ответить на него, а затем подумайте, может ли помочь контекст. Если на подвопросы можно ответить, то отвечайте как можно короче предложением. Если подвопросы не могут быть определены без информации в изображениях, пожалуйста, сформулируйте соответствующий подответ как "Неопределенно". Используйте "Неопределенно" в качестве ответа только в том случае, если оно присутствует во вложенных ответах. Все ответы ожидаются максимально лаконичными. Условие: Контекст: [Текстовое описание или любой дополнительный контекст для задачи]; Изображение: [Логический индикатор (например, "True" или "False"), указывающий, есть ли изображение]; Вопрос: [Поставленная задача]; Опции: [Список вариантов ответа]", {Модель визуальных ответов на вопросы (VQA) обрабатывает подвопросы, помеченные как неопределенные}, {Модель сочетает свои лингвистические рассуждения с результатами визуального распознавания для создания обоснований:} "Система: Вы полезный, очень умный учитель. Вы не только сделаете все возможное, чтобы привести людей к правильному ответу, но и предоставите обоснования в качестве ориентира.	**ТПР**

Продолжение таблицы

1	2	3	4	5
Decomposition Prompting	*51*	*Duty-Distinct Chain-of-Thought (DDCoT) Prompting*	Пользователь: Учитывая контекст, вопросы, варианты, дополнительную информацию, думайте шаг за шагом и отвечайте на вопросы. Обратите внимание, что нужен не только ответ, но и, обоснование получения ответа. Ожидаемая форма ответа выглядит следующим образом: Обоснование: [Обоснование]; Ответы: [Один из вариантов ответа]; Обратите внимание, что предоставленная дополнительная информация не всегда может быть действительной. Пожалуйста, выберите достоверную информацию для обоснования и выберите относительно правильный вариант в качестве ответа. Условие: Контекст: [Текстовое описание или любой дополнительный контекст для задачи]; Изображение: [Логический индикатор (например, "True" или "False"), указывающий, есть ли изображение]; Вопрос: [Поставленная задача]; Опции: [Список вариантов ответа]; Дополнительная информация: [Дополнительная информация]"	**ТПР**
	52	*Faithful Chain-of-Thought (CoT) Prompting*	"Задача: [Пример задачи]; Ведя запись того, как вы думаете шаг за шагом, разбейте исходную задачу на более простые подзадачи, на которые легко ответить: Q1: [Первая подзадача], Q2: [Вторая подзадача], Q3: [Третья подзадача]; # Предоставьте код на Python, возвращающий ответ на задачу: [Код на Python, возвращающий ответ на пример задачи]; Задача: [Поставленная задача]; Ведя запись того, как вы думаете шаг за шагом, разбейте исходную задачу на более простые подзадачи, на которые легко ответить; # Предоставьте код на Python, возвращающий ответ на задачу"	**ТПР**
	53	*Least-to-Most Prompting*	"Задача: [Пример задачи]; Ведя запись того, как вы думаете шаг за шагом, разбейте исходную задачу на более простые подзадачи, на которые легко ответить: Q1-A1: [Первая подзадача][Верный ответ на первую подзадача], Q2-A2: [Вторая подзадача][Верный ответ на вторую подзадачу], Q3-A3: [Третья подзадача][Верный ответ на третью подзадачу]; Выведи окончательный ответ на основе уже полученных ответов на подзадачи: [Верный ответ на пример задачи]; Задача: [Поставленная задача]; Ведя запись того, как вы думаете шаг за шагом, разбейте исходную задачу на более простые подзадачи, на которые легко ответить; Выведи окончательный ответ на основе полученных ответов на подзадачи"	**ТЭП**

Продолжение таблицы

1	2	3	4	5
Decomposition Prompting	*54*	*Metacognitive Prompting*	"Задача: [Поставленная задача]; Ведя запись того, как вы думаете шаг за шагом при решении поставленной задачи, выполните все следующие действия: 1. Уточните своё понимание исходного вопроса; 2. Выполните предварительное суждение своего рассуждения; 3. Критически оцените свой предварительный анализ. Если вы не уверены, попробуйте пересмотреть решение; 4. Подтвердите свое окончательное решение. 5. Оцените свою уверенность (от 0 до 100%) в своем анализе и объясните этот уровень уверенности"	**20**
	55	*Plan-and-Solve (+PS Plus) Prompting*	"Задача: [Поставленная задача]; Сначала разберитесь в проблеме и разработайте план ее решения. Затем, ведя запись того, как вы думаете шаг за шагом, выполните разработанный план", {Запись выполнения плана шаг за шагом передаётся в другую модель или чат:} "Задача: [Поставленная задача]; Таким образом, окончательным ответом на поставленную задачу будет: "	**41**
	56	*Program of Thoughts Prompting*	"Задача: [Поставленная задача]; # Предоставьте код на Python, возвращающий ответ на задачу"	**ТПР**
	57	*Recursion of Thought Prompting*	{Обученная модель по рекурсивному принципу «разделяй и властвуй»: (Каждый раз, когда модель сталкивается со сложной проблемой в середине цепочки рассуждений, он отправляет эту проблему в другой запрос, где после завершения этого процесса ответ отправляется в исходный)} "Задача: [Пример задачи]; Рассуждение: 1. …, 2. …, 3. …; Окончательный ответ: [Верный ответ на пример задачи]; Задача: [Поставленная задача]"	**ТПР**
	58	*Skeleton-of-Thought Prompting*	{Skeleton Stage:} "Вы являетесь организатором, ответственным за предоставление только скелета (а не полного содержания) для ответа на задачу. Предоставьте скелет в списке пунктов (пронумерованных 1., 2., 3., и т.д.) для ответа на задачу. Вместо того, чтобы писать полное предложение, каждый скелет должен быть очень коротким, всего 3-5 слов. Как правило, скелет должен иметь 3-10 токенов. Предоставьте скелет для следующей задачи. Задача: [Поставленная задача]", {Point-Expanding Stage:} "Вы несете ответственность расширение пунктов скелета, а также за написание ответа на поставленную задачу на его основе. Расширяя пункты скелета, расписывайте их очень кратко по 1-2 предложению на каждый. Сначала верните расширенный скелет, затем ответ. Задача: [Поставленная задача]; Скелет: [Полученный от модели скелет]"	**19**

Окончание таблицы

1	2	3	4	5
Decomposition Prompting	*59*	*Tree of Thoughts (ToT) Prompting*	"Задача: [Поставленная задача]; Создайте несколько первоначальных подходов к решению исходной задачи. Организуйте мысли в древовидную структуру, где узлы - промежуточные шаги решения, а ветви - альтернативные пути развития мыслей. Оцените перспективность каждого пути. При необходимости возвращайтесь к предыдущим узлам для исследования альтернативных путей решения задачи"	**23**

Типы промптинга, требующие программной реализации, также не могли быть реализованы в рамках стандартного интерфейса чата языковой модели DeepSeek, поскольку они предполагают интеграцию с внешними вычислительными модулями, которые либо необходимо искать в открытых источниках и подключать отдельно, либо разрабатывать самостоятельно.

Необходимо отметить, что типы промптинга, требующие предоставления экспертных примеров, в рамках данного исследования не тестировались эмпирически из-за методологических ограничений, поскольку требовали наличия заранее подготовленного набора размеченных экспертом примеров, создание которых представляет отдельную сложную задачу. Такая задача требует привлечения квалицированных предметных специалистов.

Далее представлены полученные в ходе практического исследования результаты, включая количественные и качественные оценки, анализ применимости различных типов промптов, а также выявленные закономерности.

Необходимо отметить, что для сбора информации о модификаторах промптов каждого типа использовались данных из различных источников. Одна из сложностей при проведении исследования состояла в том, что собранные данные часто были неточными, неполными и даже противоречивыми. В частности, трактовка одних и тех же типов промптинга могла сильно варьироваться от источника к источнику, одинаковые типы относились к различным стратегиям и, наоборот, один и тот же тип мог быть включен в разные стратегии промптинга в разных источниках. В связи с этим исследование и систематизация имеющихся подходов промптинга для понимания их эффективности при извлечении RDF-троек из текстов на естественном языке представляют собой нетривиальную задачу.

Классификация и формирование модификаторов типов промптинга имеют достаточно гибкую интерпретацию, поскольку каждый тип промптинга может работать в связке друг с другом, использовать параметры друг друга, включать один или несколько экспертных примеров или вообще не иметь их. Поэтому предложенная в работе классификация типов промптинга и список их модификаторов представляют собой один из возможных вариантов систематизации и не рассматривается как полностью исчерпывающее или единственно верное решение.

Количественная оценка результатов показала значительный разброс в эффективности различных типов промптинга. Среднее количество извлеченных RDF-троек для всех исследованных типов составило 34. Наибольшее количество RDF-троек (53) было извлечено с использованием Step-Back Prompting (тип 27) благодаря способности модели выявлять абстрактные принципы перед детальным анализом. Среди наиболее эффективных типов промптинга также можно выделить Zero-Shot Chain-of-Thought (тип 19) с 42 RDF-тройками и Reversing Chain-of-Thought (тип 44) с 48 RDF-тройками, демонстрирующие преимущество типов, сочетающих поэтапное рассуждение с механизмом самокоррекции. В то же время наихудшие результаты с наименьшим количеством найденных RDF-троек показали Uncertainty-Routed Chain-of-Thought Prompting (тип 30) с 13 RDF-тройками и Chain of Draft (тип 17) с 17 RDF-тройками, а также Self-Calibration Prompting (тип 45) с 20 RDF-тройками. Такие показатели подтверждают, что высокая вычислительная сложность не является гарантией значительных результатов.

Качественная оценка результатов выявила важную закономерность: даже при высоких количественных показателях некоторые типы промптинга демонстрировали склонность к генерации искаженных данных. В частности, типы, основанные на многошаговых рассуждениях, то есть те, что подразумевают длительную обработку информации, хотя и показывали в среднем высокие количественные результаты, часто порождали семантические галлюцинации — добавление информации, отсутствующей в исходном тексте, или искажение уже имеющейся. Напротив, подходы с элементами самокритики, такие как Chain-of-Verification (тип 42) и Reversing Chain-of-Thought (тип 44) демонстрировали значительно более высокую точность при сохранении хороших количественных показателей.

На основе приведенных в ходе исследования оценок эффективности типов промптинга были сформулированы практические рекомендации по их использованию. Для достижения максимальной полноты извлечения данных следует применять Step-Back Prompting (тип 27), Zero-Shot Chain-of-Thought (тип 19). Для баланса между полнотой и точностью оптимальны Rephrase and Respond (тип 3), Generated Knowledge Prompting (тип 23) и Universal Self-Consistency Prompting (тип 40). Следует избегать по возможности типов промптинга с высокими вычислительными затратами и низкой отдачей, таких как Uncertainty-Routed Chain-of-Thought (тип 30), а также типов, требующих программной реализации, при отсутствии значительных финансовых средств.

Заключение

Проведенное исследование подтвердило основную гипотезу, состоящую в том, что выбор стратегий и типов промптинга оказывает значительное влияние на эффективность извлечения структурированных данных из нехудожественных разнородных текстов. На основе полученных в процессе исследования данных была предложена классификация типов промптинга, выделены 6 базовых стратегий для 59 типов промптинга. Экспериментальная оценка 59 типов промптинга позволила не только выявить закономерности в их эффективности, но и сформировать практические рекомендации по выбору оптимальных решений для задач преобразования неструктурированного текста в целостный набор RDF-троек. Особую значимость приобретают полученные результаты в контексте разработки технологии синтеза адаптивных геосемантических изображений на основе геопространственных знаний, которая является предметом наших дальнейших исследований.

Практическая значимость работы состоит в том, что установленные показатели эффективности в дальнейшем послужат ориентиром в планировании оптимального подхода для автоматического построения графов знаний при помощи больших языковых моделей и для создания эффективных конвейеров обработки разнородных текстов, что позволит существенно повысить релевантность и информативность визуализаций текстовых описаний с географическими свойствами. Выявленные наиболее эффективные типы промптинга, например, такие как Step-Back Prompting и Zero-Shot Chain-of-Thought, представляют ценность для обеспечения полноты извлечения пространственных отношений, что важно для формирования многоуровневых геосемантических моделей. Повышенный интерес представляет адаптация выявленных закономерностей для обработки не только текстов о чрезвычайных ситуациях, но и разнородных географических описаний, что позволит расширить и без того комплексные семантические модели пространственно-временных ситуаций.

Список источников

1. Vicentiy A. V., Dikovitsky V. V., Shishaev M. G. The semantic models of arctic zone legal acts visualization for express content analysis // Computer Science On-line Conference. Springer, Cham. 2018. Vol. 763. P. 216–228.
2. Vicentiy A. Definition and formalization of the user mental model for creating adaptive geointerfaces of decision support systems // Lecture Notes in Networks and Systems. Springer, Cham. 2024. Vol. 733. P. 1095–1105.
3. Gruber T. R. A translation approach to portable ontology specifications // Knowledge Acquisition. Academic Press. 1993. Vol. 5, № 2. P. 199–220.
4. RDF 1.2 Concepts and Abstract Data Model [Электронный ресурс]. URL: https://www.w3.org/TR/rdf12-concepts (дата обращения: 31.10.2025).
5. Gumiel Y. B. et al. Temporal relation extraction in clinical texts: a systematic review // Association for Computing Machinery Computing Surveys. 2022. Vol. 54, № 7. P. 1–36.

6. Sun T. Relation extraction from financial reports: Doctoral dissertation. University of York, 2022.
7. Zhu G. et al. Relationship extraction method for urban rail transit operation emergencies records // Institute of Electrical and Electronics Engineers Transactions on Intelligent Vehicles. 2023. Vol. 8, № 1. P. 520–530.
8. Sharma T., Emmert-Streib F. Deep mining the textual gold in relation extraction // Artificial Intelligence Review. 2024. Vol. 58, № 1. P. 983–1021.
9. Wang H. et al. Deep neural network-based relation extraction: an overview // Neural Computing and Applications. 2022. Vol. 34, № 6. P. 4781–4801.
10. ChatGPT [Электронный ресурс]. URL: https://chatgpt.com (дата обращения: 01.11.2025).
11. Claude [Электронный ресурс]. URL: https://claude.com (дата обращения: 01.11.2025).
12. Gemini [Электронный ресурс]. URL: https://gemini.google.com (дата обращения: 01.11.2025).
13. GigaChat [Электронный ресурс]. URL: https://giga.chat (дата обращения: 01.11.2025).
14. YandexGPT [Электронный ресурс]. URL: https://ya.ru/ai/gpt (дата обращения: 01.11.2025).
15. Schulhoff S. et al. The prompt report: a systematic survey of prompt engineering techniques // arXiv preprint arXiv:2406.06608. 2024.
16. Liu P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing // Association for Computing Machinery Computing Surveys. 2023.Vol. 55, № 9. P. 1–35.
17. Sahoo P. et al. A systematic survey of prompt engineering in large language models: Techniques and applications // arXiv preprint arXiv:2402.07927. 2024.
18. Peng R. et al. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data // arXiv preprint arXiv:2308.03107. 2023.
19. Bisercic A. et al. Interpretable medical diagnostics with structured data extraction by large language models // arXiv preprint arXiv:2306.05052. 2023.
20. Schilling-Wilhelmi M. et al. From text to insight: large language models for chemical data extraction // Chemical Society Reviews. 2025. Vol. 54. P. 1125–1150.
21. Столкновение поездов на станции Княжая // Википедия : сайт [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Столкновение_поездов_на_станции_Княжая (дата обращения: 31.10.2025).
22. DeepSeek [Электронный ресурс]. URL: https://www.deepseek.com (дата обращения: 31.10.2025).
23. LiveBench [Электронный ресурс]. URL: https://livebench.ai (дата обращения: 01.11.2025).
24. Vellum Leaderboard [Электронный ресурс]. URL: https://www.vellum.ai/llm-leaderboard (дата обращения: 01.11.2025).
25. The Big Benchmarks Collection Leaderboard [Электронный ресурс]. URL: https://huggingface.co/collections/open-llm-leaderboard/the-big-benchmarks-collection (дата обращения: 01.11.2025).

References

1. Vicentiy A. V., Dikovitsky V. V., Shishaev M. G. The semantic models of arctic zone legal acts visualization for express content analysis. *Computer Science On-line Conference*. Springer, Cham, 2018, vol. 763, pp. 216–228.
2. Vicentiy A. Definition and formalization of the user mental model for creating adaptive geointerfaces of decision support systems. *Lecture Notes in Networks and Systems*. Springer, Cham, 2024, vol. 733, pp. 1095–1105.
3. Gruber T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*. Academic Press, 1993, vol. 5, no. 2, pp. 199–220.
4. RDF 1.2 Concepts and Abstract Data Model. Available at: https://www.w3.org/TR/rdf12-concepts (accessed 31.10.2025).
5. Gumiel Y. B. et al. Temporal relation extraction in clinical texts: a systematic review. *Association for Computing Machinery Computing Surveys*, 2022, vol. 54, no. 7, pp. 1–36.
6. Sun T. *Relation extraction from financial reports*. Doctoral dissertation. University of York, 2022.
7. Zhu G. et al. Relationship extraction method for urban rail transit operation emergencies records. *Institute of Electrical and Electronics Engineers Transactions on Intelligent Vehicles*, 2023, vol. 8, no. 1, pp. 520–530.
8. Sharma T., Emmert-Streib F. Deep mining the textual gold in relation extraction. *Artificial Intelligence Review*, 2024, vol. 58, no. 1, pp. 983–1021.
9. Wang H. et al. Deep neural network-based relation extraction: an overview. *Neural Computing and Applications*, 2022, vol. 34, no 6, pp. 4781–4801.
10. ChatGPT. Available at: https://chatgpt.com (accessed 01.11.2025).
11. Claude. Available at: https://claude.com (accessed 01.11.2025).
12. Gemini. Available at: https://gemini.google.com (accessed 01.11.2025).
13. GigaChat. Available at: https://giga.chat (accessed 01.11.2025).
14. YandexGPT. Available at: https://ya.ru/ai/gpt (accessed 01.11.2025).

15. Schulhoff S. et al. The prompt report: a systematic survey of prompt engineering techniques. arXiv preprint arXiv:2406.06608, 2024.
16. Liu P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Association for Computing Machinery Computing Surveys*, 2023, vol. 55, no 9, pp. 1-35.
17. Sahoo P. et al. A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927, 2024.
18. Peng R. et al. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. arXiv preprint arXiv:2308.03107, 2023.
19. Bisercic A. et al. Interpretable medical diagnostics with structured data extraction by large language models. arXiv preprint arXiv:2306.05052, 2023.
20. Schilling-Wilhelmi M. et al. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews*, 2025, vol. 54, pp. 1125–1150.
21. Stolknovenie poezdov na stancii Knyazhaya [Train Collision at Knyazhaya Station]. (In Russ.). Available at: https://ru.wikipedia.org/wiki/Столкновение_поездов_на_станции_Княжая (accessed 31.10.2025).
22. DeepSeek. Available at: https://www.deepseek.com (accessed 31.10.2025).
23. LiveBench. Available at: https://livebench.ai (accessed 01.11.2025).
24. Vellum Leaderboard. Available at: https://www.vellum.ai/llm-leaderboard (accessed 01.11.2025).
25. The Big Benchmarks Collection Leaderboard. Available at: https://huggingface.co/collections/open-llm-leaderboard/the-big-benchmarks-collection (accessed 01.11.2025).

Информация об авторах

Р. А. Горбунов — стажер-исследователь;
А. В. Вицентий — кандидат технических наук, старший научный сотрудник.

Information about the authors

R. A. Gorbunov — Research Assistant;
A. V. Vicentiy — Candidate of Science (Tech.), Senior Research Fellow.

Статья поступила в редакцию 21.11.2025; одобрена после рецензирования 24.11.2025; принята к публикации 25.11.2025.
The article was submitted 21.11.2025; approved after reviewing 24.11.2025; accepted for publication 25.11.2025.

Научная статья
УДК 004.932, 004.89, 549.08
doi:10.37614/2949-1215.2025.16.3.007

ЭКСПРЕСС-ТЕХНОЛОГИЯ ФОРМИРОВАНИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ ДЛЯ ПЛАНИМЕТРИЧЕСКОГО МИНЕРАЛОГИЧЕСКОГО АНАЛИЗА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Владимир Витальевич Диковицкий[1]✉, Максим Геннадьевич Шишаев[2]
[1, 2]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[1]v.dikovitsky@ksc.ru✉, https://orcid.org/0000-0003-0329-9979
[2]m.shishaev@ksc.ru, https://orcid.org/0000-0001-7070-7878

Аннотация

В работе рассматривается реализация планиметрического метода минералогического анализа как задачи классификации. Для ее решения используется машинное обучение, а для формирования обучающей выборки предложена экспресс-технология, основанная на аугментации данных. Для получения признакового вектора объектов классификации использован предобученный нейросетевой векторизатор на базе сверточной искусственной нейронной сети ResNet-18. Проведенные эксперименты на примере анализа апатитовых руд Хибинского месторождения показали, что использование такого вектора обеспечивает весьма высокую точность классификации чистых образцов апатита (до 99 % в зависимости от вида руды и размера планиметрической сетки). А с учетом специфики рассматриваемых в работе объектов классификации аналогичный подход может использоваться и для решения задачи определения удельного содержания минералов в рудах площадным методом. Предложенная в работе технология экспресс-разметки образцов обеспечивает приемлемую итоговую точность решения этой задачи при минимальных трудозатратах на создание обучающей выборки.

Ключевые слова:

минералогический анализ, машинное обучение, классификация, планиметрический метод

Благодарности:

исследование выполнено в рамках государственного задания Института информатики и математического моделирования Кольского научного центра Российской академии наук, Министерства науки и высшего образования Российской Федерации, темы научно-исследовательской работы «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределённости и риска» (шифр темы FMEZ-2025-0053).

Для цитирования:

Диковицкий В. В., Шишаев М. Г. Экспресс-технология формирования обучающей выборки для планиметрического минералогического анализа на основе методов машинного обучения // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 106–116. doi:10.37614/2949-1215.2025.16.3.007.

Original article

EXPRESS TECHNOLOGY FOR FORMING A TRAINING SET FOR PLANIMETRIC MINERALOGICAL ANALYSIS BASED ON MACHINE LEARNING METHODS

Vladimir V. Dikovitsky[1]✉, Maxim G. Shishaev[2]
[1, 2]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[1]v.dikovitsky@ksc.ru✉, https://orcid.org/0000-0003-0329-9979
[2]m.shishaev@ksc.ru, https://orcid.org/0000-0001-7070-7878

Abstract

This paper examines the implementation of the planimetric method of mineralogical analysis as a classification problem. Machine learning is used to solve this problem, and an express technology based on data augmentation is proposed for generating a training set. A pre-trained neural network vectorizer based on the ResNet-18 convolutional artificial neural network is used to obtain the feature vector for classification objects. Experiments conducted using the analysis of apatite ores from the Khibiny deposit demonstrated that using this vector ensures a very high classification accuracy for pure apatite samples (up to 99 %, depending on the ore type and the size of the planimetric grid). Given the specific nature of the classification objects considered in this paper, a similar approach can also be used to determine the specific mineral content in ores using the areal method. The proposed express sample labeling technology ensures acceptable final accuracy for this problem with minimal effort required to generate the training set.

Keywords:
mineralogical analysis, machine learning, classification, planimetric method
Acknowledgments:
The study was carried out within the framework of the Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences state assignment of the Ministry of Science and Higher Education of the Russian Federation, research topic “Methods and technologies for creating intelligent information systems to support the development of complex dynamic systems with regional specifics in conditions of uncertainty and risk” (registration number of the research topic FMEZ-2025-0053).
For citation:
Shishaev M. G., Dikovitsky V. V. Express technology for forming a training set for planimetric mineralogical analysis based on machine learning methods. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 106–116. doi:10.37614/2949-1215.2025.16.3.007.

Введение

Планиметрический минералогический анализ относится к так называемым площадным методам, когда делается оценка объемного содержания минерала в образце на основе соответствующей оценки доли минерала на плоском изображении. Площадной анализ опирается на принцип Делессе [1], в соответствии с которым предполагается, что площадная оценка эквивалентна объемной. Это предположение справедливо, если используется «правильная» проекция образца (для слоистых руд — сечение, поперечное слоям минерала) или большое количество случайных проекций, а также если измеряемая площадь сечения много больше площади сечения отдельных зерен изучаемого минерала. Отметим, что для рассматриваемых в данной работе апатитовых руд эти условия выполнимы.

Для получения площадной оценки планиметрическим методом на изображение образца руды накладывается прямоугольная сетка, после чего каждая ячейка сетки (далее по тексту именуемая также секцией) идентифицируется с тем или иным минералом. Таким образом, планиметрический анализ эквивалентен задаче классификации изображений, для решения которой широко применяются методы машинного обучения.

В целом, при использовании методов машинного обучения в площадном минералогическом анализе, задача интерпретируется чаще всего как задача сегментации изображений [2–9]. Главной проблемой такого подхода является необходимость в качественных размеченных наборах данных, получение которых является трудоемким и затратным по времени процессом. Поэтому планиметрический метод анализа может рассматриваться как альтернатива методам, основанным на сегментации изображений.

Ключевыми проблемами при решении задачи классификации с помощью методов машинного обучения являются получение набора признаков объектов, обеспечивающего хорошую разделимость классов, а также формирование размеченного набора данных (датасета) достаточно большого размера для тренировки модели. В данной работе рассматривается технология автоматизированного получения обучающей выборки для тренировки нейросетевых классификаторов, использующих для получения признакового вектора предобученную сверточную нейронную сеть ResNet. Исследования проводились на примере задачи определения удельной доли полезного компонента[1] (ПК) в апатитовой руде.

Исследование эффективности классификации апатита на базе признаковых векторов ResNet-18

В настоящее время лучшие результаты в качестве универсального векторизатора изображений для различных приложений показывают сверточные нейронные сети [10]. В нашей работе в качестве классифицирующего признака использовались векторы изображений, полученные с помощью предобученной сверточной остаточной (residual) нейронной сети ResNet-18 [11]. Отметим, что в качестве векторизатора могут рассматриваться и другие варианты сверточных сетей [12], однако на задаче

[1] Строго говоря, термин «полезный компонент» означает некоторое химическое соединение, определяющее полезные свойства руды. Однако, ввиду того что объемное содержание минерала в руде однозначно определяет содержание полезного компонента, в данной статье для упрощения изложения мы будем использовать этот термин для обозначения минерала, содержащего искомое химическое соединение.

распознавания апатита уже ResNet-18, с относительно небольшими количеством слоев и размером выходного вектора (512 компонентов), обеспечила достаточно высокое качество результата.

Для подтверждения перспективности использования такого признакового вектора был проведен предварительный анализ имеющегося экспериментального датасета с помощью метода главных компонент. Для эксперимента были отобраны ячейки, представляющие положительный («апатит») и отрицательный («не_апатит») классы, полученные наложением планиметрической сетки размером 20 × 20 пикселей (всего 100 000 объектов). Полученное распределение секций по классам на редуцированном пространстве изображено на рис. 1. По результатам эксперимента можно сделать вывод о хорошей разделимости объектов по использованному признаку, что позволяет говорить о перспективности создания классификатора для решения рассматриваемой задачи.

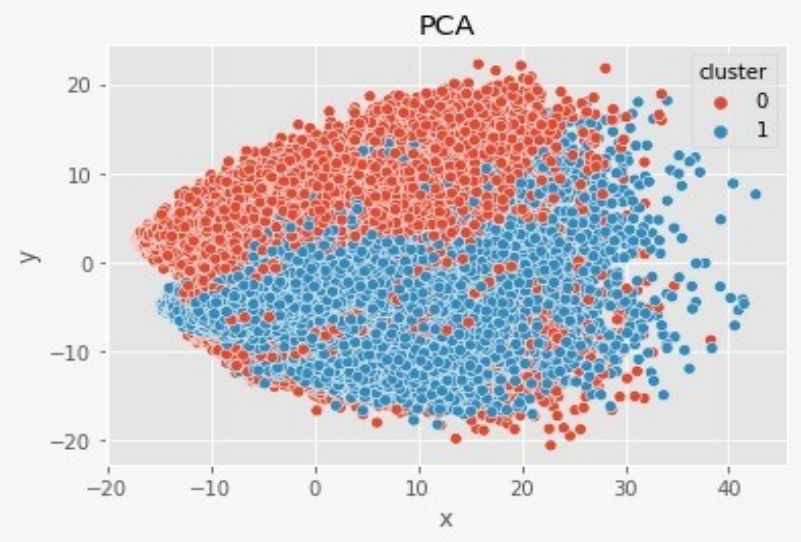


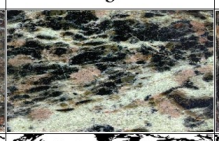
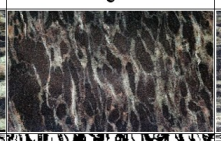
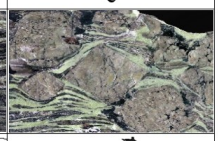
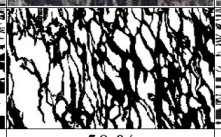
Рис. 1. Распределение объектов в редуцированном пространстве признаков

Специфической проблемой планиметрического метода анализа как задачи классификации является неоднозначность отнесения к тому или иному классу ячеек, находящихся на границе двух и более минералов («пограничных» ячеек). Строго говоря, в данном случае мы имеем дело с задачей мягкой (нечеткой) классификации, когда объект принадлежит тому или иному классу с не стопроцентной достоверностью. Степень принадлежности объекта классу, как правило, оценивают в диапазоне от 0 до 1, при этом крайние значения степени принадлежности соответствуют ячейкам, представляющим чистые положительный или отрицательный классы, а промежуточные — ячейкам, лежащим на границе минерального пятна. Далее рассмотрены эксперименты по классификации ячеек планиметрической сетки, представляющих чистые классы, и пограничных ячеек.

В экспериментах использовались датасеты, полученные наложением планиметрической сетки различных размеров на эталонные размеченные изображения образцов апатитовой руды различных типов, представленные в табл. 1. Разметка образцов осуществлялась экспертами-минералогами и представлена соответствующими изображениями-масками, на которых черный цвет соответствует апатиту. Изображения образцов были получены с помощью обычной фотокамеры, разрешение 300 dpi, размер 915 × 709 пикселей. В качестве классификатора использовалась ИНС прямого распространения с пятью полносвязными слоями. Тренировка классификатора осуществлялась на датасетах, сформированных из одной части изображений образцов руды, а для тестирования использовались объекты, полученные из изображений, не участвовавших в тренировке, то есть, наряду с кросс-валидацией, осуществлялась проверка нейросетевого классификатора на внешних данных.

Таблица 1

Использованные в экспериментах образцы апатитовой руды

Вид руды	Апатитовый уртит	Пятнисто-полосчатая	Пятнистая	Линзовидно-полосчатая	Блоковая
Индекс образца	a	b	c	d	e
Исходное фото					
Изображение-маска					
Содержание ПК	25 %	41 %	50 %	40 %	25 %

Результаты тестирования классификаторов, натренированных на различных наборах объектов (полученных из изображений руды разных типов), представлены в табл. 2. Серым цветом выделены ячейки, отражающие результаты кросс-валидации, ячейки без заливки содержат результаты проверки точности квалификации на внешних данных. Для формирования датасетов использовалась планиметрическая сетка размером 10 × 10 пикселей.

Таблица 2

Результаты тестирования бинарного классификатора на чистых классах

	Классификатор, натренированный											
	на образце “a”					на образце “b”					на образцах “a, b, d, e”	
Предъявляемые для классификации образцы	a	b	c	d	e	a	b	c	d	e	a, b, d, e	c
Точность классификации (Precision), %	98,6	93,8	36,5	85,4	51,1	59,8	99,9	38,6	95,1	93,5	98,5	98,8

Особенностью рассматриваемой задачи классификации, определяемой природой исследуемых объектов (изображений минералов), является однородность классов. Это свойство заключается в том, что при достаточно крупном размере ячейки, принадлежащей строго одному классу, ее элементы (более мелкие разбиения) будут также принадлежать этому же классу. Очевидно, что при достижении некоторого предельно малого размера ячейки это свойство перестанет выполняться. Предельный размер ячейки в этом смысле определяется морфологическими свойствами рассматриваемого минерала (размером характерных визуальных признаков). В рамках работы были проведены эксперименты по определению предельного размера ячейки, позволяющего эффективно идентифицировать апатит с использованием признакового вектора ResNet-18. Результаты, представленные в табл. 3, показывают, что для рассматриваемого минерала свойство однородности класса соблюдается вплоть до размера планиметрической ячейки 2 × 2 пикселя.

Таблица 3

Эффективность бинарной классификации при различных размерах ячейки

Размер ячейки	2 × 2	5 × 5	10 × 10	20 × 20
Точность классификации на валидационной части выборки	0,78	0,90	0,98	0,99

Таким образом, использование векторизатора ResNet-18 для получения признакового вектора образцов изображений в задаче идентификации чистых классов в рассматриваемой задаче дает высокий эффект. Классификатор, натренированный на датасете, полученном на основе 4 из 5 имеющихся

эталонных изображений («a, b, d, e»), демонстрирует очень высокую точность классификации как на валидационной выборке, так и на внешних данных (секциях, полученных из изображения «c»).

Однако, как отмечалось выше, при минералогическом анализе планиметрическим методом мы имеем дело не только с представителями чистых классов (ячейками сетки, полностью принадлежащими положительному или отрицательному классу), но и с пограничными секциями, содержащими изображения минералов как положительного (в нашем случае — апатита), так и отрицательного классов. Одним из вариантов решения задачи определения удельного содержания ПК в данном случае является непосредственная интерпретация оценки вероятности принадлежности входного объекта положительному классу, получаемой на выходе бинарного классификатора, как доли ПК в ячейке.

С целью проверки эффективности такого подхода к определению ПК планиметрическим методом использовалось размеченное изображение апатита повышенного разрешения с размером секции 20 × 20 пикселей. Бинарному нейросетевому классификатору, натренированному на размеченном датасете, включающем объекты положительного и отрицательного классов, были предъявлены 99 пограничных объектов с различным содержанием минерала, представляющих промежуточный класс. Результаты классификации пограничных объектов представлены на рис. 2 (объекты отсортированы в порядке убывания степеней принадлежности к положительному классу). Как видим, классификатор уверенно относит к положительному классу объекты с содержанием минерала выше 0,5 и практически утрачивает способность к классификации при содержании минерала менее 0,5.

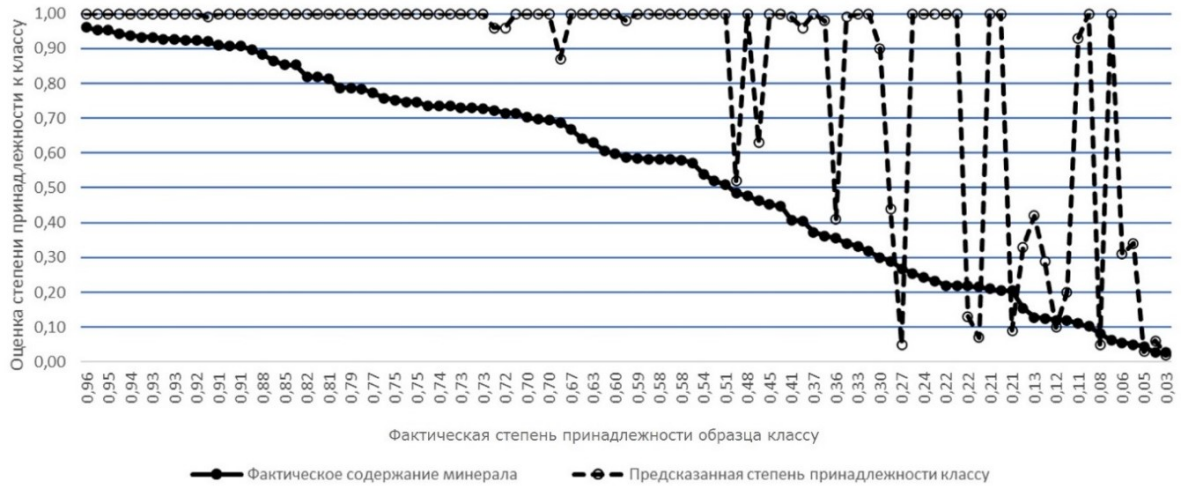


Рис. 2. Результаты классификации пограничных объектов бинарным классификатором

Расчетное значение предсказанного содержания ПК в рассмотренных пограничных объектах составило 0,84 при истинном значении 0,55, что свидетельствует о невозможности непосредственного использования бинарного классификатора для определения удельного содержания полезного компонента в руде планиметрическим методом. Однако можно сделать предположение о том, что пограничные объекты в нашем случае будут в среднем содержать идентичные объемы положительного и отрицательного классов, то есть оценка удельного содержания ПК в таких объектах будет стремиться к 0,5 с увеличением их количества. В свою очередь, количество пограничных объектов при том же изображении образца будет возрастать с уменьшением размера ячейки планиметрической сетки.

В случае справедливости данного предположения, при определении удельного содержания ПК планиметрическим методом с использованием бинарного классификатора будет иметь место систематическая ошибка, зависящая от количества пограничных объектов в датасете, сформированном из рассматриваемого образца изображения руды. В табл. 4 приведены расчетные значения содержания ПК в пограничных ячейках при наложении планиметрической сетки различного размера на имеющиеся образцы. Как видим, предположение о сбалансированности содержания ПК в пограничных ячейках в целом подтверждается.

Таблица 4

Удельное содержание ПК в пограничных ячейках

Размер ячейки	Индекс образца				
	a	b	c	d	e
5 × 5	0,50	0,48	0,52	0,51	0,52
10 × 10	0,49	0,45	0,53	0,50	0,54
15 × 15	0,48	0,42	0,54	0,48	0,53
20 × 20	0,46	0,40	0,53	0,47	0,53

Введем следующие обозначения: c — истинное удельное содержание полезного компонента в образце; $\tilde{c}$ — оценка удельного содержания ПК в образце, полученная с помощью бинарного классификатора; N^+ — количество ячеек положительного класса в образце; N^b — количество пограничных ячеек в образце; N — общее количество ячеек в образце; λ — относительная ошибка бинарного классификатора на пограничных ячейках.

Тогда с учетом сбалансированности содержания ПК в пограничных секциях фактическое содержание ПК в образце может быть рассчитано по формуле:

$$c = \frac{N^+ + N^b/2}{N}. \quad (1)$$

В свою очередь, предсказанное бинарным классификатором содержание ПК в том же образце с учетом единичной точности идентификации чистых классов и ошибки на пограничных ячейках:

$$\tilde{c} = \frac{N^+ + \lambda N^b/2}{N}. \quad (2)$$

Из формул (1) и (2) следует:

$$c = \tilde{c} - (\lambda - 1)\frac{N^b}{2N}.$$

Таким образом, при соблюдении условия сбалансированности классов в пограничных ячейках и известном количестве последних, бинарный классификатор может быть использован для оценки удельного содержания полезного компонента в образцах руды с применением планиметрического метода.

Экспресс-технология формирования обучающей выборки

Проведенные эксперименты показывают эффективность применения машинного обучения (нейросетевого классификатора) для определения удельного содержания минерала в руде площадным методом. Основным сдерживающим фактором широкого применения данного подхода на практике является необходимость обучения классификатора на образцах, наиболее полно представляющих возможные варианты анализируемой руды с точки зрения ее цвето-яркостных и морфологических характеристик. Разумно предположить, что затруднительно построить классификатор, универсальный для всех типов и разновидностей руд. С другой стороны, наши эксперименты показывают, что использование при обучении классификатора более представительной выборки, включающей образцы руды тех типов, которые будут анализироваться на этапе использования классификатора (в нашем случае это различные апатитовые руды Хибинского месторождения), обеспечивает определенный уровень универсальности при сохранении высоких показателей точности классификации.

Тем не менее при изменении типа или разновидности анализируемой руды возникает необходимость обучения нового классификатора. Это, в свою очередь, требует наличия качественно размеченных образцов, используемых для формирования обучающей выборки. Ручная разметка изображения руды является трудоемким и продолжительным по времени процессом. Однако с учетом свойства однородности классов для рассматриваемого типа руд, отмеченного ранее, возможно сократить трудоемкость и длительность этого процесса путем разметки лишь части изображения

образца руды с последующей аугментацией полученной обучающей выборки. Схема соответствующей экспресс-технологии представлена на рис. 3 и включает в себя следующие основные этапы:

1. Формирование базового набора объектов выборки (альтернативно: либо способом (a), либо способом (b)):

a. Выбор и ручная разметка малого фрагмента изображения руды с последующим наложением планиметрической сетки на фрагмент и его маску и автоматизированным формированием множества размеченных объектов.

b. Наложение планиметрической сетки на исходное изображение целиком с последующим ручным заданием набора планиметрических ячеек, достаточно полно представляющих положительный, отрицательный и (при многоклассовой постановке задачи) промежуточные классы.

2. Аугментация выборки путем смещения и/или поворота ячеек относительно геометрического центра.

3. Тренировка классификатора.

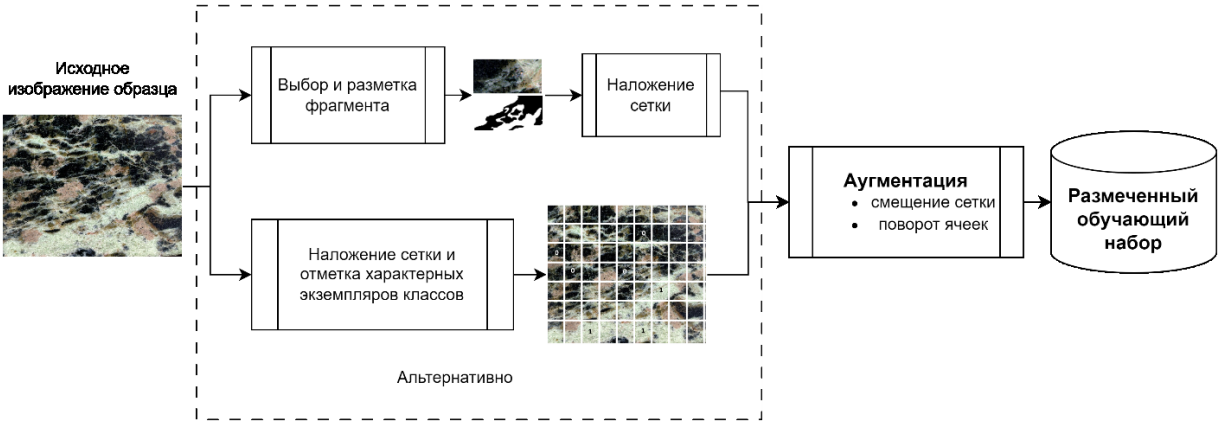


Рис. 3. Технология формирования обучающего набора с помощью аугментации

Технология была опробована на примере определения содержания ПК планиметрическим методом в образце пятнисто-полосчатой апатитовой руды Хибинского месторождения. Для формирования выборки использовался способ (a): на исходном изображении был выделен и размечен фрагмент 200 × 100 пикселей (рис. 4), на который затем была наложена планиметрическая сетка размером 5 × 5 пикселей. Далее полученный набор объектов положительного и отрицательного классов подвергся аугментации путем последовательного сдвига планиметрической сетки. Полученный в результате набор объектов был сбалансирован по классам методом случайного сокращения мажоритарного класса (субдискретизации, under-sampling) [13].

a *b*

Рис. 4. Использованные для формирования обучающей выборки фрагмент изображения (*a*) и соответствующая ему маска (*b*)

В итоге обучение проводилось в 8 эпох на 3181 образце, 796 образцов использовались для валидации. Значение точности на обучающем наборе составило 0,9053, на валидационном наборе — 0,8935. Результаты обучения представлены на рис. 5.

Полученный классификатор затем был использован для разметки полного исходного изображения. Результаты разметки маски минеральных пятен апатита в образце и расчета на ее основе процентного содержания искомого минерала при различных способах разметки представлены в табл. 5.

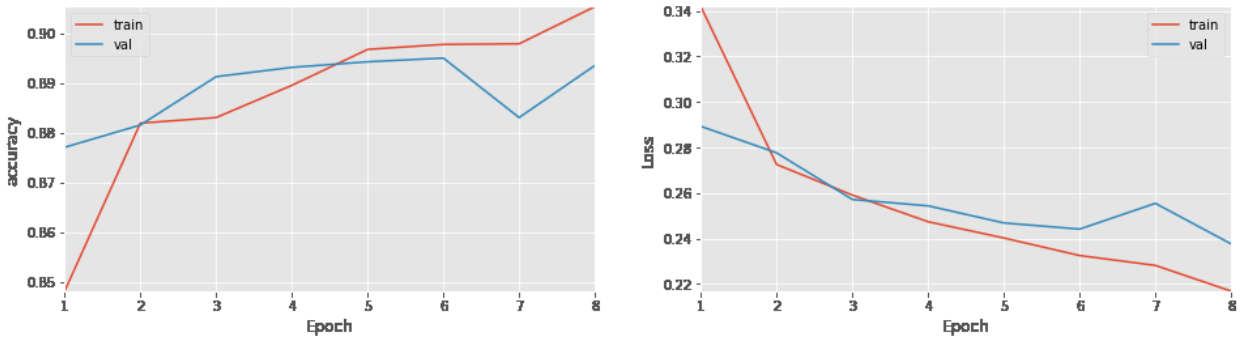
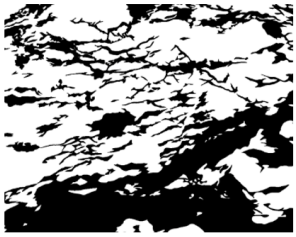


Рис. 5. Результаты тренировки классификатора на аугментированном наборе данных

Таблица 5

Результаты ручной и автоматизированной разметки образца апатитовой руды

	Ручная разметка	Разметка классификатором, натренированным на всех образцах апатитовой руды	Разметка классификатором, натренированным на обучающем наборе, полученном экспресс-методом
Маска минеральных пятен апатита			
Расчетное содержание ПК	33,40 %	33,73 %	35,97 %

Таким образом, при использовании экспресс-метода формирования обучающего набора точность определения содержания ПК образце, в сравнении с классификатором, обученным на полном наборе образцов, снизилась лишь на 2,24 %. Это является, на наш взгляд, вполне адекватной «платой» за кардинальное сокращение временных затрат на ручную минералогическую разметку образцов.

Следует также отметить, что применение в рамках экспресс-технологии второго способа формирования обучающей выборки возможно лишь при использовании планиметрического подхода к площадному анализу. При этом возможна отметка в качестве образцов положительного и отрицательного классов произвольных прямоугольных областей, которые в нашем случае, в силу свойства однородности классов, могут быть разделены на более мелкие фрагменты для формирования более объемной обучающей выборки.

Заключение

В результате проделанной работы исследованы возможность и эффективность реализации количественного минералогического анализа планиметрическим методом как задачи классификации с применением методов машинного обучения, а также предложена технология формирования соответствующих обучающих выборок экспресс-методом на основе аугментации.

Для получения признакового вектора, идентифицирующего целевой минерал по изображению, в работе использована предобученная сверточная ИНС ResNet-18. В целом использование сверточных

ИНС для извлечения признаков соответствует современным трендам в применении методов машинного зрения в минералогическом анализе. При этом наши эксперименты показывают, что предобученные сверточные ИНС могут быть использованы для извлечения признаков в контексте рассматриваемой прикладной задачи без каких-либо изменений в архитектуре или дообучения.

Проведенные эксперименты с образцами апатитовой руды различных видов продемонстрировали весьма высокую эффективность классификации чистых образцов апатита на основе признакового вектора, полученного с помощью ИНС ResNet-18. Причем высокая точность классификации сохраняется даже при относительно малых размерах идентифицируемого изображения ячейки планиметрической сетки и существенно деградирует лишь при размере ячейки 2 × 2 пикселя. Очевидно, что оптимальный размер используемой планиметрической сетки определяется, с одной стороны, морфологическими свойствами рассматриваемой руды, а с другой — разрешением используемых изображений. Больший размер сетки обеспечивает более высокую точность идентификации целевого минерала, но при этом усложняет получение обучающей выборки достаточного объема. Также верхняя граница размера ячейки определяется возрастающей систематической ошибкой определения удельного содержания ПК в силу увеличения количества объектов, включающих смесь целевого минерала с прочими компонентами руды.

Основной проблемой при решении задачи автоматизированного планиметрического анализа в мультиклассовой постановке является получение обучающей выборки, в которой в достаточной мере представлены объекты всех рассматриваемых классов. В качестве решения этой проблемы в работе предложена и опробована экспериментально экспресс-технология формирования обучающего набора данных с использованием аугментации и частичной разметки изображения образца. В рамках технологии предложены два способа частичной разметки — полная минералогическая разметка фрагмента изображения образца с последующим автоматическим формированием размеченного датасета и непосредственное указание меток классов произвольным фрагментам изображения образца, соответствующим ячейкам наложенной планиметрической сетки. При использовании первого способа появляется возможность получения набора данных для мультиклассовой классификации. Однако в этом случае фрагмент изображения должен быть в достаточной мере представителен в смысле возможных вариаций визуализации каждого идентифицируемого класса. Обеспечить это условие не всегда возможно: вследствие неравномерности освещения и иных причин изображение целевого минерала и иных компонентов руды на различных участках изображения может существенно отличаться. Этого недостатка лишен второй способ, при котором пользователь может отметить любые элементы исходного изображения, обеспечив представленность в выборке различных вариаций изображения идентичных компонентов (классов) образца. В рамках данной работы осуществлено предварительное тестирование технологии, показавшее перспективность предложенного подхода в формировании обучающих наборов данных для рассматриваемой категории задач анализа изображений. В целом возможность получения обучающей выборки с невысокими трудозатратами создает условия для быстрой адаптации технологии к новым видам минералов и руд и повышает универсальность предложенного подхода.

Продолжение работы возможно в различных направлениях. Очевидно, что требуется экспериментальная проверка гипотезы о применимости рассмотренного подхода к количественному анализу других типов руд, а также к площадному анализу изображений в рамках иных прикладных задач. Анализ использованного в данной работе экспериментального датасета показал, что при наложении планиметрической сетки расчетная ошибка определения удельного содержания минерала является близкой к симметричной — в пограничных ячейках сетки доли компонентов, соответствующих положительному и отрицательному классам, примерно равны. С учетом этого перспективным представляется использование многоэтапной (иерархической) классификации, при которой на стартовых этапах анализируются макропараметры образцов руды — вид руды с точки зрения минерального состава, морфологические особенности и т. п., а на последующих этапах в соответствии с идентифицированными макропараметрами производится детальный минералогический анализ. При этом в зависимости от результатов классификации на макроуровне для последующего анализа могут применяться различные модели и алгоритмы.

Список источников

1. Weibel E. R., Elias H. Introduction to stereologic principles // Quantitative Methods in Morphology / Quantitative Methoden in der Morphologie / eds. E. R. Weibel, H. Elias. Berlin, Heidelberg: Springer, 1967. P. 89–98.
2. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks / A. Chattopadhay [и др.] // 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018. Grad-CAM++. С. 839–847.
3. Srinivas S., Fleuret F. Full-Gradient Representation for Neural Network Visualization // Advances in Neural Information Processing Systems. Curran Associates, Inc., 2019. Vol. 32.
4. Baklanova O., Shvets O. Cluster analysis methods for recognition of mineral rocks in the mining industry // 2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA) 2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA). 2014. P. 1–5.
5. Baklanova O. E., Baklanov M. A. Methods and Algorithms of Image Recognition for Mineral Rocks in the Mining Industry // Advances in Swarm Intelligence: Lecture Notes in Computer Science / eds. Y. Tan, Y. Shi, L. Li. Cham: Springer International Publishing, 2016. P. 253–262.
6. Launeau P., Cruden A., Bouchez J. Mineral recognition in digital images of rocks: a new approach using multichannel classification // Canadian Mineralogist. 1994. Mineral recognition in digital images of rocks. No. 32. P. 919–933.
7. Review of Nodule Mineral Image Segmentation Algorithms for Deep-Sea Mineral Resource Assessment / W. Song [et al.] // Computers, Materials and Continua. 2022. Vol. 73. No. 1. P. 1649–1669.
8. Deep learning-based method for SEM image segmentation in mineral characterization, an example from Duvernay Shale samples in Western Canada Sedimentary Basin / Z. Chen [et al.] // Computers & Geosciences. 2020. Vol. 138. P. 104450.
9. Deep neural networks for improving physical accuracy of 2D and 3D multi-mineral segmentation of rock micro-CT images / Y. D. Wang [et al.] // Applied Soft Computing. 2021. Vol. 104. P. 107185.
10. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects / Z. Li [et al.] // IEEE Transactions on Neural Networks and Learning Systems. 2022. Vol. 33. A Survey of Convolutional Neural Networks. No. 12. P. 6999–7019.
11. Deep Residual Learning for Image Recognition / K. He [et al.] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 770–778.
12. MineralImage5k: A benchmark for zero-shot raw mineral visual recognition and description / S. Nesteruk [et al.] // Computers & Geosciences. 2023. Vol. 178. MineralImage5k. P. 105414.
13. Yen S.-J., Lee Y.-S. Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset // Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006: Lecture Notes in Control and Information Sciences / eds. D.-S. Huang, K. Li, G. W. Irwin. Berlin, Heidelberg: Springer, 2006. P. 731–74.

References

1. Weibel E. R., Elias H. Introduction to stereologic principles. *Quantitative Methods in Morphology / Quantitative Methoden in der Morphologie* / eds. E. R. Weibel, H. Elias. Berlin, Heidelberg, Springer, 1967, pp. 89–98.
2. Chattopadhay A. et al. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV) 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, Grad-CAM++, pp. 839–847.
3. Srinivas S., Fleuret F. Full-Gradient Representation for Neural Network Visualization. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019, vol. 32.
4. Baklanova O., Shvets O. Cluster analysis methods for recognition of mineral rocks in the mining industry. *2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA) 2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2014, pp. 1–5.
5. Baklanova O. E., Baklanov M. A. Methods and Algorithms of Image Recognition for Mineral Rocks in the Mining Industry. *Advances in Swarm Intelligence: Lecture Notes in Computer Science* / eds. Y. Tan, Y. Shi, L. Li. Cham, Springer International Publishing, 2016, pp. 253–262.
6. Launeau P., Cruden A., Bouchez J. Mineral recognition in digital images of rocks: a new approach using multichannel classification. *Canadian Mineralogist*, 1994, Mineral recognition in digital images of rocks, no. 32, pp. 919–933.

7. Song W. et al. Review of Nodule Mineral Image Segmentation Algorithms for Deep-Sea Mineral Resource Assessment. *Computers, Materials and Continua*, 2022, vol. 73, no. 1, pp. 1649–1669.
8. Chen Z. et al. Deep learning-based method for SEM image segmentation in mineral characterization, an example from Duvernay Shale samples in Western Canada Sedimentary Basin. *Computers & Geosciences*, 2020, vol. 138, p. 104450.
9. Wang Y. D. et al. Deep neural networks for improving physical accuracy of 2D and 3D multi-mineral segmentation of rock micro-CT images. *Applied Soft Computing*, 2021, vol. 104, p. 107185.
10. Li Z. et al. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, vol. 33, A Survey of Convolutional Neural Networks, no. 12, pp. 6999–7019.
11. He K. et al. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
12. Nesteruk S. et al. MineralImage5k: A benchmark for zero-shot raw mineral visual recognition and description. *Computers & Geosciences*, 2023, vol. 178, MineralImage5k, p. 105414.
13. Yen S.-J., LeeY.-S. Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset. *Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006: Lecture Notes in Control and Information Sciences* / eds. D.-S. Huang, K. Li, G. W. Irwin. Berlin, Heidelberg: Springer, 2006, pp. 731–740.

Информация об авторах

В. В. Диковицкий — кандидат технических наук, старший научный сотрудник;
М. Г. Шишаев — доктор технических наук, главный научный сотрудник.

Information about the authors

V. V. Dikovitsky — Candidate of Science (Tech.), Senior Research Fellow;
M. G. Shishaev — Doctor of Science (Tech.), Chief Research Fellow.

Статья поступила в редакцию 30.10.2025; одобрена после рецензирования 14.11.2025; принята к публикации 17.11.2025.
The article was submitted 30.10.2025; approved after reviewing 14.11.2025; accepted for publication 17.11.2025.

Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 117–130.
Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2025. Vol. 16, No. 3. P. 117–130.

Научная статья
УДК 004.832
doi:10.37614/2949-1215.2025.16.3.008

РЕШЕНИЕ ЗАДАЧ ГЕНЕРАТИВНОГО ДИЗАЙНА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ УДОВЛЕТВОРЕНИЯ ОГРАНИЧЕНИЙ

Полина Владимировна Таран[1], Александр Анатольевич Зуенко[2]✉
[1, 2]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[1]p.taran@ksc.ru, https://orcid.org/0009-0003-9485-7004
[2]zuenko@ksc.ru✉, https://orcid.org/0000-0002-7165-6651

Аннотация

Данная работа посвящена рассмотрению методов решения задачи генеративного дизайна. В настоящее время подобные задачи решаются, как правило, с использованием нейросетевого подхода. В представленных исследованиях предлагается задачу генеративного дизайна ставить как задачу удовлетворения ограничений и решать с использованием технологии программирования в ограничениях. Предлагаемый подход иллюстрируется на примере задачи проектирования двумерной пространственной среды с учетом разнородных требований к взаимному расположению объектов среды.

Ключевые слова:

генеративный дизайн, задача удовлетворения ограничений, программирование в ограничениях, комбинаторный поиск, распространение ограничений

Благодарности:

работа выполнена в рамках темы НИР «Методы и информационные технологии мониторинга и управления региональными критическими инфраструктурами Арктической зоны Российской Федерации» (FMEZ-2025-0054).

Для цитирования:

Таран П. В., Зуенко А. А. Решение задач генеративного дизайна с использованием методов удовлетворения ограничений // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 117–130. doi:10.37614/2949-1215.2025.16.3.008.

Original article

SOLVING GENERATIVE DESIGN PROBLEMS USING CONSTRAINT SATISFACTION METHODS

Polina V. Taran[1], Aleksandr A. Zuenko[2]✉
[1, 2]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[1]p.taran@ksc.ru, https://orcid.org/0009-0003-9485-7004
[2] a.zuenko@ksc.ru✉, https://orcid.org/0000-0002-7165-6651

Abstract

The paper is devoted to the consideration of methods for solving the generative design problems. Traditionally such problems are solved using a neural network approach. In the study, it is proposed to represent the problem of generative design as a constraint satisfaction problem and solving it using constraint programming technology. The proposed approach is illustrated by the example of the problem of two-dimensional spatial scena design, taking into account the heterogeneous requirements for the relative location of objects.

Keywords:

generative design, constraint satisfaction problem, constraint programming, combinatorial search, constraint propagation

Acknowledgments:

The work was carried out within the framework of the current research topic "Methods and information technologies for monitoring and management of regional critical infrastructures in the Arctic zone of the Russian Federation" (registration number FMEZ-2025-0054).

For citation:

Taran P. V., Zuenko A. A. Solving generative design problems using constraint satisfaction methods. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 117–130. doi:10.37614/2949-1215.2025.16.3.008.

Введение

Процесс получения решения в рамах технологии генеративного дизайна (проектирования), иногда называемой технологией порождающего проектирования, заключается в том, что пользователь задает технические требования и параметры задачи, а программа генерирует варианты ее решения [1]. Пространство поиска при этом достаточно велико, чтобы человек вручную мог корректно построить все интересующие альтернативы, а затем их проанализировать. Алгоритм построения решений, как правило, не оперирует принципиально новыми идеями, а комбинирует хорошо зарекомендовавшие себя варианты реализации отдельных компонентов проектируемой системы и предлагает решения, оптимальные по определенной системе критериев.

Среди методов генеративного дизайна можно выделить следующие группы [2–4]: методы с использованием грамматик формы [5; 6], L-системы [7], клеточные автоматы [8], методы на основе эволюционных вычислений, генетических алгоритмов, роевого интеллекта [9; 10].

В последнее время подобные задачи решаются с использованием нейросетевого подхода, а именно с помощью генеративных нейронных сетей [11; 12]. В представленных исследованиях предлагается задачу генеративного дизайна ставить как задачу удовлетворения ограничений и решать с использованием технологии программирования в ограничениях [13]. Предлагаемый подход иллюстрируется на примере задачи проектирования двумерной пространственной среды с учетом разнородных требований к взаимному расположению объектов среды.

Задача генеративного дизайна

Генеративный дизайн (Generarive Desing) — парадигма проектирования, в которой процесс создания проектных решений автоматизирован и делегирован вычислительной системе [4]. В отличие от традиционного компьютерного моделирования, в котором инженер вручную создает проектное решение, использование парадигмы генеративного дизайна позволяет описать задачу, задать целевую функцию и ограничения, а алгоритм автоматически сгенерирует множество решений, удовлетворяющих заданным условиям. Автоматическая генерация множества решений позволяет сократить время проектирования и выбрать из предложенных альтернатив лучшую согласно различным критериям.

Алгоритмы генеративного дизайна изначально использовались в области архитектуры и строительства, но в настоящее время данный подход применяется в различных областях промышленного и потребительского дизайна [4].

Ключевыми этапа генеративного дизайна являются: 1) описание задачи (определение проектной области, задание ограничений и целевой функции); 2) поиск всех решений на основе описанной задачи; 3) анализ решений (интерпретация полученных решений); 4) оценка решений по заданным критериям; 5) постобработка, уточнение предметной области путем задания новых ограничений.

В настоящее время второй этап, который может быть автоматизирован, как правило, реализован с помощью нейросетевого подхода (нейроморфные методы). В настоящей работе предлагается описывать и решать задачу генеративного дизайна как задачу удовлетворения ограничений.

Задача удовлетворения ограничений

Задача удовлетворения ограничений — это задача поиска решений для сети ограничений, которая формализуется в виде трех множеств [13]:

$X = \{X_1, X_2, \ldots X_n\}$ — множество переменных, представляющих элементы задачи;

$D = \{D_1, D_2, \ldots D_n\}$ — множество областей определения (доменов) переменных, причем каждый домен D_i определяет множество допустимых значений, которые может принимать переменная X_i;

$C = \{C_1, C_2, \ldots C_n\}$ — множество ограничений, где каждое ограничение C_j — это отношение, определяющее допустимые комбинации значений для некоторого подмножества переменных.

Существует три вида ограничений. *Унарное* — ограничение затрагивает домен единственной переменной. Например, ограничение $X_1 < m$, где m — любое число. *Бинарное* — ограничение, связывающее между собой две переменные. Например, ограничение $X_1 > X_2$. *Ограничения высшего порядка* — ограничение, связывающее между собой три и более переменные. Каждое ограничение высокого порядка с конечной областью определения можно свести к множеству бинарных

ограничений, введя достаточное количество вспомогательных переменных. Примером ограничения высшего порядка является ограничение *alldiff* [13], которое требует, чтобы все переменные задачи принимали различные значения.

Процесс поиска решения состоит из двух чередующихся этапов:

1) распространение — общее название методов вывода на ограничениях, сводящихся к пошаговому удалению из доменов заведомо недопустимых значений. Изменения в доменах одних переменных по принципу «цепной реакции» приводит к усечению доменов других переменных. Когда пространство задачи больше нельзя сократить за счет распространения, используется ветвление;

2) ветвление. Суть процедуры ветвления заключается в разделении пространства задачи на подпространства меньшей размерности, соответствующие подзадачам. Одним из способов ветвления является присваивание рассматриваемой переменной выбранного значения и дальнейшее распространение для корректировки доменов остальных переменных.

Путем чередования шагов ветвления и распространения каждой переменной задачи присваивается значение, которое является одним из *решений* задачи.

В случае, когда в процессе поиска возникает противоречие (домен одной из переменных становится пустым), используется алгоритм поиска с возвратом, при котором осуществляется возврат из текущего состояния задачи до предыдущего непротиворечивого, и исследуется другая ветвь дерева поиска.

Правило, устанавливающее, какую переменную и значение выбирать при ветвлении, называется *эвристикой*. Удачный выбор эвристики позволяет значительно сократить дерево поиска.

Представление задачи генеративного дизайна как задачи удовлетворения ограничений

В настоящей работе предлагается задачу генеративного дизайна представлять и решать в виде задачи удовлетворения ограничений. В качестве примера задачи генеративного дизайна здесь и далее рассматривается задача проектирования двумерной среды с учетом разнородных требований к взаимному расположению объектов среды.

Для представления данной задачи в рамках задачи удовлетворения ограничений необходимо задать сеть ограничений. В качестве переменных в задаче выступают объекты, которые необходимо расположить в двумерном пространстве. В работе двумерное пространство представлено в виде прямоугольной сетки, разделенной на пронумерованные клетки (ячейки) одинакового размера. Объект описывается в виде набора клеток пространства, а также параметров длины и ширины. Каждое значение домена переменной — это номер клетки, в которой может размещаться верхняя левая клетка объекта. Ниже приведено графическое представление пространства, объекта и его домена (рис. 1).

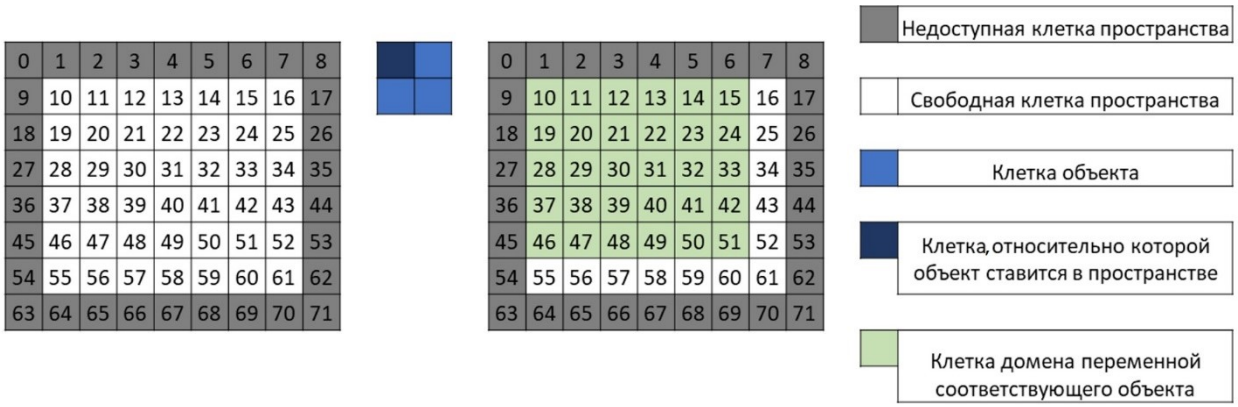


Рис. 1. Графическое представление двумерной сцены, объекта и его домена

Серым на рис. 1 отмечены клетки, которые нельзя занимать, в данном случае это только границы пространства, но незанимаемые клетки могут располагаться и внутри пространства. *Белым* отмечены свободные клетки, в которых можно расположить клетки объектов. *Синим* отмечены клетки объекта, а *темно-синим* — клетка, относительно которой объект ставится в пространстве. *Зеленым* отмечены клетки, в которые можно поставить показанный на рис. 1 объект верхней левой клеткой.

В рамках работы было разработано четыре вида ограничений:

1) ограничение *GDConstraint*, согласно которому в одной клетке пространства не может располагаться два объекта одновременно. Данное ограничение обязательное и является глобальным, т. е. затрагивает все переменные задачи;

2) ограничение *CornerCosntraint* — унарное ограничение, согласно которому объект должен располагаться в углу (т. е. с двух сторон от объекта должны быть недоступные клетки);

3) ограничение *DistanceToWall* — унарное ограничение, согласно которому объект должен располагаться на заданном расстоянии от заданной стены (стеной считается граница пространства или набор недоступных клеток внутри пространства). Расстояние считается в клетках пространства;

4) ограничение *NearConstraint* — бинарное ограничение, связывающее два объекта правилом, что они должны располагаться рядом друг с другом на расстоянии, не превышающем заданное (в клетках пространства).

Все ограничения, кроме первого вида, задаются пользователем факультативно для некоторых из размещаемых объектов.

Для каждого из ограничений был разработан свой алгоритм распространения. Данные алгоритмы подробно описываются далее.

Описание разработанных алгоритмов удовлетворения ограничений

Глобальное ограничение GDConstraint

После того, как у одной из переменных задачи определяется значение (т. е. объект окажется «привязан» к некоторой клетке пространства, расположен в пространстве), необходимо для всех нерасставленных объектов определить недопустимые клетки, чтобы исключить ситуацию, когда несколько объектов занимают одну область пространства

Для этого необходимо из доменов переменных, соответствующих нерасставленным объектам, удалить четыре типа клеток:

клетки, занимаемые поставленным объектом:

$$p + i + width_g \times j, \text{ при } i \in [0, width_1), j \in [0, height_1); \quad (1)$$

клетки выше поставленного объекта:

$$p + i - width_g \times j, \text{ при } i \in [0, width_1), j \in [1, height_2); \quad (2)$$

клетки левее поставленного объекта:

$$p - i + width_g \times j, \text{ при } i \in [1, width_2), j \in [0, height_1); \quad (3)$$

клетки выше и левее поставленного объекта:

$$p - I - width_g \times j, \text{ при } I \in [1, width_2), j \in [1, height_2). \quad (4)$$

В формулах используются следующие обозначения: $width_1$, $height_1$ — ширина и высота поставленного объекта; $width_2$, $height_2$ — ширина и высота объекта, у которого пересчитывается домен; $width_g$ — ширина пространства; p – ячейка, где расположена верхняя левая клетка объекта.

Пример: есть два объекта *Y* и *Z* (рис. 2), которые необходимо расставить в пространстве (рис. 3).

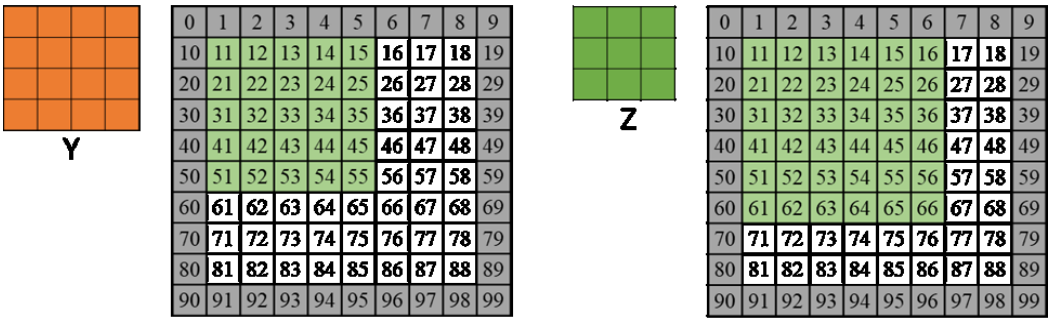
Пусть объект *Y* поставили в клетку 55. Необходимо удалить из домена переменной *Z* номера тех клеток, при постановке в которые объекта *Z* будет пересечение с объектом *Y*.

Согласно формуле 1 рассчитываются клетки, которые занимает объект *Y*:

$Del_1 = \{55, 56, 57, 58, 65, 66, 67, 68, 75, 76, 77, 78, 85, 86, 87, 88\}$.

Новым доменом переменной *Y* будет разность множества домена *Z* и множества Del_1 $D_Y = D_Y \setminus Del_1$. В результате из домена удалятся клетки {55,56,65,66}.

Аналогично, согласно формулам 2–4 рассчитываются недопустимые клетки вокруг *Y*: $Del_2 = \{35, 36, 37, 38, 45, 46, 47, 48\}$; $Del_3 = \{53, 54, 63, 64, 73, 74, 83, 84\}$; $Del_4 = \{33, 34, 43, 44\}$. В результате разности домена переменной *Y* с множествами Del_2, Del_3, Del_4 получится домен $D_Y = \{11, 12, 14, 15, 16, 21, 22, 23, 24, 25, 26, 31, 32, 41, 42, 51, 52, 61, 62\}$ (рис. 4).



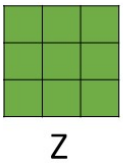
0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	**16**	**17**	**18**	19
20	21	22	23	24	25	**26**	**27**	**28**	29
30	31	32	33	34	35	**36**	**37**	**38**	39
40	41	42	43	44	45	**46**	**47**	**48**	49
50	51	52	53	54	55	**56**	**57**	**58**	59
60	**61**	**62**	**63**	**64**	**65**	**66**	**67**	**68**	69
70	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	79
80	**81**	**82**	**83**	**84**	**85**	**86**	**87**	**88**	89
90	91	92	93	94	95	96	97	98	99

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	**17**	**18**	19
20	21	22	23	24	25	26	**27**	**28**	29
30	31	32	33	34	35	36	**37**	**38**	39
40	41	42	43	44	45	46	**47**	**48**	49
50	51	52	53	54	55	56	**57**	**58**	59
60	61	62	63	64	65	66	**67**	**68**	69
70	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	79
80	**81**	**82**	**83**	**84**	**85**	**86**	**87**	**88**	89
90	91	92	93	94	95	96	97	98	99

Рис. 2. Переменные *Y*, *Z* и их домены

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99

Рис. 3. Пространство 10 × 10



0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99

Рис. 4. Домен переменной Z после распространения

Ограничение CornerConstraint

Так как данное ограничение унарное, то оно будет удовлетворено при первоначальном распространении. Для этого для каждого значения из домена переменной проверяется, будет ли объект, поставленный в эту клетку, располагаться в углу. Также можно задавать расположение объекта у конкретного угла (или нескольких углов).

Для каждого из четырех видов углов проверяются две опорные клетки, которые должны быть незанимаемыми. Если хоть одна из них свободна, то, значит, объект не будет располагаться в проверяемом углу.

Для ***верхнего левого*** угла опорными являются клетка выше $p - width_g$ и клетка левее $p - 1$.

Для ***верхнего правого*** угла вначале вычисляется верхняя правая клетка объекта $p_1 = p + width_{obj} - 1$, а после проверяются клетка выше $p_1 - width_g$ и клетка правее $p_1 + 1$.

Для ***нижнего левого*** угла вначале вычисляется нижняя левая клетка объекта $p_2 = p + (height_{obj} - 1) \times width_g$, а затем проверяются клетка ниже $p_2 + width_g$ и клетка левее $p_2 - 1$.

Для ***нижнего правого*** угла вначале вычисляется нижняя правая клетка объекта $p_3 = p + (height_{obj} - 1) \times width_g + width_{obj} - 1$, а после проверяются клетка ниже $p_3 + width_g$ и клетка правее $p_3 + 1$.

Если при проверке всех углов, объект не располагается ни в одном из углов, следовательно, проверяемая клетка p не удовлетворяет ограничению и удаляется из домена.

Пример: есть объект *Y* (см. рис. 2), который необходимо поставить в любой из углов пространства 10 × 10 (см. рис. 3).

При распространении проверяется каждая клетка, но в данном примере показывается алгоритм при рассмотрении клетки 51. Вначале необходимо определить четыре крайние клетки объекта:

верхняя левая клетка $p = 51$;

верхняя правая клетка $p_1 = p + width_{obj} - 1 = 54$;

нижняя левая клетка $p_2 = p + (height_{obj} - 1) \times width_g = 81$;

нижняя правая клетка $p_3 = p + (height_{obj} - 1) \times width_g + width_{obj} - 1 = 84$.

Затем для каждой из крайних точек проверяются две опорные клетки:

для верхней левой клетки $p = 51$: клетки выше ($p - width_g = 41$) и левее ($p - 1 = 50$). Так как клетка 41 является незанимаемой, то объект Y не будет располагаться в верхнем левом углу пространства.

для верхней правой клетки $p_1 = 54$: клетки выше ($p_1 - width_g = 44$) и правее ($p_1+1 =55$). Обе клетки являются свободными, поэтому условие на расположение в верхнем правом углу не выполняется.

для нижней левой клетки $p_2 = 81$: клетки ниже ($p_2 + width_g = 91$) и левее ($p_2 - 1 = 80$). Обе клетки являются незанимаемыми, а значит, выполняется условие на расположение объекта в нижнем левом углу.

для нижней правой клетки $p_3 = 84$: клетки ниже ($p_3 + width_g = 94$) и правее ($p_3 + 1 = 85$). Так как клетка 85 свободна, то условие на расположение объекта в нижнем правом углу не выполняется.

В результате проверки получилось, что если объект Y поставить в клетку 51, то объект будет располагаться в нижнем левом углу, что удовлетворяет ограничению, а значит, значение {51} остается в домене. В случае, когда при проверке значения выявляется, что объект не будет располагаться в каком-то из углов пространства, то проверяемая клетка удаляется из домена.

В результате распространения в домене переменной останутся только те значения, которые удовлетворяют ограничению (рис. 5).

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99

Рис. 5. Домен переменной Y после распространения

Ограничение DistanceToWallConstraint

Данное ограничение также является унарным и удовлетворяется при первоначальном распространении.

У ограничения два параметра — направление (*direction*) и расстояние (*distance*). Параметр направления (*direction*) может принимать четыре значения: «*North*» — вверх; «*South*» — низ; «*East*» — право; «*West*» — лево. Параметр расстояния (*distance*) измеряется в клетках пространства и может задаваться двумя вариантами: ***жестко*** ограничивать максимальное расстояние до «стены» или ***гибко*** задавать, что расстояние до стены должно быть не меньше заданного.

При создании пространства для каждого из направлений создается матрица, в элементах которой хранится информация о расстоянии от соответствующей клетки до стены. Например, ниже показано графическое представление пространства и матрицы для направления «*South*» (рис. 6).

При распространении рассматривается каждое значение p из домена переменной: вычисляется набор клеток $\{p_k\}$, которые займет объект, если поставить его в проверяемую клетку, по следующей формуле:

$$p + i + j \times width_g, \text{ при } i \in [0, width_1), j \in [0, height_1). \quad (5)$$

Далее для каждого значения $\{p_k\}$ из полученного набора вычисляются координаты x и y. Координатой y является результат целочисленного деления p_i на ширину пространства ($y = p_i \; DIV \; width_g$); а x — это остаток от деления p_i на ширину пространства ($x = p_i \; MOD \; width_g$).

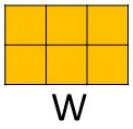
Затем из полученных пар координат из матрицы расстояний вычисляются соответствующие элементы и берется минимальное из них. Если полученное значение не равно заданному, значит, проверяемая клетка p не удовлетворяет ограничению и удаляется из домена.

0	1	2	3	4	5	6	7	8	9
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	19
20	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	29
30	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	39
40	41	42	43	**44**	**45**	**46**	**47**	**48**	49
50	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	59
60	**61**	**62**	**63**	**64**	**65**	**66**	**67**	**68**	69
70	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	79
80	81	82	83	84	85	86	87	88	89

-1	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**
-1	**2**	**2**	**2**	**6**	**6**	**6**	**6**	**-1**
-1	**1**	**1**	**1**	**5**	**5**	**5**	**5**	**-1**
-1	**0**	**0**	**0**	**4**	**4**	**4**	**4**	**-1**
-1	**-1**	**-1**	**-1**	**3**	**3**	**3**	**3**	**-1**
-1	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**-1**
-1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**-1**
-1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**-1**
-1	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**	**-1**

Рис. 6. Пространство и матрица расстояний для направления *«South»*

Пример: есть объект *W* (рис. 7) и пространство (см. рис. 6). Задано ограничение, что объект *W* должен располагаться строго на расстоянии 1 от «стены» по направлению «*South*».



0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89

Рис. 7. Переменная *W* и ее домен

При распространении проверяется каждое значение из домена, но в данном примере рассматривается клетка 13.

Вначале необходимо получить список $\{p_i\}$ всех клеток, которые займет объект при постановке его в клетке 13, согласно формуле 5: $p + i + j \times width_g$, при $i \in [0, 3)$, $j \in [0, 2)$. Получится список $p_i = \{13, 14, 15, 23, 24, 25\}$.

Для каждого из значения необходимо рассчитать координаты *x* и *y* и получить по полученным координатам значение d_j из матрицы расстояний для *«South»*:

$$\{13\}: x = 3; y = 1 \rightarrow d_{13} = 2;$$
$$\{14\}: x = 4; y = 1 \rightarrow d_{14} = 6;$$
$$\{15\}: x = 5; y = 1 \rightarrow d_{15} = 6;$$
$$\{23\}: x = 3; y = 2 \rightarrow d_{23} = 1;$$
$$\{24\}: x = 4; y = 2 \rightarrow d_{24} = 5;$$
$$\{25\}: x = 5; y = 2 \rightarrow d_{25} = 5.$$

Из полученных d_j берется минимальное ($d_{23} = 1$) и сравнивается с параметром *distance*. Так как d_{23} не превышает *distance*, то рассматриваемая клетка 13 удовлетворяет ограничению и не удаляется из домена. Если бы параметр *distance* был 0 или 2+, то ограничение не удовлетворялось бы и клетка 13 удалилась бы из домена. После проверки всех значений из домена *W* в домене останутся только те значения, которые удовлетворяют ограничению (рис. 8).

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89

Рис. 8. Домен переменной *W* после распространения

Ограничение NearConstraint

У данного ограничения два параметра: список из пары объектов и максимальное расстояние между ними (*distance*).

После того, как одна из переменных ограничения определила значение (т. е. объект оказался поставлен в пространстве), необходимо для второго объекта выделить область пространства, в которую можно его поставить, соблюдая ограничение на максимальное расстояние. Для области вычисляются координаты четырех углов:

$X_d = p \ MOD \ width_g$;
$Y_d = p \ DIV \ width_g$;
$X_{left} = X_d - width_u - distance$, если $X_{left} \leq 0$, то $X_{left} = 1$;
$X_{right} = X_d + width_d + distance$, если $X_{right} \geq width_g$, то $X_{right} = width_g - width_u$;
$Y_{top} = Y_d - height_u - distance$, если $Y_{top} \leq 0$, *то* $Y_{top} = 1$;
$Y_{bottom} = Y_d + height_d + distance$, если $Y_{bottom} \geq height_g$, то $Y_{bottom} = height_g - height_u$.

Обозначения: $width_g$ — ширина пространства; X_d, Y_d — координаты клетки p, в которую поставили объект; $width_d$, $height_d$ — ширина и высота поставленного объекта; $width_u$, $height_u$ — ширина и высота объекта, для которого строится область.

По полученным координатам строится область p_s из клеток:

$$p_s = y \times width_g + x, \text{ где } x \in [X_{left}, X_{right}], \text{ у } \in [Y_{top}, Y_{bottom}].$$

Полученное множество p_s содержит все клетки, которые удовлетворяют ограничению, и из домена переменной удаляются значения, которых нет в p_s.

Пример: есть переменные *W*, *Y*, *Z (*рис. 9) и пространство 10 × 10 с поставленным объектом *W* (рис. 10). Задано ограничение, что объект *W* должен стоять вплотную к объекту *Y* на расстоянии 0, а *Z* должен стоять на расстоянии не больше одной клетки от объекта *W*. Объект *W* был поставлен в клетку 44.

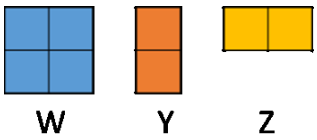


Рис. 9. Объекты *W*, *Y*, *Z*

0	1	2	3	4	5	6	7	8	9
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	19
20	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	29
30	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	39
40	**41**	**42**	**43**	44	45	**46**	**47**	**48**	49
50	**51**	**52**	**53**	54	55	**56**	**57**	**58**	59
60	**61**	**62**	**63**	**64**	**65**	**66**	**67**	**68**	69
70	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	79
80	**81**	**82**	**83**	**84**	**85**	**86**	**87**	**88**	89
90	91	92	93	94	95	96	97	98	99

Рис. 10. Пространство с поставленным объектом *W*

После постановки объекта *W* необходимо определить допустимую область для объектов *Y*, *Z*.

Область для Y:

Координаты области: $X_d = 44\ MOD\ 10 = 4$; $Y_d = 44\ DIV\ 10 = 4$; $X_{left} = 4 - 1 - 0 = 3$; $X_{right} = 4 + 2 + 0 = 6$; $Y_{top} = 4 - 2 - 0 = 2$; $Y_{bottom} = 4 + 2 + 0 = 6$.

Клетки области $p_s = y \times width_g + x$, где $x \in [3, 6]$, у $\in [2, 6]$, показаны на рисунке ниже (рис. 11).

Область для Z:

Координаты области: $X_d = 44\ MOD\ 10 = 4$; $Y_d = 44\ DIV\ 10 = 4$; $X_{left} = 4 - 2 - 1 = 1$; $X_{right} = 4 + 2 + 1 = 7$; $Y_{top} = 4 - 1 - 1 = 2$; $Y_{bottom} = 4 + 2 + 1 = 7$.

Клетки области $p_s = y \times width_g + x$, где $x \in [1, 7]$, у $\in [2, 7]$, показаны на рисунке выше (рис. 12).

Полученные области становятся доменом переменных и могут содержать значения, которые не удовлетворяют ограничению *GDConstaint*, но они будут удалены при распространении ограничения *GDConstaint*.

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99

Рис. 11. Область допустимых значений для переменной *Y*

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99

Рис. 12. Область допустимых значений для переменной *Z*

Программная реализация

Процедура решения рассматриваемой в работе задачи удовлетворения ограничений была реализована с применением библиотеки программирования в ограничениях *Choco* на языке программирования *Java*. Реализация включала формализацию переменных, их доменов и системы ограничений средствами используемой библиотеки. Поиск решений задачи выполнялся стандартными для решателя стратегиями поиска с возвратом, а распространение ограничений осуществлялось с применением оригинальных процедур вывода, разработанных в ходе исследований.

При ветвлении в качестве переменной выбиралась переменная с наименьшим размером домена, а значения переменной выбирались по порядку. Для каждого ограничения был реализован свой класс-распространитель.

Ниже показана общая схема работы программы (рис. 13).

Программа имеет пользовательский интерфейс, внутри которого пользователь создает пространство (рис. 14), объекты (рис. 15) и задает для них ограничения (рис. 16). После на основе полученных данных формируется задача в рамках модели *Choco*.

Первым этапом поиска решений является первичное распространение, которое удовлетворяет унарные ограничения, сокращает домены переменных, сужая пространство поиска, и в некоторых случаях позволяет сразу определить, что решений нет.

Если первоначальное распространение не выявило противоречий в задаче, то запускается ветвление. Среди переменных, у которых не определено значение, выбирается та переменная, у которой размер домена минимален. Значения переменной выбираются по порядку.

После ветвления запускается распространение для сокращения доменов неопределенных переменных. В результате распространения проверяется текущее состояние системы. У каждого ограничения есть параметр *ESat*, который принимает одно из трех значений:

ESat.TRUE — если ограничение удовлетворено;

ESat.FALSE — если ограничение не может быть удовлетворено на данной ветви поиска;

ESat.UNDEFINED — если ограничение пока не удовлетворено.

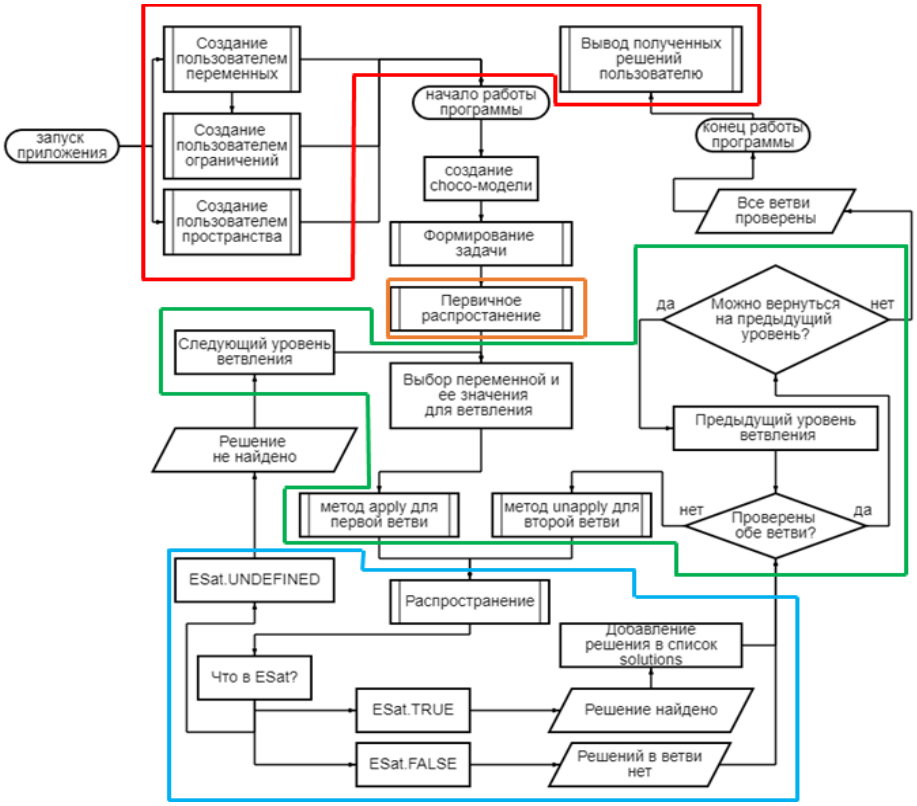


Рис. 13. Общая схема работы программы

Если у всех ограничений *ESat.TRUE*, значит, все объекты расставлены в соответствии со всеми ограничениями и найдено решение. Решение записывается в список и происходит возврат к предыдущему шагу поиска.

Если параметр *ESat* принял значение *UNDEFINED*, то решение пока не найдено и выполняется ветвление.

Если у ограничения в процессе распространения параметр *ESat* становится равен *FALSE*, то в данной ветви решений нет, задача откатывается в предыдущее состояние.

Пользователь может создать пространство (см. рис. 14) двумя способами: ***первый*** — задать ширину и высоту, а программа сама создаст пространство нужной размерности; ***второй*** — вручную описать каждую клетку, задав ей одно из состояний («0» — пустая клетка, «1» — незанимаемая клетка (стена), «2» — клетка, занятая объектом) или считать с файла описанное ранее пространство.

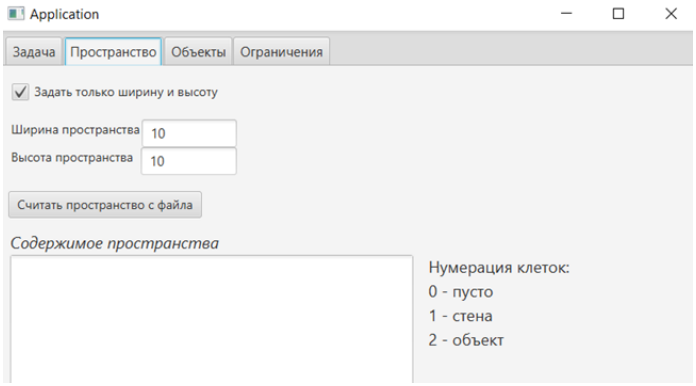


Рис. 14. Пользовательский интерфейс создания пространства

При создании объекта (см. рис. 15) пользователю необходимо задать имя объекта и его содержимое. Под содержимым подразумевается описание каждой клетки объекта его состоянием: «1» — клетка самого объекта; «0» — клетка, которая описывает не сам объект, а клетку, которая должна оставаться свободной для подхода к объекту (например, на рис. 15 показано создание объекта «кровать», для удобного подхода к которому необходимо дополнительно сделать некоторые клетки точно свободными).

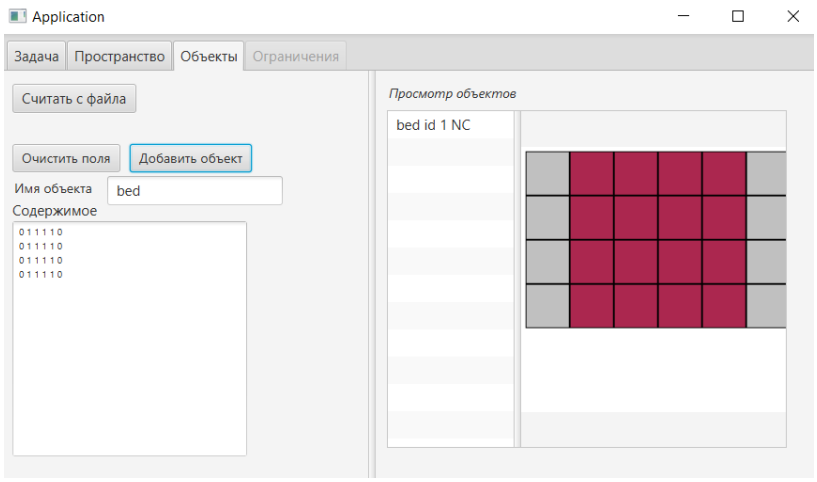


Рис. 15. Пользовательский интерфейс создания объектов

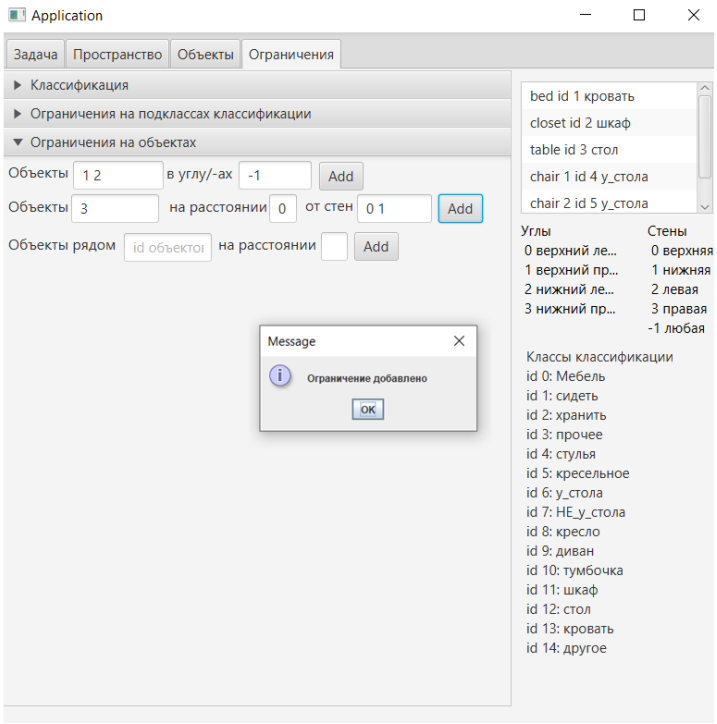


Рис. 16. Пользовательский интерфейс создания ограничений

При создании объектов им будет присвоен уникальный ID, по которому пользователь будет создавать ограничения. Также при создании объектов можно посмотреть их графическое представление. Помимо этого, объекты можно считать с ранее созданного текстового файла.

Для создания ограничений (см. рис. 16) необходимо заполнять текстовые поля:

для ограничения на расположение объекта в углу необходимо задать ID объекта\объектов и код угла\углов, в которых необходимо расположить заданный объект\объекты («0» — верхний левый угол; «1» — верхний правый; «2» — нижний левый; «3» — нижний правый; «-1» — любой);

для ограничения на расположение объекта на заданном расстоянии от заданной стены необходимо задать ID объекта\объектов, расстояние в клетках и код стены («0» — верхняя («*North*»); «1»— нижняя («*South*»); «2» — левая («*West*»); «3» — правая («*East*»); «-1» — любая);

для ограничения на расположение объектов рядом друг с другом необходимо задать ID объектов и максимальное расстояние между ними.

Если выполнение возврата в дереве поиска невозможно, то это означает, что всё дерево поиска пройдено и найдены все решения. После завершения поиска полученные решения выводятся пользователю в интерфейсной части программы (рис. 17).

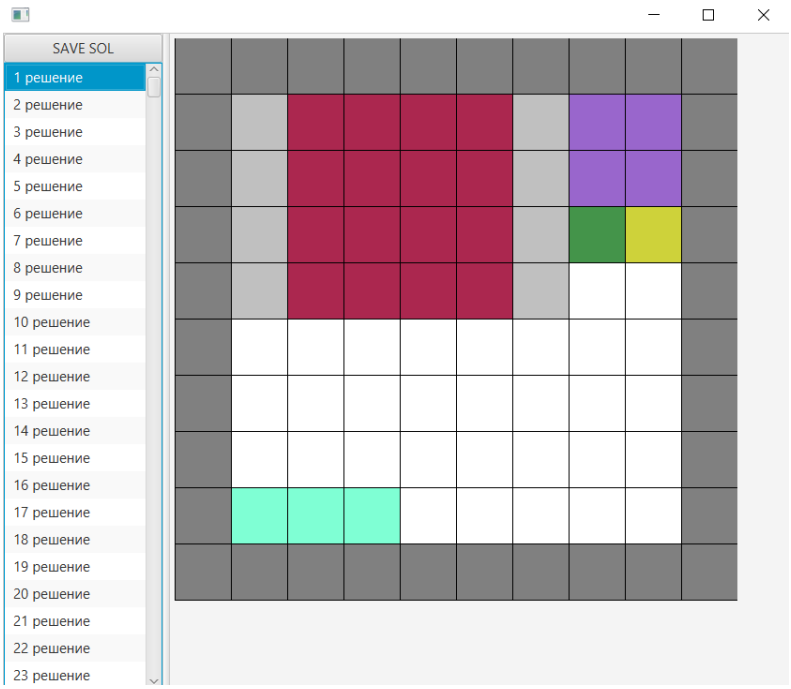


Рис. 17. Пользовательский интерфейс просмотра решений

Ограничения на группы объектов

Также в программе реализована возможность накладывания ограничений не только на отдельные объекты (см. рис. 16), но и на группы объектов (рис. 18).

При создании задачи загружается файл, в котором описывается дерево классификации. Это позволяет при создании объектов отнести их к какому-либо классу и при формировании ограничений задавать ограничения на группу объектов.

Дерево классификации хранится в текстовом файле и загружается пользователем. После загрузки файла в программе появляется графическое представление дерева и пользователю необходимо проклассифицировать объекты, т. е. соотнести с тем или иный классом (см. рис. 18, А). Также каждому классу присваивается свой идентификатор ID, который используется при создании ограничений. Для задания ограничений на подклассах пользователю необходимо в выбранной вкладке заполнить те же поля, что и при создании ограничений для объектов, только вместо ID объекта необходимо указать ID подкласса.

Например, на рис. 18, Б показан процесс создания ограничения, которое можно сформулировать как «все объекты, отнесенные к классу «стул, стоящий у стола» (ID класса 6) должны стоять вплотную (distance = 0) к объекту «стол» (ID объекта 3)».

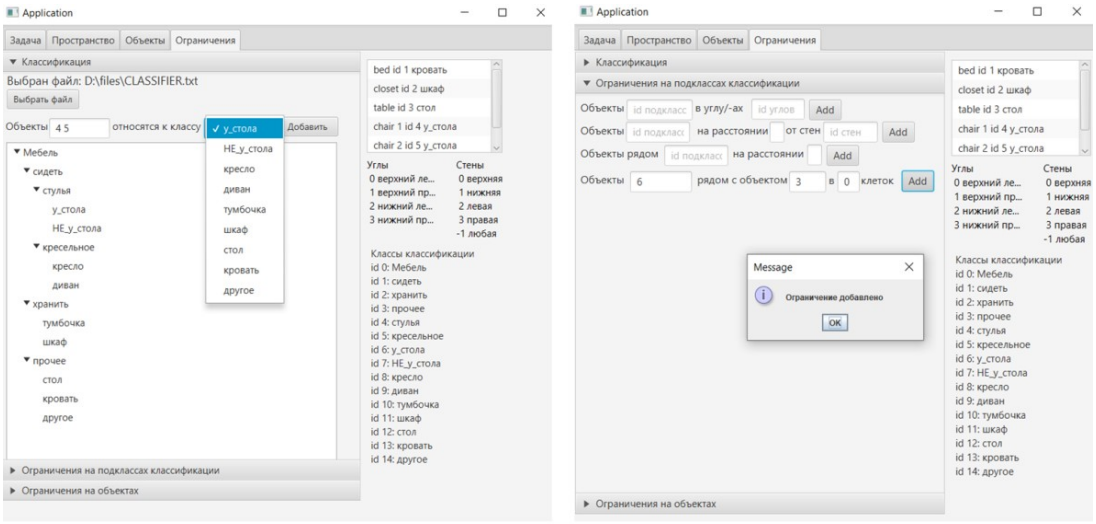


Рис. 18. Пользовательский интерфейс классификации объектов (А) и создания ограничений на подклассах (Б)

Заключение

Парадигма программирования в ограничениях широко применяется для решения задач комбинаторного поиска и комбинаторной оптимизации. Отличительной особенностью данной парадигмы является широкое применение методов рассуждения на ограничениях (методов распространения ограничений).

В работе рассмотрена применимость методов удовлетворения ограничений для решения задач генеративного дизайна. В качестве примера задач рассматриваемого класса в статье подробно разбирается задача проектирования двумерной пространственной среды с учетом требований к взаимному расположению объектов. Подобные требования включают, в частности, запрет на расположение нескольких объектов в одной и той же области двумерной сцены. Также на расположение объекта могут быть наложены ограничения следующих типов: объект должен располагаться в углу; объект должен располагаться на определенном расстоянии до границы пространства; объект должен располагаться на определенном удалении от другого объекта. Соблюдение всех перечисленных ограничений достигается за счет удаления из домена соответствующего объекта клеток пространства, которые не удовлетворяют предъявляемым требованиям. Исключение из доменов недопустимых клеток реализуется разработанными авторами процедурами распространения ограничений. Кроме того, обеспечивается поддержка возможности задавать ограничения не только на отдельные объекты, но и на группы объектов.

Представленные результаты исследований продемонстрировали применимость парадигмы программирования в ограничениях для решения задач генеративного дизайна.

Список источников

1. Krish S. A practical generative design method // Computer-Aided Design. 2011. Vol. 43, № 1. P. 88–100.
2. Mountstephens J., Teo J. Progress and Challenges in Generative Product Design: A Review of Systems // Computers. 2020. Vol. 9. P. 80.
3. Метелик Т. С. Генеративный метод проектирования и способы его реализации в графическом дизайне // Бизнес и дизайн ревью. 2017. Т. 1, № 2 (6). 11 с.
4. Jaisawal R., Agrawal V. Generative Design Method (GDM)—A State of Art // Proceedings of the ICOTRIME 2020 IOP Conf. Series: Materials Science and Engineering. 2021. Vol. 1104, № 012036.
5. Gu N., Behbahani A. P. Shape Grammars: A Key Generative Design Algorithm. Handbook of the Mathematics of the Arts and Sciences. 2021. P. 1385–1405.
6. McCormack J. P., Dorin A. Innocent T. C. Generative design: a paradigm for design research // Futureground. Monash University. 2005. Vol. 2.
7. Lindenmayer A., Rozenberg G. Developmental Systems and languages // Proceedings of the Fourth Annual ACM Symposium on Theory of Computing. New York, NY, USA, Springer International Publishing. Cham, Switzerland, 2012. P. 214–221.

8. Herr C. M., Ford R. C. Cellular automata in architectural design: From generic systems to specific design tools // Automation in Construction. 2016. Vol. 72, Part 1. P. 39–45.
9. Gero J. S., Kazakov V. A. Genetic engineering approach to genetic algorithms // Evol. Comput. 2001. № 9. P. 71–92.
10. Kennedy J., Eberhart R. C., Shi Y. Swarm Intelligence. Elsevier. Amsterdam, The Netherlands, 2001.
11. Родин И. С., Нестерова А. В., Кожевяткина М. А., Кечин Е. С. Применение глубокого обучения в генеративном дизайне: анализ и оптимизация алгоритмов // Наука молодых: Сборник статей по материалам XVI Всероссийской научно-практической конференции. Арзамас; Нижний Новгород: Нижегородский государственный технический университет им. Р. Е. Алексеева, 2024. С. 235–240.
12. Chao Q., Ren T., Wenjing Y. An adaptive artificial neural network-based generative design method for layout designs // International Journal of Heat and Mass Transfer. 2022. P. 1–31.
13. Poole D., Mackworth A. Artificial Intelligence: Foundations of Computational Agents. Cambridge University Press, 2023. 903 p.

References

1. Krish S. A practical generative design method. *Computer-Aided Design*, 2011, vol. 43, no 1, pp. 88–100.
2. Mountstephens J., Teo J. Progress and Challenges in Generative Product Design: A Review of Systems. *Computers*, 2020, vol. 9, p. 80.
3. Metelik T. S. Generativnyy metod proyektirovaniya i sposoby yego realizatsii v graficheskom dizayne [Generative design method and ways of its implementation in graphic design]. *Biznes i dizayn rev'yu* [Business and design review], 2017, vol. 1, no 2 (6), 11 p. (In Russ.).
4. Jaisawal R., Agrawal V. Generative Design Method (GDM)—A State of Art. *Proceedings of the ICOTRIME 2020 IOP Conf. Series: Materials Science and Engineering*, 2021, vol. 1104, no 012036.
5. Gu N., Behbahani A. P. *Shape Grammars: A Key Generative Design Algorithm*. Handbook of the Mathematics of the Arts and Sciences, 2021, pp. 1385–1405.
6. McCormack J. P., Dorin A. Innocent T. C. Generative design: a paradigm for design research. *Futureground*. Monash University, 2005, vol. 2.
7. Lindenmayer A., Rozenberg G. Developmental Systems and languages. *Proceedings of the Fourth Annual ACM Symposium on Theory of Computing*. New York, NY, USA, Springer International Publishing. Cham, Switzerland, 2012, pp. 214–221.
8. Herr C. M., Ford R. C. Cellular automata in architectural design: From generic systems to specific design tools. *Automation in Construction*, 2016, vol. 72, Part 1, pp. 39–45.
9. Gero J. S., Kazakov V. A. Genetic engineering approach to genetic algorithms. Evol. *Comput*., 2001, no 9, pp. 71–92.
10. Kennedy J., Eberhart R. C., Shi Y. *Swarm Intelligence*. Elsevier. Amsterdam, The Netherlands, 2001.
11. Rodin I. S., Nesterova A. V., Kozhevyatkina M. A., Kechin Ye. S. Primeneniye glubokogo obucheniya v generativnom dizayne: analiz i optimizatsiya algoritmov [Application of deep learning in generative design: analysis and optimization of algorithms]. *Nauka molodykh: Sbornik statey po materialam XVI Vserossiyskoy nauchno-prakticheskoy konferentsii* [Science of the young: Collection of articles based on the materials of the XVI All-Russian scientific and practical conference]. Arzamas, Nizhniy Novgorod, Nizhegorodskiy gosudarstvennyy tekhnicheskiy universitet im. R. Ye. Alekseyeva [Arzamas, Nizhny Novgorod, Nizhny Novgorod State Technical University named after R. E. Alekseev], 2024, pp. 235–240. (In Russ.).
12. Chao Q., Ren T., Wenjing Y. An adaptive artificial neural network-based generative design method for layout designs. *International Journal of Heat and Mass Transfer*, 2022, pp. 1–31.
13. Poole D., Mackworth A. *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, 2023, 903 p.

Информация об авторах

П. В. Таран — стажер-исследователь;
А. А. Зуенко — кандидат технических наук, ведущий научный сотрудник.

Information about the authors

P. V. Taran — Intern Researcher;
A. A. Zuenko — Candidate of Science (Tech.), Leading Researcher.

Статья поступила в редакцию 01.11.2025; одобрена после рецензирования 24.11.2025; принята к публикации 28.11.2025.
The article was submitted 01.11.2025; approved after reviewing 24.11.2025; accepted for publication 28.11.2025.

Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 131–139.
Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2025. Vol. 16, No. 3. P. 131–139.

Научная статья
УДК 004.942, 62-9
doi:10.37614/2949-1215.2025.16.3.009

CFD-МОДЕЛЬ АГРЕГИРОВАНИЯ ФЕРРОМАГНИТНЫХ ЧАСТИЦ ПОД ДЕЙСТВИЕМ МАГНИТНОГО ПОЛЯ В ВОСХОДЯЩЕМ ВОДНОМ ПОТОКЕ

Валерий Валентинович Бирюков[1]✉, Андрей Григорьевич Олейник[2]
[1, 2]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[1]v.biryukov@ksc.ru✉, https://orcid.org/0000-0002-0495-2928
[2]a.oleynik@ksc.ru, https://orcid.org/0000-0002-7612-5999

Аннотация

В работе рассматривается использование методов CFD-моделирования для исследования процессов разделения тонких ферромагнитных частиц в восходящих водных потоках с наложением слабонеоднородных магнитных полей. На основе эмпирических данных получена зависимость снижения гидродинамического сопротивления агрегатов частиц в водных потоках. Создана имитационная модель течения ферромагнитной суспензии в рабочем объеме разделительного аппарата, прогнозирующая процесс пространственной сегрегации частиц с учетом агрегирования. Результаты могут быть использованы для разработки эффективных технологий и аппаратов переработки минерального сырья.

Ключевые слова:

CFD-моделирование, UDF-модуль, ферромагнитная суспензия, разделительный аппарат, агрегирование, пространственная сегрегация

Благодарности:

работа выполнена в рамках НИР «Методы и информационные технологии мониторинга и управления региональными критическими инфраструктурами Арктической зоны Российской Федерации» (проект № FMEZ-2025-0054).

Для цитирования:

Бирюков В. В., Олейник А. Г. CFD-модель агрегирования ферромагнитных частиц под действием магнитного поля в восходящем водном потоке // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 131–139. doi:10.37614/2949-1215.2025.16.3.009.

Original article

CFD-MODEL FOR FERROMAGNETIC PARTICLE AGGREGATION UNDER THE MAGNETIC FIELD INFLUENCE IN AN ASCENDING WATER FLOW

Valery V. Biryukov[1]✉, Andrey A. Oleynik[2]
[1, 2]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[1]v.biryukov@ksc.ru✉, https://orcid.org/0000-0002-0495-2928
[2]a.oleynik@ksc.ru, https://orcid.org/0000-0002-7612-5999

Abstract

The use of CFD modeling methods to study the separation of fine ferromagnetic particles in ascending water flows with the application of weakly inhomogeneous magnetic fields is considered in the paper. A dependence for the reduction in hydrodynamic drag of particle aggregates in water flows is obtained on the base of empirical data. A simulation model of the ferromagnetic suspension flow in the working volume of a separation apparatus is created, which predict the process of spatial segregation of particles taking into account aggregation. The results can be used to develop efficient technologies and equipment for processing mineral raw materials.

Keywords:

CFD modeling, UDF module, ferromagnetic suspension, separation apparatus, aggregation, spatial segregation

Acknowledgements:

The study was carried out within the framework of the research project “Methods and information technologies for monitoring and managing regional critical infrastructures of the Arctic zone of the Russian Federation” (project No. FMEZ-2025-0054).

For citation:
Biryukov V. V., Oleynik A. A. CFD-model for ferromagnetic particle aggregation under the magnetic field influence in an ascending water flow. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 131–139. doi:10.37614/2949-1215.2025.16.3.009.

Введение

От устойчивого и эффективного функционирования предприятий горнопромышленного комплекса во многом зависит благополучие социально-экономических систем целого ряда регионов Российской Федерации. На северо-западе страны такими регионами являются Мурманская область и Республика Карелия [1]. Совершенствование используемых технологий добычи и первичной переработки полезных ископаемых способствует не только обеспечению конкурентоспособности этих предприятий в условиях истощения запасов богатых руд, но и снижению оказываемой этими предприятиями нагрузки на окружающую среду. Для разработки и исследования потенциальных возможностей новых, более эффективных технологических процессов и аппаратов широко применяются различные методы моделирования и вычислительные эксперименты. В частности, Горным институтом Кольского научного центра Российской академии наук проводятся конференции под названием «Цифровые технологии в горном деле», посвященные моделированию объектов и процессов горного производства, технологиям решения задач эффективного извлечения полезных компонентов из минерального сырья, а также решению экономических и экологических проблем горной отрасли. Доклады, представленные на конференции, публикуются в специальных выпусках профильных периодических изданий [2; 3]. Наряду с прикладными решениями в области информационно-аналитической поддержки деятельности горнопромышленных предприятий, основывающими на известных и апробированных математических методах и моделях, возникают задачи, для которых эти методы не позволяют получить эффективное с точки зрения используемых временных и вычислительных ресурсов решение. Это обуславливает необходимость разработки и использования новых подходов и информационных технологий получения результата. В качестве примера можно отметить развиваемые в Институте информатики и математического моделирования КНЦ РАН технологии на основе парадигмы удовлетворения ограничений [4; 5].

Практически на всех предприятиях по переработке минерального сырья реализуются процессы снижения крупности минеральных частиц для раскрытия их физических и физико-химических свойств с последующим разделением компонентов суспензий в продукты с контрастными свойствами в силовых полях различной природы. Действие массовых сил, таких как гравитационная, центробежная или магнитная, приводит к пространственной сегрегации частиц суспензий и последующему выводу их из пространства разделения. В случае тонкой вкрапленности полезных минералов в рудной массе возникает необходимость уменьшения крупности минеральных частиц до размеров менее $5 \cdot 10^{-5}$ м, что приводит к снижению контрастности их физических и физико-химических свойств и, как следствие, к снижению относительных скоростей их движения в жидкости и взаимному засорению продуктов разделения. Интенсификация разделительных процессов для тонких минеральных частиц возможна с использованием эффектов их агрегирования под действием химических реагентов или магнитных и электрических полей [6].

Агрегирование тонких минеральных частиц позволяет изменять скорость их осаждения в жидкостях. Воздействие химических реагентов, намагничивание и электризация частиц делают возможным объединение отдельных тонких частиц в пространственные агрегаты — флоккулы, имеющие более высокие скорости движения под действием физических сил в жидкой среде.

Моделирование движения компонентов минеральных суспензий в пространстве разделения реализуется активно развивающимися в настоящее время методами вычислительной гидродинамики (CFD) на основе теории многофазного многоскоростного континуума (ММК), основы которой разработал Р. И. Нигматулин [7; 8], с использованием специализированного программного обеспечения [9; 10]. Однако, несмотря на богатый арсенал моделей и средств их реализации, они не обеспечивают учета всей полноты физических и физико-химических эффектов взаимодействия

компонентов многофазных сред, влияющих на их пространственную сегрегацию. В настоящей работе предлагается модель агрегирования тонких ферромагнитных частиц под действием магнитных полей. Данная модель позволяет адекватно учитывать эффекты агрегирования частиц в ферромагнитных суспензиях при моделировании процессов их разделения.

Базовая CFD-модель многофазных многокомпонентных сред

Континуальный подход к моделированию многокомпонентной смеси заключается в представлении каждого компонента суспензии как отдельной сжимаемой псевдожидкости, движущейся в едином объеме с другими компонентами под действием массовых и поверхностных физических сил. Модель движения сжимаемой псевдожидкости называется моделью Эйлера и состоит из комплекта уравнений, включающих уравнения законов сохранения массы (уравнения непрерывности), количества движения и энергии по необходимости. Различаются Эйлер-Эйлеровы и Эйлер-Лагранжевы модели [11–13]. В Эйлер-Эйлеровых моделях сжимаемыми псевдожидкостями являются общий континуум и континуумы отдельных компонентов. В Эйлер-Лагранжевых моделях движение компонентов моделируется уравнениями динамики отдельных частиц под действием физических сил, что позволяет исследовать их траектории.

CFD-модель движения многокомпонентной смеси в пространстве разделения состоит из комплекта Эйлеровых уравнений сохранения для всех q компонентов:

1. Уравнения непрерывности для q компонента:

$$\frac{\partial}{\partial t}\alpha_q\rho_q + \nabla\cdot\left(\alpha_q\rho_q\vec{u}_q\right) = \sum_{p=1}^{n}\dot{m}_{pq},$$

где α_q, ρ_q $\vec{u}_q$ — удельный объем, плотность и скорость q компонента; $\sum_{p=1}^{n}\dot{m}_{pq}$ — источниковый член.

2. Уравнение баланса количества движения для q компонента:

$$\frac{\partial}{\partial t}\left(\alpha_q\rho_q\vec{u}_q\right) + \nabla\cdot\left(\alpha_q\rho_q\vec{u}_q\vec{u}_q\right) = -\alpha_q\nabla p + \alpha_q\rho_q\vec{g} + \nabla\cdot\bar{\bar{\tau}}_q + \sum_{p=1}^{n}\left(\vec{R}_{pq} + \dot{m}_{pq}\vec{u}_q\right) + \alpha_q\rho_q\left(\vec{F}_q + \vec{F}_{lift,q} + \vec{F}_{vm,q}\right),$$

где ∇p — градиент давления; $\nabla\cdot\bar{\bar{\tau}}_q$ — напряжения в сплошной среде; $\vec{R}_{pq}$ — межфазные, межкомпонентные взаимодействия; $\vec{F}_q + \vec{F}_{lift,q} + \vec{F}_{vm,q}$ — внешние силы, подъемная сила и виртуальная массовая сила.

Уравнения обмена импульсом между отдельными сжимаемыми p и q компонентами имеют следующий вид:

$$\vec{R}_{pq} = K_{sl}\left(\vec{u}_p - \vec{u}_q\right),$$

где $K_{sl} = \frac{\alpha_s\rho_s f}{\tau_s}$ — коэффициент обмена между твердой (s) и жидкой (l) фазами; $\tau_s = \frac{\rho_s d_s^2}{18\mu_l}$ — время релаксации — время, за которое твердая частица диаметром d_s и плотностью ρ_s приобретает расчетную скорость $\vec{u}_p$ в среде с динамической вязкостью μ_l.

Разделение компонентов ферромагнитной суспензии происходит в гидродинамических условиях плотного кипящего слоя, что является основанием выбора уравнений Huilin & Gidaspow [14] для моделирования обмена импульсом между твердыми и жидкими компонентами:

$$\mathrm{K_{sl}} = \frac{3}{4}\mathrm{C_D}\frac{\alpha_\mathrm{s}\alpha_\mathrm{l}\rho_\mathrm{l}|\vec{\mathrm{v}}_\mathrm{s} - \vec{\mathrm{v}}_\mathrm{l}|}{\mathrm{d_s}}\alpha_\mathrm{l}^{-2.65},$$

где $\mathrm{C_D}$ — коэффициент сопротивления среды ($\mathrm{C_D} = \frac{24}{\alpha_\mathrm{l}\mathrm{Re_s}}[1 + 0.15(\alpha_\mathrm{l}\mathrm{Re_s})^{0.687}]$); α_l, α_s — объемные фракции жидкой и частиц твердой фазы; Re — числа Рейнольдса.

Модель Huilin & Gidaspow пригодна как для плотных кипящих слоев с объемной долей жидкости $\alpha_l < 0.8$, так и для разреженных суспензий с долей объема твердого вещества меньше 0.2.

$$\text{Для } \alpha_l < 0.8: \qquad \mathrm{K_{sl}} = 150\frac{\alpha_s(1-\alpha_l)\mu_l}{\alpha_l d_s^2} + 1.75\frac{\rho_l \alpha_s |\vec{v}_s - \vec{v}_l|}{d_s};$$

$$\text{Для } \alpha_l < 0.2: \qquad \mathrm{K_{sl}} = \psi \mathrm{K_{sl-Ergun}} + (1-\psi)\mathrm{K_{sl-Wen\&Yu}}.$$

Плавный переход между состояниями с различной концентрацией частиц описывается функцией-«переключателем» ψ:

$$\psi = \frac{1}{2} + \frac{\arctan\left(262.5(\alpha_s - 0.2)\right)}{\pi}.$$

Для моделирования обмена импульсами между отдельными дисперсными твердыми компонентами используется уравнение [15]:

$$\mathrm{K_{hs}} = \frac{3(1+e_{hs})\left(\frac{\pi}{2} + C_{fr,hs}\frac{\pi^2}{8}\right)\alpha_s \rho_s \alpha_h \rho_h (d_h + d_s)^2 g_{0,hs}}{2\pi\left(\rho_h d_h^3 + \rho_s d_s^3\right)} |\vec{v}_h - \vec{v}_s|,$$

где e_{hs} — коэффициент восстановления; $C_{fr,hs}$ — коэффициент трения; d_h — диаметр твердых частиц фракции h; $g_{0,hs}$ — радиальный коэффициент распределения.

Течение жидкости в различных частях разделительного аппарата числа Рейнольдса варьируются в достаточно широком диапазоне значений $Re = 2 \div 4000$, что требует использования стандартной полуэмпирической модели турбулентности k-ε, основанной на уравнениях переноса для кинетической энергии турбулентности (k) и ее скорости диссипации (ε) [16], и имеющей достаточную точность при моделировании потоков переходного и турбулентного режимов.

Моделирование преобразования кинетической энергии потока жидкости в кинетическую энергию турбулентности и ее рассеяния проводится на основе уравнений переноса:

$$\frac{\partial}{\partial t}(\rho k) + \frac{\partial}{\partial x_i}(\rho k u_i) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_j}\right] + G_k + G_b - \rho\varepsilon - Y_M + S_k;$$

$$\frac{\partial}{\partial t}(\rho \varepsilon) + \frac{\partial}{\partial x_i}(\rho \varepsilon u_i) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_j}\right] + C_{1\varepsilon}\frac{\varepsilon}{k}(G_k + C_{3\varepsilon}G_b) - C_{2\varepsilon}\rho\frac{\varepsilon^2}{k} + S_\varepsilon,$$

где G_k — скорость перехода кинетической энергии потока в кинетическую энергию турбулентности; G_b — скорость перехода кинетической энергии потока при естественной конвекции (при наличии гравитационного потенциала); k — кинетическая энергия турбулентности; ε — скорость турбулентной диссипации; σ_ε — турбулентное число Прандтля; $C_{1\varepsilon}$, $C_{2\varepsilon}$, $C_{3\varepsilon}$, — константы модели турбулентности; μ_t — турбулентная вязкость.

Приведенный выше математический аппарат относится к традиционному описанию многокомпонентных смесей в рамках Эйлер-Эйлерова подхода и имеет реализацию в программных комплексах CFD-моделирования.

Модель учета агрегирования частиц в магнитном поле

Отличительной особенностью предлагаемой модели является возможность учета эффектов агрегирования частиц в магнитном поле.

При попадании в вертикальное слабонеоднородное магнитное поле ферромагнитные частицы агрегируются в виде цепочек, ориентированных вдоль силовых линий (рис. 1). В разрабатываемой CFD-модели разделения частиц суспензии по их гравитационным и магнитным свойствам агрегирование частиц предлагается рассматривать как процесс массопереноса частиц одной псевдожидкой Эйлеровой фазы в частицы другой Эйлеровой фазы.

Уравнения непрерывности для массопереноса между фазой p и фазой q в прямом и обратном направлениях: $m_p = -m_{p^i q^j}$ и $m_q = m_{p^i q^j}$.

Уравнения передачи импульса: $m_p \vec{u}_p = -m_{p^i q^j}\vec{u}_p$ и $m_q \vec{u}_q = m_{p^i q^j}\vec{u}_p$.

Вновь образованные Эйлеровы фазы взаимодействуют в объеме с другими фазами и между собой. При этом уравнения модели Huilin & Gidaspow обмена импульсом между жидкими и твердыми компонентами должны быть модифицированы.

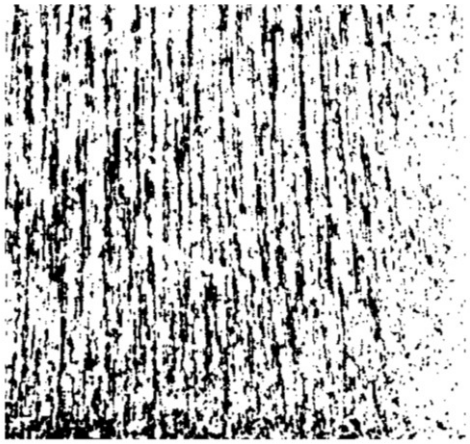


Рис. 1. Фотография вертикальных агрегатов магнетитовых частиц в ожиженном слое суспензии в магнитном поле

Модель Huilin & Gidaspow разработана на основе представлений о фильтрации жидкости через псевдоожиженный слой частиц [17]. В любом псевдоожиженном слое уравновешены вес слоя, отнесенный к единице поверхности, и перепад давления Δp.

Величина перепада давления определяется зависимостью:

$$\Delta p = \lambda \frac{1}{d_э} \cdot \frac{\rho \omega^2}{2}, \omega = \frac{\omega_0}{\varepsilon},$$

где λ — общий коэффициент сопротивления, включающий сопротивление трения и сумму дополнительных местных сопротивлений; $d_э$ — эквивалентный диаметр, соответствующий суммарному поперечному сечению каналов в зернистом слое; ω, ω_0 — действительная и фиктивная скорости фильтрации жидкости в слое соответственно; ε — порозность слоя.

Для учета влияния формы частиц вводится фактор формы Ф (для шарообразных частиц Ф = 1) и d — диаметр эквивалентного шара, имеющего тот же объем, что и частица. В результате выражение для перепада давления принимает следующий вид:

$$\Delta p = \frac{3(1-\varepsilon)}{2\varepsilon^3 \Phi} \lambda \frac{H}{d} \cdot \frac{\rho \omega_0^2}{2}.$$

Коэффициент сопротивления λ зависит от гидродинамического режима течения, определяемого значением числа Рейнольдса $Re = \frac{4W}{\alpha \mu}$, где W — массовая скорость жидкости, отнесенная к 1 м^2 сечения аппарата, кг/(м^2·сек).

Для расчета коэффициента сопротивления λ ожиженного слоя имеется эмпирическая зависимость, работающая при всех числах Рейнольдса: $\lambda = \frac{133}{Re} + 2{,}34$.

Результатом агрегирования частиц при воздействии магнитного поля является снижение коэффициента сопротивления, то есть λ является функцией от напряженности магнитного поля:

$$\lambda(H) = \left(\frac{133}{Re} + 2{,}34\right) \cdot F(H_m).$$

Приравнивая перепад давления к весу псевдоожиженного слоя, получаем выражение, модифицирующее модель Huilin & Gidaspow для компонентов ферромагнитной суспензии, агрегированной под действием магнитного поля:

$$\Delta p = \frac{3}{2} \cdot \frac{1-\varepsilon}{\varepsilon^3 \Phi} \left(\frac{133}{Re} + 2{,}34 \right) \frac{H}{d} \cdot \frac{\rho \omega_0^2}{2} F(H_m);$$

$$\Delta p = \left(150 \frac{(1-\varepsilon)^2 \mu H \omega_0}{\varepsilon^3 \Phi^2 d^2} + 1{,}75 \frac{1-\varepsilon}{\varepsilon^3 \Phi} H \frac{\rho \omega_0^2}{d} \right) \cdot F(H_m).$$

Для определения параметров множителя $F(H_m)$ была проведена серия лабораторных экспериментов. В экспериментах измерялась высота псевдоожиженного слоя ферромагнитных частиц узкого класса крупности в вертикально ориентированном магнитном поле различной напряженности. На основе полученных в экспериментах данных (рис. 2) с использованием программы MathCad15 был определен вид зависимости:

$$F(H_m) = e^{-3.77 \cdot 10^{-5} \cdot Hm}$$, где H_m измеряется в А/м.

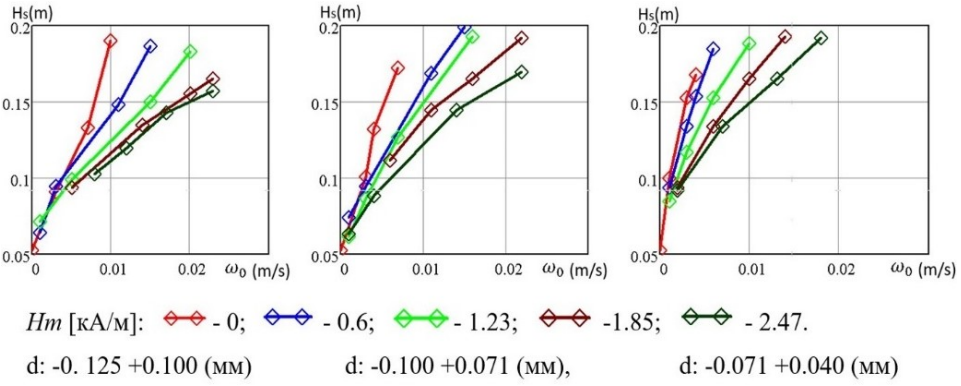


Рис. 2. Зависимости высоты слоя H_s[м] ферромагнитных частиц разной крупности от фиктивной скорости псевдоожижения в магнитном поле с различной напряженностью H_m [кА/м]

По результатам проведенных исследований на языке С++ был создан программный модуль для расчета $F(H_m)$. Интеграция данного модуля с уже существующими в среде моделирования ANSYS Fluent инструментами позволяет получить имитационные модели процессов сепарации, происходящих в разделительных аппаратах, учитывающие влияние на эти процессы агрегирования тонких ферромагнитных частиц под действием магнитных полей. Схема проведения вычислительных экспериментов с такими моделями с указанием используемых инструментальных средств реализации показана на рис. 3.

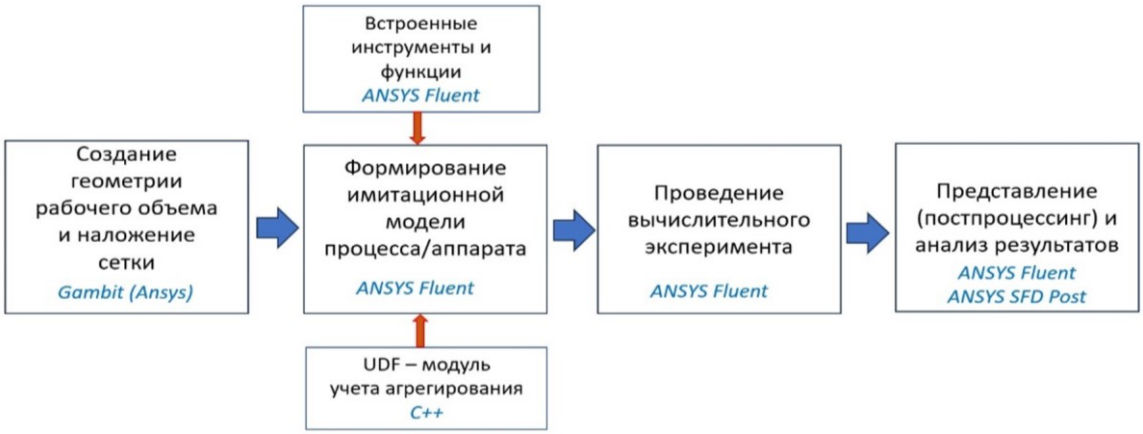


Рис. 3. Укрупненная схема проведения вычислительного эксперимента с использованием модуля учета агрегирования

Наряду со стандартными функциями учета пристеночных взаимодействий, расчета числа Рейнольдса Re и времени релаксации τ_s, разработанный модуль включает функцию расчета коэффициента межфазного взаимодействия с учетом агригирования частиц в магнитном поле:

```
/************************************************************************
 UDF for customizing the default Syamlal drag law in ANSYS Fluent
 ************************************************************************/
 #include "udf.h"
 #define pi 4.*atan(1.)
 #define diam2 3.e-4
 DEFINE_EXCHANGE_PROPERTY(custom_drag,cell,mix_thread,s_col,f_col)
 {
…..
. /*compute drag and return drag coeff, k_g_s*/
   afac = pow(void_g,4.14);
   if(void_g <= 0.85)
      bfac = 0.281632*pow(void_g, 1.28);
   else
      bfac = pow(void_g, 9.076960);
    vfac = 0.5*(afac-0.06*reyp+sqrt(0.0036*reyp*reyp+0.12*reyp*(2.*bfac-
         afac)+afac*afac));
    fdrgs = void_g*(pow((0.63*sqrt(reyp)/
          vfac+4.8*sqrt(vfac)/vfac),2))/24.0;
    k_g_s = (1.-void_g)*pow(2.718,-0.0000377*H_magn)*rho_s*fdrgs/taup;
   return k_g_s;
}
```

Заключение

В настоящей статье рассмотрен случай агрегирования ферромагнитных частиц под действием вертикально ориентированного слабонеоднородного магнитного поля в восходящем водном потоке. Кроме процессов разделения ферромагнитных компонентов в постоянных магнитных полях эффекты агрегирования имеют место в процессах, основанных на других физических и физико-химических принципах: например, флокуляция мелких частиц под действием химических реагентов, агрегирование при контактной электризации или кратковременном воздействии магнитного поля. Представленные в настоящей работе результаты могут рассматриваться как один из блоков имитационных моделей изменения физических и физико-химических свойств высокодисперсных компонентов минеральных суспензий в объемах разделительных аппаратов. В ходе дальнейших исследований предполагается расширить спектр моделей, позволяющих учитывать эффекты агрегирования и в других процессах. Ожидаемые результаты ориентированы на исследование и разработку более эффективных технологий и аппаратов переработки минерального сырья для повышения конкурентоспособности предприятий горнопромышленного комплекса.

Список источников

1. Интернет-портал СНГ: Горнорудная промышленность России и СНГ. Профессиональная дискуссия для развития отрасли [Электронный ресурс]. URL: https://e-cis.info/news/566/118268/ (дата обращения: 08.10.2025).
2. Цифровые технологии в горном деле: Горный информационно-аналитический бюллетень (научно-технический журнал). 2019. № 11 (специальный выпуск 37). 664 с.
3. Горная промышленность. 2023. № S5. 149 с.
4. Зуенко А. А. Олейник Ю. А. Составление расписаний как задача удовлетворения ограничений (на примере планирования открытых горных работ) // Информатика и автоматизация. 2024. Т. 23, № 5. С. 1290–1310.

5. Зуенко А. А., Олейник А. Г., Олейник Ю. А. Интеграция концептуального моделирования и программирования в ограничениях для синтеза схем технологических процессов // Информационные технологии и вычислительные системы. 2022. № 3. С. 95–107.
6. Бирюков В. В., Никитин Р. М., Скороходов В. Ф., Степанникова А. С. Использование элементов теории популяционного баланса для интенсификации процессов агрегирования тонкодисперсных частиц // Современные проблемы переработки труднообогатимых руд и техногенного сырья (Плаксинские чтения 2017): Материалы Международной конференции. Красноярск, 12–15 сентября 2017 г. Красноярск: Сибирский федеральный университет, 2017. С. 234–236.
7. Нигматулин Р. И. Динамика многофазных сред. М.: Наука. Гл. ред. физ.-мат. лит., 1987. Ч. 1. 464 с.
8. Zhu C, Fan L-S, Yu Z. Continuum Modeling of Multiphase Flows. In: Dynamics of Multiphase Flows. Cambridge Series in Chemical Engineering. Cambridge University Press, 2021. P. 212–260.
9. Моделирование и цифровые двойники: Портал [Электронный ресурс]. URL: https://www.cadfem-cis.ru/products/ansys/fluids/fluent/ (дата обращения: 09.10.2025).
10. АО «Ай-Джи-Эй Технологии». ЛОГОС Аэро-Гидро [Электронный ресурс]. URL: https://igatec.com/software-logos/logos-aero-gidro/ (дата обращения: 09.10.2025).
11. Loha, C., Chattopadhyay, H., & Chatterjee, P. K. Euler-Euler CFD modeling of fluidized bed: Influence of specularity coefficient on hydrodynamic behavior // Particuology. 2013. 11 (6). P. 673–680.
12. Hamidouche, Z., Dufresne, Y., Pierson, J., et al. DEM/CFD simulations of a pseudo-2D fluidized bed: Comparison with experiments // Fluids. 2019. 4 (1). 51. [Электронный ресурс]. URL: https://www.mdpi.com/2311-5521/4/1/51. (дата обращения: 12.10.2025).
13. Ku, X., Li, T., & Lovas, T. Influence of drag force correlations on periodic fluidization behavior in Eulerian-Lagrangian simulation of a bubbling fluidized bed // Chemical Engineering Science. 2013. Vol. 95. P. 94–106.
14. Huilin L. and Gidaspow D. Hydrodynamics of binary fluidization in a riser: CFD simulation using two granular temperatures // Chemical Engineering Science. 2003. V. 58, Issue 16. P. 3777–3792.
15. Syamlal M. The Particle-Particle Drag Term in a Multi-Particle Model of Fluidization. 1987. URL: https://www.researchgate.net/publication/241929833TheParticle-Particle_Drag_Term_in_a_Multiparticle_Model_of_Fluid ization (дата обращения: 02.10.2025).
16. Launder B. E. and Spalding D. B. Lectures in Mathematical Models of Turbulence. Academic Press, London, England. 1972. 169 p.
17. Jackson R. The Dynamics of Fluidized Particles. Cambridge Monographs on Mechanics, 2000. 352 p.

References

1. Internet-portal SNG: Gornorudnaya promyshlennost[1] Rossii i SNG. Professional'naya diskussiya dlya razvitiya otrasli [CIS Internet Portal: Mining Industry of Russia and the CIS. Professional Discussion for Industry Development]. Available at: https://e-cis.info/news/566/118268/ (accessed 08.10.2025) (In Russ.).
2. *Tsifrovyye tekhnologii v gornom dele: Gornyy informatsionno-analiticheskiy byulleten' (nauchno-tekhnicheskiy zhurnal)* [Digital Technologies in Mining: Mining Information and Analytical Bulletin (Scientific and Technical Journal)], 2019, no. 11 (special issue 37), 664 p. (In Russ.).
3. *Gornaya promeshlennost'* [Russian Mining Industry], 2023, no 5S, 149 p. (In Russ.).
4. Zuenko A., Oleynik Yu. Sostavleniye raspisaniy kak zadacha udovletvoreniya ogranicheniy (na primere planirovaniya otkrytykh gornykh rabot) [Scheduling as a Constraint Satisfaction Problem (Using the Example of Open-Pit Mine Production Scheduling Problem)]. *Informatika i avtomatizatsia* [Informatics and Automation], 2024, Vol. 23, no. 5, pp. 1290–1310. (In Russ.).
5. Zuenko A. A., Oleynik A. G., Oleynik Y. A. Integratsiya kontseptual'nogo modelirovaniya i programmirovaniya v ogranicheniyakh dlya sinteza skhem tekhnologicheskikh protsessov [Integration of Conceptual Modeling and Constraint Programming for the Synthesis of Flowsheets]. *Informatsionnyye tekhnologii i vychislitel'nyye sistemy* [Information technology and computing systems], 2022, no. 3, pp. 95–107. (In Russ.).
6. Biryukov V. V., Nikitin R. M., Skorokhodov V. F., Stepannikova A. S. Ispol'zovaniye elementov teorii populyatsionnogo balansa dlya intensifikatsii protsessov agregirovaniya tonkodispersnykh chastits [Using elements of population balance theory to intensify the processes of fine particles aggregation]. *Sovremennyye problemy pererabotki trudnoobogatimykh rud i tekhnogennogo syr'ya (Plaksinskiye chteniya 2017): Materialy Mezhdunarodnoy konferentsii. Krasnoyarsk, 12–15 sentyabrya 2017 g*. [Current Problems of Processing Refractory Ores and Man-Made Raw Materials (Plaksin Readings 2017): Proceedings of the International Conference. Krasnoyarsk, September 12–15, 2017]. Krasnoyarsk, Sibirskiy federal'nyy universitet, 2017, pp. 234–236. (In Russ.).

7. Nigmatulin R. I. Dinamika mnogofaznykh sred [Dynamics of multiphase media]. Moscow, Nauka, Ed. in chief of physical and mathematical literature, 1987, Chast' I, 464 p. (In Russ.).
8. Zhu C, Fan L-S, Yu Z. Continuum Modeling of Multiphase Flows. In: *Dynamics of Multiphase Flows*. Cambridge Series in Chemical Engineering. Cambridge University Press, 2021, pp. 212–260.
9. Portal Modelirovaniye i tsifrovyye dvoyniki [Modeling and Digital Twins Portal]. Available at: https://www.cadfem-cis.ru/products/ansys/fluids/fluent/ (accessed 09.10.2025).
10. AO "Ay-Dzhi-EyTekhnologii". LOGOS Aero-Gidro [IGA Technologies. LOGOS Aero-Hydro]. Available at: https://igatec.com/software-logos/logos-aero-gidro/ (accessed 09.10.2025). (In Russ.).
11. Loha, C., Chattopadhyay, H., & Chatterjee, P. K. Euler-Euler CFD modeling of fluidized bed: Influence of specularity coefficient on hydrodynamic behavior. *Particuology*, 2013, 11 (6), pp. 673–680.
12. Hamidouche, Z., Dufresne, Y., Pierson, J., et al. DEM/CFD simulations of a pseudo-2D fluidized bed: Comparison with experiments. *Fluids*, 2019, 4 (1), 51. Available at: https://www.mdpi.com/2311-5521/4/1/51 (accessed 12.10.2025).
13. Ku, X., Li, T., & Lovas, T. Influence of drag force correlations on periodic fluidization behavior in Eulerian-Lagrangian simulation of a bubbling fluidized bed. *Chemical Engineering Science*, 2013, Vol. 95, pp. 94–106.
14. Huilin L. and Gidaspow D. Hydrodynamics of binary fluidization in a riser: CFD simulation using two granular temperatures. *Chemical Engineering Science*, 2003, V. 58, Issue 16, pp. 3777–3792.
15. Syamlal M. The Particle-Particle Drag Term in a Multi-Particle Model of Fluidization. 1987. Available at: https:\www.researchgate.net\publication\241929833TheParticle-Particle_Drag_Term_in_a_Multiparticle_Model_of_Fluidi zation (accessed 02.10.2025).
16. Launder B. E. and Spalding D. B. *Lectures in Mathematical Models of Turbulence*. Academic Press, London, England, 1972, 169 p.
17. Jackson R. *The Dynamics of Fluidized Particles*. Cambridge Monographs on Mechanics, 2000, 352 p.

Информация об авторах

В. В. Бирюков — ведущий инженер;
А. Г. Олейник — доктор технических наук, главный научный сотрудник.

Information about the authors

V. V. Biryukov — Lead Engineer;
A. G. Oleynik — Doctor of Science (Tech.), Chief Research Fellow.

Статья поступила в редакцию 15.11.2025; одобрена после рецензирования 21.11.2025; принята к публикации 24.11.2025.
The article was submitted 15.11.2025; approved after reviewing 21.11.2025; accepted for publication 24.11.2025.

Научная статья
УДК 004.822, 004.89
doi:10.37614/2949-1215.2025.16.3.010

МОДЕЛИРОВАНИЕ ПРОСТРАНСТВЕННЫХ СИТУАЦИЙ КАК ГЕОСЕМАНТИЧЕСКИХ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ГЕОПРОСТРАНСТВЕННОГО ГРАФА ЗНАНИЙ

***Александр Владимирович Вицентий*[1, 2]✉**
[1]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия, alx_2003@mail.ru✉, https://orcid.org/0000-0003-1331-4749
[2]Филиал Мурманского арктического университета в г. Апатиты, Апатиты, Россия

Аннотация

Моделирование пространственных ситуаций является одним из эффективных средств информационной поддержки принятия решений, в том числе и для задач управления сложными динамическими системами с учетом региональной специфики. В настоящее время популярным инструментом для моделирования пространственных ситуаций являются географические информационные системы. Современные системы такого класса моделируют ситуации с помощью электронных карт, а также предоставляют пользователям набор инструментов для анализа пространственных данных. Одной из проблем данного подхода является сложность учета семантических свойств и связей отображаемых объектов, которые не могут быть напрямую записаны в базу данных географической информационной системы. Еще одна сложность, обусловленная подходом к моделированию ситуаций и способом организации данных, связана с добавлением новой семантической информации об уже визуализированных объектах. Данная работа посвящена решению проблемы автоматизированного построения моделей пространственных ситуаций для решения задач информационной поддержки принятия решений с учетом географических и семантических свойств пространственных объектов. Предложен способ семантического обогащения геопространственных данных на основе графа знаний. Выполнен краткий обзор существующих подходов к использованию графов знаний для формирования картографических изображений. Описаны модель пространственной ситуации и основные компоненты технологии автоматизированного построения моделей пространственных ситуаций в виде геосемантических изображений. Представлены результаты моделирования пространственной ситуации как геосемантического изображения, построенного с использованием картографической основы и геопространственного графа знаний.

Ключевые слова:

геопространственный граф знаний, пространственная ситуация, геосемантическое изображение, ситуационное моделирование, пространственный анализ, системы информационной поддержки принятия решений

Благодарности:

исследование выполнено в рамках государственного задания Института информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук от Министерства науки и высшего образования Российской Федерации, тема научно-исследовательской работы «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределенности и риска» (регистрационный номер 1023032300374-0-2.2.1).

Для цитирования:

Вицентий А. В. Моделирование пространственных ситуаций как геосемантических изображений на основе геопространственного графа знаний // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 140–153. doi:10.37614/2949-1215.2025.16.3.010.

Original article

MODELING OF SPATIAL SITUATIONS AS GEOSEMANTIC IMAGES BASED ON THE GRAPH OF GEOSPATIAL KNOWLEDGE

***Alexander V. Vicentiy*[1, 2]✉**
[1]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia, alx_2003@mail.ru✉, https://orcid.org/0000-0003-1331-4749
[2]Apatity branch of Murmansk Arctic State University, Apatity, Russia

Abstract

Spatial situations modeling is an effective means of providing information support for decision-making, including for managing complex dynamic systems while taking into account regional specifics. Geographic information systems

are currently a popular tool for modeling spatial situations. Modern systems of this class model situations using electronic maps and also provide users with a set of tools for analyzing spatial data. One of the challenges of this approach is the difficulty of accounting for the semantic properties and relationships of displayed objects, which cannot be directly recorded in the geographic information system database. Another challenge, caused by the approach to modeling situations and the method of data organization, is associated with the addition of new semantic information about already visualized objects. This paper addresses the problem of automated construction of spatial situation models for solving information support problems for decision-making, taking into account the geographic and semantic properties of spatial objects. A method for semantic enrichment of geospatial data based on a knowledge graph is proposed. A brief overview of existing approaches to using knowledge graphs to generate cartographic images is provided. The spatial situation model and the main components of the technology for automated construction of spatial situation models in the form of geosemantic images are described. The results of modeling the spatial situation as a geosemantic image constructed using a cartographic framework and a geospatial knowledge graph are presented.

Keywords:

geospatial knowledge graph, spatial situation, geosemantic image, situational modeling, spatial analysis, decision support information systems

Acknowledgments:

the study was carried out within the framework of the Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences state assignment of the Ministry of Science and Higher Education of the Russian Federation, research topic “Methods and technologies for creating intelligent information systems to support the development of complex dynamic systems with regional specifics in conditions of uncertainty and risk” (registration number of the research topic 1023032300374-0-2.2.1).

For citation:

Vicentiy A. V. Modeling of spatial situations as geosemantic images based on the graph of geospatial knowledge. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 140–153. doi:10.37614/2949-1215.2025.16.3.010.

Введение

В настоящее время ситуационное моделирование для информационной поддержки принятия решений в целях обеспечения развития сложных динамических региональных систем является одним из актуальных подходов, обеспечивающих оперативность и высокое качество принимаемых решений [1–3]. Современный этап цифровой трансформации характеризуется экспоненциальным ростом объемов геопространственных данных. Эти данные генерируется не только людьми, являющимися специалистами в области картографии и пространственного анализа, но и обычными пользователями различных информационных систем, а также в результате автоматической обработки спутниковых данных, устройствами интернета вещей, социальными медиа и геосервисами [4]. В условиях растущей сложности управляемых систем и динамичности происходящих в них процессов, требующих оперативного принятия решений, моделирование ситуаций играет ключевую роль в процессе обеспечения информацией лиц, принимающих решения (ЛПР) [5; 6].

В общем случае под моделированием ситуации понимают создание некоторого формализованного представления части реального мира, описываемого совокупностью взаимосвязанных объектов, их значимых характеристик и состояний, отношений между ними, а также, событий, происходящих в определенные интервалы времени, в случае необходимости учета динамики ситуации. Такое представление называется моделью ситуации и служит адекватной абстракцией, позволяющей проводить анализ, оценку и прогнозирование ее потенциального развития. Таким образом, моделирование ситуаций — это процесс представления реальных или гипотетических условий, объектов и обстоятельств в формализованном виде с целью анализа и понимания сложных явлений, которые невозможно или сложно изучать без использования моделей [7; 8]. Оно позволяет исследовать взаимодействия элементов системы, выявлять закономерности и прогнозировать возможные исходы развития событий.

Под ситуационным моделированием также понимают метод исследования, заключающийся не только в построении моделей ситуаций, но и в проведении экспериментов, отработке управленческих решений, прогнозировании, планировании и проверке гипотез без риска для реальных объектов [9; 10]. Ситуационное моделирование — это непрерывный динамический процесс, включающий в себя не только создание модели, но и ее постоянную актуализацию на основе

поступающих данных, идентификацию текущей ситуации, оценку ее соответствия целям управления, прогнозирование возможных сценариев развития и выработку управляющих воздействий [11].

Для создания моделей конкретных ситуаций, возникающих в определенных пространственно-временных условиях и характеризующихся множеством факторов и взаимосвязей, в ситуационном моделировании применяется широкий спектр методов и подходов. Среди наиболее распространенных можно выделить детерминированное и стохастическое моделирование, имитационное моделирование, логико-лингвистическое моделирование, сценарное моделирование, когнитивное моделирование, моделирование, основанное на знаниях, и геопространственное моделирование. В нашей работе мы предлагаем способ моделирования пространственных ситуаций как геосемантических изображений на основе комбинации геопространственного, когнитивного и основанного на знаниях подходов.

В настоящее время наиболее распространенными инструментами моделирования, обработки и анализа пространственных ситуаций являются географические информационные системы (ГИС). В большинстве современных ГИС, таких, например, как ArcGIS [12], MapInfo [13] и QGIS [14], модели пространственных ситуаций представляются в виде электронных карт и некоторого набора инструментов, позволяющего выполнять базовые манипуляции с ними. В общем виде под моделью пространственной ситуации понимается формализованное упрощенное и управляемое представление реальных или абстрактных объектов, явлений и процессов, которое явным образом учитывает их пространственное положение, форму, взаимное расположение и пространственные отношения. Наибольшее распространение получила реализация модели в виде многослойного картографического изображения. Пример представления ландшафта, содержащего как искусственные, так и природные объекты, с помощью такого типа модели представлен на рис. 1 [15].

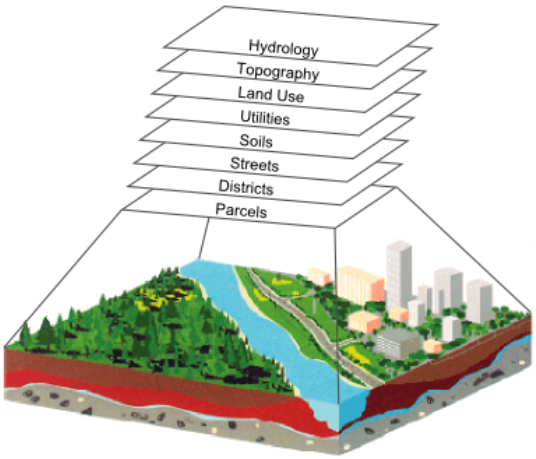


Рис. 1. Модель, состоящая из тематических слоев данных [15]

Такие модели создаются в ГИС с помощью набора тематических слоев, содержащих информацию о пространственных объектах (зданиях, дорогах, реках и т. д.), наложенных друг на друга в определенном порядке. Слои могут быть векторными либо растровыми, элементы слоя могут содержать различные атрибуты местоположения. Модели такого типа позволяют учитывать ряд пространственных отношений. Например, с помощью топологических отношений описываются смежность, пересечение, включение объектов, а метрические отношения позволяют задавать расстояния и направления. Семантическая информация задается преимущественно с помощью атрибутов (качественных характеристик) объектов, которые не должны дублировать их пространственные свойства. В модели атрибутивная информация может быть представлена с помощью текстовых или числовых значений.

Несмотря на широкое распространение, представление моделей пространственных ситуаций описанным выше способом имеет ряд недостатков, наиболее существенным из которых является невозможность добавления в модель семантической информации, которая не может быть описана

совокупностью атрибутов пространственных объектов. Примером такой семантической информации может быть информация, которая напрямую не относится к географической информации и обычно не сохраняется в базе данных ГИС, но необходима для задания контекста модели пространственной ситуации [16]. Кроме того, часто такая информация не может быть представлена с помощью комбинации текстовых или числовых значений атрибутов пространственных объектов, хранимых в базе данных ГИС [17]. Еще одной проблемой традиционного представления пространственной ситуации в виде электронной карты является невозможность адаптировать получаемые таким образом геоизображения под конкретные информационные потребности различных категорий пользователей без использования дополнительных средств [18; 19]. Также имеет место сложность, обусловленная классическим подходом к моделированию ситуаций, описанным выше, и способом организации данных в ГИС, связанная с добавлением новой семантической информации об уже визуализированных объектах. Как правило, оперативное внесение изменений в базу данных ГИС и визуализированное геоизображение не всегда возможно, кроме того, не все типы данных могут поддерживаться и корректно визуализироваться ГИС.

Перспективным подходом к решению указанных выше проблем представляется интеграция методов классического моделирования пространственных ситуаций с методами онтологического проектирования и построения графов знаний [20; 21], а также с когнитивными технологиями [22]. Совместное использование указанных методов и технологий позволяет решить проблему семантического обогащения геопространственных изображений, используемых для моделирования пространственных ситуаций без необходимости существенного изменения архитектуры ГИС. Такой подход также позволяет создавать на базе традиционных ГИС когнитивно-адаптированные системы, способные не только отражать структуру предметной области в виде семантических сетей (онтологий и графов знаний), но и учитывать контекст, извлекать скрытые зависимости и предоставлять персонализированные геоизображения для различных категорий пользователей на основе учета модели пользователя [23].

В нашей работе мы предлагаем способ построения моделей пространственных ситуаций как геосемантических изображений, представленных множеством описаний геообъектов и объектов, не хранящихся в базе данных ГИС, пространственных и семантических отношений, релевантных моделируемой ситуации. Общая схема построения такого изображения с использованием геопространственного графа знаний (ГГЗ) представлена на рис. 2.

Рис. 2. Общая схема построения геосемантического изображения

Геосемантическое изображение строится на основе интеграции геоданных, получаемых из ГИС, семантических данных, извлекаемых из ГГЗ, и формализованного запроса пользователя,

дополнительно обработанного с помощью методов расширения контекста таким образом, чтобы лучше отражать его информационную потребность [24]. Результатом является геосемантическое изображение, выдаваемое пользователю системы информационной поддержки принятия решений, описывающее исследуемую пользователем пространственную ситуацию на основе представления пертинентных запросу множеств геообъектов, пространственных и семантических отношений, а также дополнительной семантической информации.

Материалы и методы

Для того, чтобы представить модель пространственной ситуации в виде геосемантического изображения, необходимо обеспечить представление информации не только об объектах и отношениях, хранящихся в базе данных ГИС, но и об объектах и отношениях, информация о которых хранится во внешних по отношению к ГИС ресурсах. В качестве таких внешних ресурсов информации могут выступать, например, различные базы знаний, онтологии [25] и графы знаний (ГЗ) [26]. В последние годы применение именно ГЗ для семантического обогащения геоизображений вызывает наибольший интерес [27–29]. Таким образом, ГЗ, ассоциированный с ГИС и используемый для семантического обогащения формируемых геоизображений (цифровых карт) можно условно представить как дополнительный семантический слой цифровой карты. Причем для конечного пользователя, как правило, не имеет значения из какого источника, ГИС или ГЗ, была получена информация для синтеза геосемантического изображения. Для конечного пользователя важно только то, насколько хорошо сформированное геоизображение соответствует его информационной потребности [30; 31].

В этом контексте под ГЗ можно понимать структурированную модель данных, которая представляет информацию в виде семантической сети взаимосвязанных сущностей (узлов графа) и их отношений (рёбер графа). ГГЗ — это вид ГЗ, позволяющий реализовать структурированное представление геопространственных данных. Узлами такого графа, как правило, являются сущности реального мира, пространственные объекты и события, а ребрами — связи между ними [32; 33]. За счет использования онтологических принципов для формального определения концептов, свойств и связей, ГГЗ обеспечивает эффективную интеграцию разнородных данных, а также ускоряет их поиск и анализ по сравнению с традиционными базами данных. Кроме того, использование геопространственных ГЗ позволяет реализовывать принципы FAIR (Findable, Accessible, Interoperable, and Reusable), что существенно облегчает анализ и управление геоданными в информационных системах [34]. В последние годы технология графов знаний находит широкое применение в решении научных [35–37] и прикладных задач в таких областях, как прогнозирование стихийных бедствий [38; 39], сельское хозяйство [40; 41], строительство сложных инженерных сооружений [42; 43], обработка геоданных [44], прогнозирование погодных явлений [45] и многих других.

В нашей работе мы также используем геопространственный ГЗ для семантического обогащения моделируемой пространственной ситуации. Одной из главных проблем при создании ГЗ в настоящее время является проблема получения данных о релевантных сущностях и связях из разнородных источников, их очистка и приведение в необходимый формат для включения в граф [46]. Предлагаемая нами технология автоматизированного построения моделей пространственных ситуаций в виде геосемантических изображений предполагает использование средств больших языковых моделей (LLM — Large Language Model) для извлечения данных из текстов на естественном языке, необходимых для наполнения графа. Такой подход позволяет существенно снизить трудоемкость и повысить оперативность создания и пополнения геопространственного ГЗ [47–50]. В качестве большой языковой модели для извлечения данных мы выбрали DeepSeek-V3.2. Выбор данной LLM был основан на результатах предварительного тестирования нескольких моделей в задачах извлечения RDF-троек из текстов на естественном языке. Для предварительного тестирования были выбраны такие модели, как ChatGPT, DeepSeek, GigaChat и YandexGPT. Сопоставимые результаты показали модели ChatGPT и DeepSeek, GigaChat показал результат хуже, чем DeepSeek, но лучше, чем YandexGPT. Учитывая, что компания OpenAI, которой принадлежит ChatGPT, официально не работает в России, окончательный выбор большой языковой модели был сделан в пользу DeepSeek в последней доступной версии.

Для хранения и обработки сформированного геопространственного ГЗ была выбрана СУБД, поддерживающая графовые структуры данных. Как правило, подобные СУБД уже оптимизированы для хранения данных, представленных в виде сетей и графов, что обеспечивает лучшую производительность по сравнению с традиционными СУБД [51]. На основе сравнения нескольких графовых СУБД, представленного в работе [52], выбор был сделан в пользу Neo4j, которая показала наилучшие результаты как в области потребления ресурсов, так и в части скорости обработки и реализации специфичных функций обработки больших объемов данных.

Учитывая тот факт, что потенциальная семантика моделируемых пространственных ситуаций может очень сильно различаться, невозможно предложить такие модель пространственной ситуации и технологию автоматизированного построения ситуаций в виде геосемантических изображений, которые были бы адекватны для любой предметной области. В связи с этим на данном этапе исследования мы решили ограничить предметную область рассматриваемых пространственных ситуаций чрезвычайными ситуациями на железнодорожном транспорте. Этот класс ситуаций будет являться основным для рассмотрения, но предложенные в рамках работы модель и технология обладают некоторой универсальностью, а подходы и принципы, положенные в их основу, могут быть использованы при моделировании схожих классов пространственных ситуаций. Для демонстрации некоторых элементов разработанных модели и технологии используется описание конкретной чрезвычайной ситуации (кейса) столкновения поездов на станции Княжая в Мурманской области [53].

В качестве источников информации для наполнения геопространственного ГЗ использованы тексты и мультимедийные материалы (видео, фото, схемы, аудиозаписи) открытого доступа. Основными источниками данных являлись официальные сайты различных министерств, ведомств, органов исполнительной власти, надзорных органов, например, Министерства транспорта Российской Федерации, открытого акционерного общества «Российские железные дороги», губернатора Мурманской области, транспортной прокуратуры и других. Также отдельным источником информации являлись сообщения в релевантных сообществах виртуальных социальных сетей и в популярных мессенджерах. Собранные таким образом данные были разделены на текстовые данные и мультимедийные данные. Текстовые данные были обработаны посредством большой языковый модели с целью извлечения RDF-троек для последующего формирования геопространственного ГЗ. Мультимедиа-данные также прошли предварительную обработку и были добавлены в геопространственный ГЗ в виде отдельных узлов средствами Neo4j.

Результаты

Основными результатами данной работы являются модель пространственной ситуации, представленная множеством геообъектов, пространственных и семантических отношений в виде фрагмента геосемантического изображения, а также технология автоматизированного построения моделей пространственных ситуаций в виде геосемантических изображений. Предложенная модель основана на онтологии ситуации, используемой в качестве концептуального шаблона для создания геопространственного ГЗ, включающего как географическую информацию об объектах, пространственных и семантических отношениях между ними, так и дополнительную семантическую информацию. Под дополнительной информацией понимается информация из внешних источников, которая напрямую не относится к пространственной ситуации, но семантически связана с моделируемой пространственной ситуацией. Например, для чрезвычайной ситуации столкновения поездов такой информацией может быть количество пассажирских мест в железнодорожном вагоне, погодные условия в момент аварии, количество пострадавших, категории пассажиров, тип груза и т. д. Благодаря комбинированию разнородных данных (геоданные, текст, графика, видео) в структуре единого геопространственного ГЗ, модель полезна для различных категорий пользователей, привлекаемых к выработке решений в рамках моделируемой пространственной ситуации. Предложенная модель может быть использована для решения широкого круга прикладных задач поддержки развития сложных региональных систем, для которых пространственные данные имеют большое значение, таких, например, как реагирование на чрезвычайные ситуации, смягчение последствий чрезвычайных ситуаций, мониторинг заболеваний, анализ городских потоков и других.

Схема использования предложенной модели для моделирования пространственной ситуации столкновения поездов на станции Княжая как геосемантического изображения на основе геопространственного графа знаний представлена на рис. 3.

В левой части схемы изображены геоданные, используемые для классического представления модели пространственной ситуации в виде набора тематических слоев цифровой карты. В правой части схемы изображен геопространственный ГЗ, содержащий дополнительную информацию, семантически связанную с моделируемой ситуацией и сформированный на основе разнородных источников данных о столкновении поездов. В центре схемы находится модель пространственной ситуации, представленная множеством геообъектов, пространственных и семантических отношений в виде фрагмента геосемантического изображения.

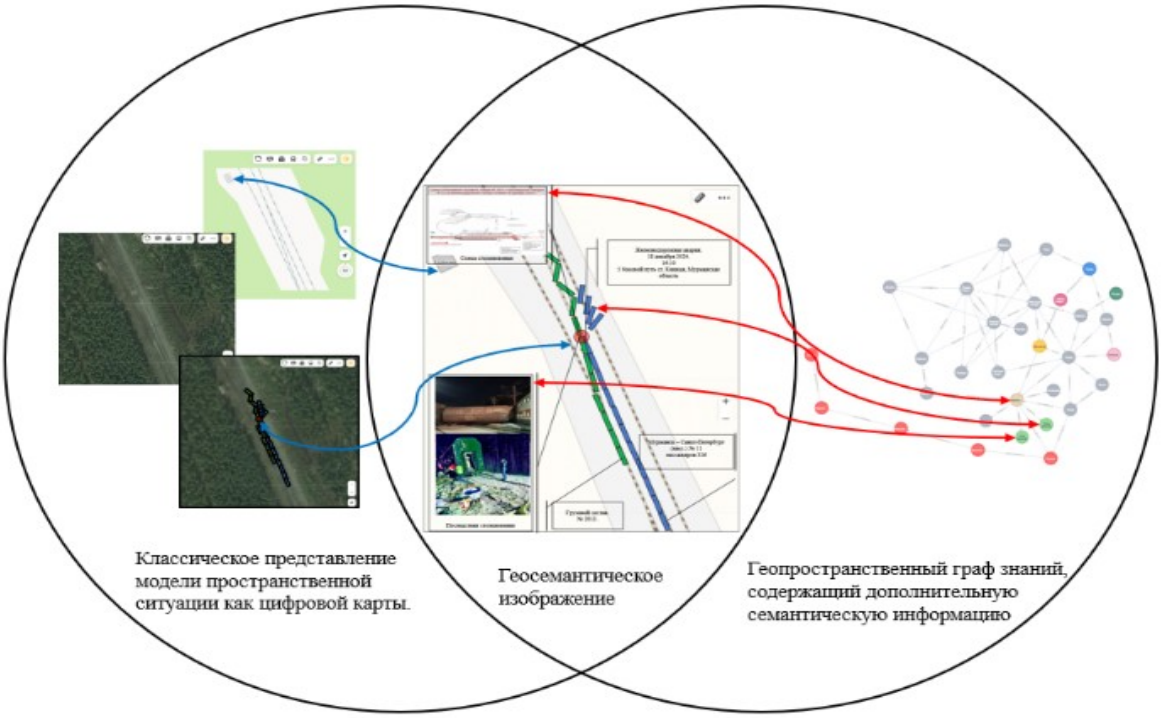


Рис. 3. Схема моделирования пространственной ситуации столкновения поездов на станции Княжая как геосемантического изображения на основе геопространственного графа знаний

Геосемантическое изображение изначально формируется как классическое геоизображение в виде цифровой карты, состоящей из нескольких слоев. На этом этапе визуализируются основные геопространственные объекты, релевантные рассматриваемой ситуации, составляющие основу геоизображения. В результате выполнения этого этапа формируется цифровая карта подобно тому, как это происходит в большинстве современных географических информационных систем. На втором этапе осуществляется обогащение сформированной цифровой карты дополнительной информацией, семантически связанной с моделируемой ситуацией. Для этой цели используется предварительно построенный геопространственный ГЗ. Из графа извлекаются текстовые описания, фото-, видео-, аудиоматериалы, ассоциированные с тем или иным геообъектом на цифровой карте. Например, при выборе на карте сошедшего с рельсов вагона пассажирского поезда геосемантическое изображение обогащается фотографиями последствий аварии, иллюстрирующими положение в пространстве сошедшего с рельсов вагона, его повреждения и взаимное расположение относительно других объектов на месте аварии. При визуальном анализе этих снимков можно сделать вывод о том, что вагон не просто сошел с рельс, но и заблокировал не мене двух рельсовых колей, что может существенно затруднить подъезд к месту аварии и работу ремонтного поезда. Кроме того, ГЗ содержит также

семантическую информацию о типах вагонов обоих поездов, количестве пассажиров, перевозимом грузе, количестве пострадавших, машинистах и их помощниках и другую информацию, релевантную рассматриваемой ситуации. В этом контексте важно отметить, что именно сформированный с помощью предложенной в работе технологии автоматизированного построения моделей пространственных ситуаций в виде геосемантических изображений геопространственный ГЗ оказывает решающее влияние на эффективность семантического обогащения.

Разработанная технология автоматизированного построения моделей пространственных ситуаций в виде геосемантических изображений позволяет на основе заданной онтологией схемы сформировать геопространственный ГЗ, описывающий соответствующую ситуацию. Объекты, их географические и семантические свойства и отношения извлекаются из разнородных источников данных, содержащих описание моделируемой ситуации, с помощью выделения RDF-триплетов средствами больших языковых моделей. Затем полученные триплеты объединяются в RDF-граф и сохраняются в графовой СУБД, образуя единый геопространственный ГЗ. Результатом работы технологии является модель пространственной ситуации, представленная машиночитаемым геопространственным ГЗ, сохраненным в графовую СУБД, что допускает его многократное повторное использование различными пользователями и приложениями. Предложенная технология может быть использована для решения задач семантического обогащения геопространственных данных с целью повышения качества проведения пространственного анализа для информационной поддержки принятия решений. Обобщенная схема технологии представлена на рис. 4.

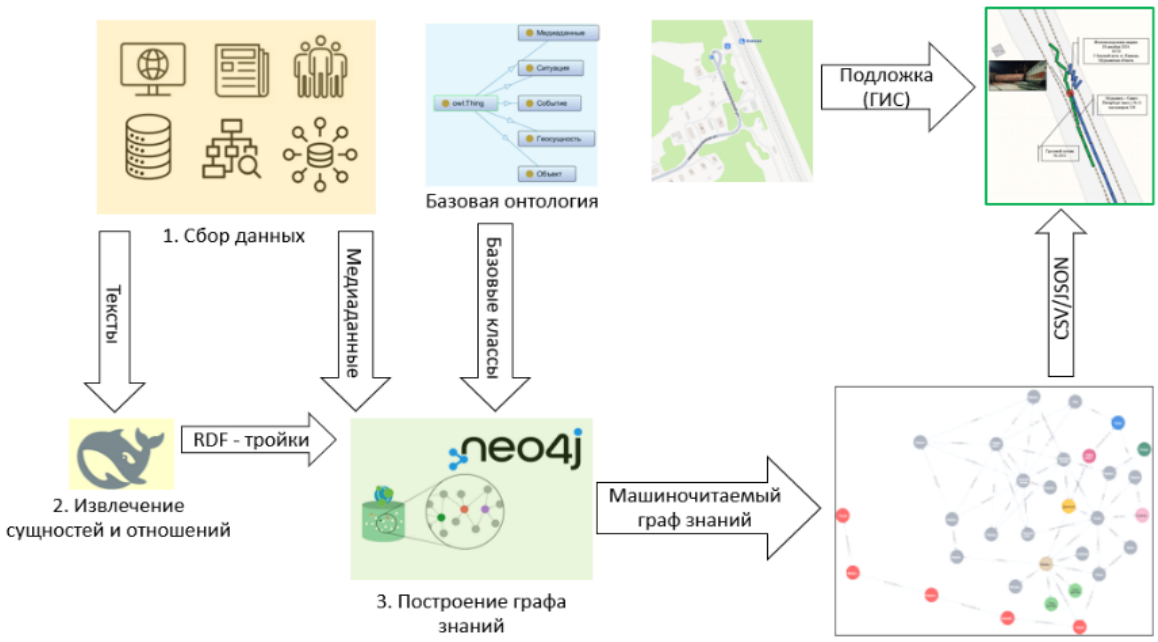


Рис. 4. Обобщенная схема технологии автоматизированного построения моделей пространственных ситуаций в виде геосемантических изображений

На первом этапе технологии осуществляется сбор разнородных данных для последующей обработки с целью формирования геопространственного ГЗ. В качестве источников используются как официальные сайты различных министерств, ведомств, органов исполнительной власти, надзорных органов, так и сообщения в сообществах виртуальных социальных сетей и в популярных мессенджерах. На этом этапе все собранные данные делятся на два основных вида — текстовые и мультимедийные. К текстовым данным относятся непосредственно тексты, а к мультимедийным — фотографии, видеоматериалы, аудиозаписи, схемы и другая релевантная информация, не являющаяся текстом. На втором этапе происходит обработка собранных текстовых материалов средствами большой языковой модели с целью извлечения из текстов основных сущностей и отношений в виде RDF-триплетов.

Для построения машиночитаемого геопространственного ГЗ на третьем этапе технологии используется графовая СУБД, позволяющая хранить и отображать данные в виде семантической сети или графа, а также обеспечивающая инструменты работы с такими структурами данных. Кроме того, выбранная СУБД Neo4j позволяет хранить в качестве узлов графа мультимедийные данные, что существенно облегчает доступ к ним и реализацию связывания этих данных с другими узлами графа. Результатом данного этап технологии автоматизированного построения моделей пространственных ситуаций в виде геосемантических изображений является геопространственный ГЗ, допускающий многократное повторное использование. Далее сформированный граф используется для семантического обогащения цифровой карты информацией, связанной с моделируемой ситуацией, но отсутствующей в базе данных ГИС и выполняет роль как бы дополнительного семантического слоя цифровой карты. Пример построенного таким образом графа для текста, описывающего ситуацию столкновения поездов на станции Княжая, представлен на рис. 5.

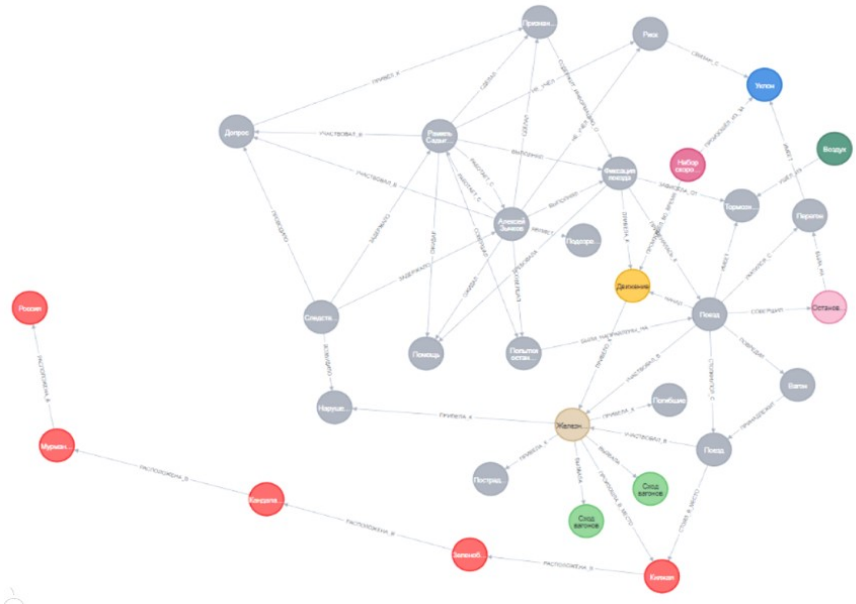


Рис. 5. Пример графа знаний для текстового описания ситуации столкновения поездов на станции Княжая

Заключение

В настоящее время для решения задач информационной поддержки принятия решений для управления сложными динамическими системами с региональной спецификой, наряду с прочими, активно используются подходы на основе ситуационного моделирования. Несмотря на значительное развитие программных средств, таких как географические информационные системы, позволяющие создавать модели пространственных ситуаций в виде цифровых карт, ряд проблем, связанных с семантическим наполнением таких изображений, решается недостаточно эффективно.

Данная работа посвящена решению проблемы моделирования пространственных ситуаций для задач информационной поддержки принятия решений с учетом географических и семантических свойств пространственных объектов. Для решения этой проблемы предложены модель пространственной ситуации, представленная множеством геообъектов, пространственных и семантических отношений в виде геосемантического изображения, а также технология автоматизированного построения моделей пространственных ситуаций в виде геосемантических изображений. В основе данного изображения лежит формируемый в автоматизированном режиме средствами большой языковой модели геопространственный граф знаний, используемый для семантического обогащения изображения. Для демонстрации некоторых элементов модели и технологии в качестве примера используется описание конкретной чрезвычайной ситуации столкновения поездов на станции Княжая в Мурманской области.

Полученные в данной работе результаты являются основой для следующих работ в области разработки метода управления геовизуализацией с целью синтеза геосемантических изображений, обладающих прагматическими свойствами для информационной поддержки развития сложных динамических систем.

Список источников

1. Suduc A. M., Bizoi M. A Quantitative Perspective on the Evolution of Decision Support Systems // Stud. Syst. Decis. Control. Springer Science and Business Media Deutschland GmbH. 2024. Vol. 534. P. 93–108.
2. Fridman A. Y. Experience of Intellectualization of Situational Modeling Methods for Discrete Time-Varying Spatial Objects // Autom. Remote Control 2022 836. Springer. 2022. Vol. 83, No. 6. P. 946–959.
3. Vicentiy A. V. The Geoimage Generation Method for Decision Support Systems Based on Natural Language Text Analysis // Lect. Notes Networks Syst. Springer Science and Business Media Deutschland GmbH. 2021. Vol. 230. P. 609–619.
4. Wu J. et al. Geospatial Big Data: Survey and Challenges // IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. Institute of Electrical and Electronics Engineers Inc. 2024. Vol. 17. P. 17007–17020.
5. Desku F. The Role of Information Systems in the Decision-making Processes of the Enterprises in Kosovo // Springer Proc. Bus. Econ. Springer Nature. 2023. P. 801–814.
6. Вицентий А. В. Разработка формальной модели конфликтной ситуации для мультипредметной системы поддержки управления пространственно-распределенными социально-экономическими системами // Информатизация и связь. 2019. Vol. 4. P. 64–69.
7. Wu C., Liu Y. Mathematical modeling in human–machine system design and evaluation // Handbook of human factors and ergonomics. Wiley, 2021. P. 685–703.
8. Paul G. E. Modeling and Simulation of Human Systems // Handb. Hum. Factors Ergon. Wiley, 2021. P. 704–735.
9. Nicheporchuk V. V., Nozhenkov A. I. Emergencies situational modeling technology for territorial management support // Procedia Struct. Integr. Elsevier, 2019. Vol. 20. P. 248–253.
10. Wang J. et al. Situation modeling and evaluation for complex systems: A case study of satellite attitude control system // Adv. Eng. Informatics. Elsevier, 2024. Vol. 61. P. 102505.
11. Ivanov V. V., Sarkisyants Y. K. Situation modeling for decision making in international contracts // Russ. Foreign Econ. Bull. 2019. No. 9. P. 80–94.
12. Desktop GIS Software | Mapping Analytics | ArcGIS Pro [Electronic resource]. URL: https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview (accessed 05.11.2025).
13. MapInfo Pro: A complete, desktop mapping GIS software solution [Electronic resource]. URL: https://www.precisely.com/product/precisely-mapinfo/mapinfo-pro/ (accessed 05.11.2025).
14. Spatial without Compromise QGIS Web Site [Electronic resource]. URL: https://qgis.org/ (accessed 05.11.2025).
15. Solve spatial problems with representation and process models—ArcGIS Pro | Documentation [Electronic resource]. URL: https://pro.arcgis.com/en/pro-app/latest/help/analysis/spatial-analyst/model-solve-spatial-problems/solving-spatial-problems-with-representation-and-process-models.htm (accessed 05.11.2025).
16. Zhu J., Wu P. BIM/GIS data integration from the perspective of information flow // Autom. Constr. Elsevier, 2022. Vol. 136. P. 104166.
17. Zhou C. Exploring future GIS visions in the era of the scientific and technological revolution // Inf. Geogr. Elsevier, 2025. Vol. 1, No. 1. P. 100007.
18. Singh H. et al. Modelling human-centric aspects of end-users with iStar // J. Comput. Lang. 2022. Vol. 68. P. 101091.
19. Vicentiy A. V. Modelling the user's information needs based on the assessment of pragmatic informativeness of data // Informatiz. Commun. Informatization and Communication Journal Editorial Board. 2024. Vol. 1. P. 65–69.
20. Hogan A. et al. Knowledge Graphs // ACM Comput. Surv. Association for Computing Machinery. 2021. Vol. 54, No. 4.
21. Weikum G. et al. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases // Found. Trends® Databases. Now Publishers, Inc., 2021. Vol. 10, No. 2–4. P. 108–490.
22. Vicentiy A. V. Development of methods and tools to support regional management in the Arctic zone of the Russian Federation based on cognitive interfaces // IOP Conf. Ser. Earth Environ. Sci. IOP Publishing. 2019. Vol. 302, No. 1. P. 012139.
23. Vicentiy A. Definition and Formalization of the User Mental Model for Creating Adaptive Geointerfaces of Decision Support Systems // Lecture Notes in Networks and Systems. Springer Science and Business Media Deutschland GmbH, 2024. Vol. 733. P. 1095–1105.
24. Vicentiy A. V. Methods for estimating the quantity of information in the context of satisfying the user's information need // Informatiz. Commun. Informatization and Communication Journal Editorial Board. 2024. Vol. 2.

25. Gruber T. R. A translation approach to portable ontology specifications // Knowl. Acquis. Academic Press, 1993. Vol. 5, No. 2. P. 199–220.
26. Wu J. et al. Geospatial Big Data: Survey and Challenges // IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. Institute of Electrical and Electronics Engineers Inc., 2024. Vol. 17. P. 17007–17020.
27. Lilis G. N. et al. BIM-based semantic enrichment and knowledge graph generation via geometric relation checking // Autom. Constr. Elsevier, 2025. Vol. 173. P. 106081.
28. Zhu R. Geospatial Knowledge Graphs // https://arxiv.org/abs/2405.07664. 2024.
29. Zhu R. et al. The KnowWhereGraph: A Large-Scale Geo-Knowledge Graph for Interdisciplinary Knowledge Discovery and Geo-Enrichment // arXiv Prepr. arXiv2502.13874. 2025.
30. Hogan A. et al. Large Language Models, Knowledge Graphs and Search Engines: A Crossroads for Answering Users' Questions // http://arxiv.org/abs/2501.06699. 2025.
31. Вицентий А. В. Моделирование информационной потребности пользователя на основе оценки прагматической информативности данных // Информатизация и связь. 2024. Т. 1. С. 65–69.
32. Li T. et al. HGeoKG: A Hierarchical Geographic Knowledge Graph for Geographic Knowledge Reasoning // ISPRS Int. J. Geo-Information. Multidisciplinary Digital Publishing Institute, 2025. Vol. 14, No. 1.
33. Xu H. et al. From Maps to Geospatial Knowledge Graph: Geospatial Knowledge Representation and Reasoning // Abstr. ICA. Copernicus GmbH, 2023. Vol. 6. P. 1–2.
34. Wang S. et al. Review, framework, and future perspectives of Geographic Knowledge Graph (GeoKG) quality assessment // Geo-spatial Inf. Sci. Taylor & Francis, 2025. Vol. 28, No. 4. P. 1701–1721.
35. Zou X. A Survey on Application of Knowledge Graph // J. Phys. Conf. Ser. Institute of Physics Publishing, 2020. Vol. 1487, No. 1.
36. Liu H. et al. Knowledge graph reasoning: Mainstream methods, applications and prospects // Eng. Appl. Artif. Intell. Pergamon, 2025. Vol. 159. P. 111625.
37. Zhao X. et al. A hierarchical spatio-temporal object knowledge graph model for dynamic scene representation // Trans. GIS. John Wiley and Sons Inc, 2023. Vol. 27, No. 7. P. 1992–2016.
38. Ge X. et al. Disaster Prediction Knowledge Graph Based on Multi-Source Spatio-Temporal Information // Remote Sens. Multidisciplinary Digital Publishing Institute, 2022. Vol. 14, No. 5.
39. Du W. et al. OFPO & KGFPO: Ontology and knowledge graph for flood process observation // Environ. Model. Softw. Elsevier, 2025. Vol. 185. P. 106317.
40. Lin Y. et al. A reasoning method for rice fertilization strategy based on spatiotemporal knowledge graph // Trans. GIS. John Wiley and Sons Inc, 2024. Vol. 28, No. 4. P. 902–924.
41. Ge W. et al. A recommendation model of rice fertilization using knowledge graph and case-based reasoning // Comput. Electron. Agric. Elsevier, 2024. Vol. 219. P. 108751.
42. Wang L. et al. Multimodal knowledge graph construction for risk identification in water diversion projects // J. Hydrol. Elsevier, 2024. Vol. 635. P. 131155.
43. Lai J. et al. Dynamic data-driven railway bridge construction knowledge graph update method // Trans. GIS. John Wiley and Sons Inc, 2023. Vol. 27, No. 7. P. 2099–2117.
44. Cai K. et al. Construction of Earth Observation Knowledge Hub Based on Knowledge Graph // Trans. GIS. John Wiley and Sons Inc, 2024. Vol. 28, No. 7. P. 2445–2462.
45. Wang A. et al. The Construction of a Multimodal Knowledge Graph for WRF Simulation Knowledge // Trans. GIS. John Wiley and Sons Inc, 2025. Vol. 29, No. 7. P. e70134.
46. Mai G. et al. Towards the next generation of Geospatial Artificial Intelligence // Int. J. Appl. Earth Obs. Geoinf. Elsevier, 2025. Vol. 136. P. 104368.
47. Liang J. et al. Design and application of a semantic-driven geospatial modeling knowledge graph based on large language models // Geo-spatial Inf. Sci. Taylor and Francis Ltd., 2025. P. 1–20.
48. Meyer L. P. et al. LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT // Inform. aktuell. Springer Science and Business Media Deutschland GmbH, 2024. P. 103–115.
49. Zhang B., Soh H. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction // EMNLP 2024 — 2024 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf. Association for Computational Linguistics (ACL). 2024. P. 9820–9836.
50. Krishna Kommineni V. et al. From human experts to machines: An LLM supported approach to ontology and knowledge graph construction // arXiv. 2024. P. arXiv:2403.08345.
51. Liu Y. et al. Construction and Application of User Check-in Spatiotemporal Knowledge Graph Based on Neo4j // Procedia Comput. Sci. Elsevier, 2024. Vol. 242. P. 609–616.

52. Monteiro J. et al. Experimental Evaluation of Graph Databases: JanusGraph, Nebula Graph, Neo4j, and TigerGraph // Appl. Sci. Multidisciplinary Digital Publishing Institute, 2023. Vol. 13, No. 9. P. 5770
53. Столкновение поездов на станции Княжая // Википедия : сайт. URL: https://ru.wikipedia.org/wiki/Столкновение_поездов_на_станции_Княжая (дата обращения: 11.11.2025).

References

1. Suduc A. M., Bizoi M. A Quantitative Perspective on the Evolution of Decision Support Systems. *Stud. Syst. Decis. Control*. Springer Science and Business Media Deutschland GmbH, 2024, Vol. 534, pp. 93–108.
2. Fridman A. Y. Experience of Intellectualization of Situational Modeling Methods for Discrete Time-Varying Spatial Objects. *Autom. Remote Control* 2022 836. Springer, 2022, Vol. 83, no. 6, pp. 946–959.
3. Vicentiy A. V. The Geoimage Generation Method for Decision Support Systems Based on Natural Language Text Analysis. Lect. *Notes Networks Syst*. Springer Science and Business Media Deutschland GmbH, 2021, Vol. 230, pp. 609–619.
4. Wu J. et al. Geospatial Big Data: Survey and Challenges. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. Institute of Electrical and Electronics Engineers Inc., 2024, Vol. 17, pp. 17007–17020.
5. Desku F. The Role of Information Systems in the Decision-making Processes of the Enterprises in Kosovo. *Springer Proc. Bus. Econ*. Springer Nature, 2023, pp. 801–814.
6. Vicentiy A. V. Razrabotka formal'noj modeli konfliktnoj situacii dlya mul'tipredmetnoj sistemy podderzhki upravleniya prostranstvenno-raspredelennymi social'no-ekonomicheskimi sistemami [Development of a Formal Model of a Conflict Situation for a Multidisciplinary Management Support System for Spatially Distributed Socioeconomic Systems]. *Informatizaciya i svyaz'* [Informatization and Communications], 2019, no. 4, pp. 64–69. (In Russ.).
7. Wu C., Liu Y. Mathematical modeling in human–machine system design and evaluation. *Handbook of human factors and ergonomics*. Wiley, 2021, pp. 685–703.
8. Paul G. E. Modeling and Simulation of Human Systems. *Handb. Hum. Factors Ergon*. Wiley, 2021, pp. 704–735.
9. Nicheporchuk V. V., Nozhenkov A. I. Emergencies situational modeling technology for territorial management support. *Procedia Struct. Integr*. Elsevier, 2019, Vol. 20, pp. 248–253.
10. Wang J. et al. Situation modeling and evaluation for complex systems: A case study of satellite attitude control system. *Adv. Eng. Informatics*. Elsevier, 2024, Vol. 61, pp. 102505.
11. Ivanov V. V., Sarkisyants Y. K. Situation modeling for decision making in international contracts. *Russ. Foreign Econ. Bull.*, 2019, Vol. 9, pp. 80–94.
12. Desktop GIS Software | Mapping Analytics | ArcGIS Pro [Electronic resource]. Available at: https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview (accessed 05.11.2025).
13. MapInfo Pro: A complete, desktop mapping GIS software solution. Available at: https://www.precisely.com/product/precisely-mapinfo/mapinfo-pro/ (accessed 05.11.2025).
14. Spatial without Compromise QGIS Web Site. Available at: https://qgis.org/ (accessed 05.11.2025).
15. Solve spatial problems with representation and process models—ArcGIS Pro | Documentation. Available at: https://pro.arcgis.com/en/pro-app/latest/help/analysis/spatial-analyst/model-solve-spatial-problems/solving-spatial-problems-with-representation-and-process-models.htm (accessed 05.11.2025).
16. Zhu J., Wu P. BIM/GIS data integration from the perspective of information flow. *Autom. Constr*. Elsevier, 2022, Vol. 136, p. 104166.
17. Zhou C. Exploring future GIS visions in the era of the scientific and technological revolution. *Inf. Geogr*. Elsevier, 2025, Vol. 1, no. 1, p. 100007.
18. Singh H. et al. Modelling human-centric aspects of end-users with iStar. *J. Comput. Lang.*, 2022, Vol. 68, p. 101091.
19. Vicentiy A. V. Modelling the user's information needs based on the assessment of pragmatic informativeness of data. *Informatiz. Commun. Informatization and Communication Journal Editorial Board*, 2024, Vol. 1, pp. 65–69.
20. Hogan A. et al. Knowledge Graphs. *ACM Comput. Surv. Association for Computing Machinery*, 2021, Vol. 54, no. 4.
21. Weikum G. et al. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Found. Trends® Databases*. Now Publishers, Inc., 2021, Vol. 10, no. 2, pp. 108–490.
22. Vicentiy A. V. Development of methods and tools to support regional management in the Arctic zone of the Russian Federation based on cognitive interfaces. *IOP Conf. Ser. Earth Environ. Sci*. IOP Publishing, 2019, Vol. 302, no. 1, p. 012139.
23. Vicentiy A. Definition and Formalization of the User Mental Model for Creating Adaptive Geointerfaces of Decision Support Systems. *Lecture Notes in Networks and Systems*. Springer Science and Business Media Deutschland GmbH, 2024, Vol. 733. pp. 1095–1105.

24. Vicentiy A. V. Methods for estimating the quantity of information in the context of satisfying the user's information need. *Informatiz. Commun. Informatization and Communication Journal Editorial Board*, 2024, Vol. 2.
25. Gruber T. R. A translation approach to portable ontology specifications. *Knowl. Acquis*. Academic Press, 1993, Vol. 5, no. 2, pp. 199–220.
26. Wu J. et al. Geospatial Big Data: Survey and Challenges. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. Institute of Electrical and Electronics Engineers Inc., 2024, Vol. 17, pp. 17007–17020.
27. Lilis G. N. et al. BIM-based semantic enrichment and knowledge graph generation via geometric relation checking. *Autom. Constr*. Elsevier, 2025, Vol. 173, p. 106081.
28. Zhu R. Geospatial Knowledge Graphs. Available at: https://arxiv.org/abs/2405.07664 (accessed 05.11.2025).
29. Zhu R. et al. The KnowWhereGraph: A Large-Scale Geo-Knowledge Graph for Interdisciplinary Knowledge Discovery and Geo-Enrichment. Available at: https://arXiv2502.13874 (accessed 05.11.2025).
30. Hogan A. et al. Large Language Models, Knowledge Graphs and Search Engines: A Crossroads for Answering Users' Questions. Available at: http://arxiv.org/abs/2501.06699 (accessed 05.11.2025).
31. Vicentiy A. V. Modelirovanie informacionnoj potrebnosti pol'zovatelya na osnove ocenki pragmaticheskoj informativnosti dannyh [Modeling the user's information needs based on the assessment of the pragmatic informativeness of data]. *Informatizaciya i svyaz'* [Informatization and Communication], 2024. no. 1, pp. 65–69. (In Russ.).
32. Li T. et al. HGeoKG: A Hierarchical Geographic Knowledge Graph for Geographic Knowledge Reasoning. *ISPRS Int. J. Geo-Information*, 2025, Vol. 14, Multidisciplinary Digital Publishing Institute, 2025, Vol. 14, no. 1.
33. Xu H. et al. From Maps to Geospatial Knowledge Graph: Geospatial Knowledge Representation and Reasoning. *Abstr. ICA*. Copernicus GmbH, 2023, Vol. 6, pp. 1–2.
34. Wang S. et al. Review, framework, and future perspectives of Geographic Knowledge Graph (GeoKG) quality assessment. *Geo-spatial Inf. Sci*. Taylor & Francis, 2025, Vol. 28, no. 4, pp. 1701–1721.
35. Zou X. A Survey on Application of Knowledge Graph. *J. Phys. Conf. Ser*. Institute of Physics Publishing, 2020, Vol. 1487, no. 1.
36. Liu H. et al. Knowledge graph reasoning: Mainstream methods, applications and prospects. *Eng. Appl. Artif. Intell. Pergamon*, 2025, Vol. 159, p. 111625.
37. Zhao X. et al. A hierarchical spatio-temporal object knowledge graph model for dynamic scene representation. *Trans. GIS*. John Wiley and Sons Inc, 2023, Vol. 27, no. 7, pp. 1992–2016.
38. Ge X. et al. Disaster Prediction Knowledge Graph Based on Multi-Source Spatio-Temporal Information. *Remote Sens.*, 2022, Vol. 14, Multidisciplinary Digital Publishing Institute, 2022, Vol. 14, no. 5.
39. Du W. et al. OFPO & KGFPO: Ontology and knowledge graph for flood process observation. *Environ. Model. Softw*. Elsevier, 2025, Vol. 185, p. 106317.
40. Lin Y. et al. A reasoning method for rice fertilization strategy based on spatiotemporal knowledge graph. *Trans. GIS*. John Wiley and Sons Inc, 2024, Vol. 28, no. 4, pp. 902–924.
41. Ge W. et al. A recommendation model of rice fertilization using knowledge graph and case-based reasoning. *Comput. Electron. Agric*. Elsevier, 2024, Vol. 219, p. 108751.
42. Wang L. et al. Multimodal knowledge graph construction for risk identification in water diversion projects. *J. Hydrol*. Elsevier, 2024, Vol. 635, p. 131155.
43. Lai J. et al. Dynamic data-driven railway bridge construction knowledge graph update method. *Trans. GIS*. John Wiley and Sons Inc, 2023, Vol. 27, no. 7, pp. 2099–2117.
44. Cai K. et al. Construction of Earth Observation Knowledge Hub Based on Knowledge Graph. *Trans. GIS*. John Wiley and Sons Inc, 2024, Vol. 28, no. 7, pp. 2445–2462.
45. Wang A. et al. The Construction of a Multimodal Knowledge Graph for WRF Simulation Knowledge. *Trans. GIS*. John Wiley and Sons Inc, 2025, Vol. 29, no. 7, p. e70134.
46. Mai G. et al. Towards the next generation of Geospatial Artificial Intelligence. *Int. J. Appl. Earth Obs. Geoinf*. Elsevier, 2025, Vol. 136, p. 104368.
47. Liang J. et al. Design and application of a semantic-driven geospatial modeling knowledge graph based on large language models. *Geo-spatial Inf. Sci*. Taylor and Francis Ltd., 2025, pp. 1–20.
48. Meyer L. P. et al. LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT. *Inform. aktuell*. Springer Science and Business Media Deutschland GmbH, 2024, pp. 103–115.
49. Zhang B., Soh H. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. *EMNLP 2024 — 2024 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf. Association for Computational Linguistics (ACL)*, 2024, pp. 9820–9836.
50. Krishna Kommineni V. et al. From human experts to machines: An LLM supported approach to ontology and knowledge graph construction. Available at: https://arXiv:2403.08345 (accessed 11.11.2025).

51. Liu Y. et al. Construction and Application of User Check-in Spatiotemporal Knowledge Graph Based on Neo4j. *Procedia Comput. Sci*. Elsevier, 2024, Vol. 242, pp. 609–616.
52. Monteiro J. et al. Experimental Evaluation of Graph Databases: JanusGraph, Nebula Graph, Neo4j, and TigerGraph. *Appl. Sci*., 2023, Vol. 13, Multidisciplinary Digital Publishing Institute, 2023. Vol. 13, no. 9, p. 5770.
53. Stolknovenie poezdov na stancii Knyazhaya—Vikipediya [Train collision at Knyazhaya station—Wikipedia]. Available at: https://ru.wikipedia.org/wiki/Stolknovenie_poezdov_na_stancii_Knyazhaya (accessed 11.11.2025).

Информация об авторе

А. В. Вицентий — кандидат технических наук, старший научный сотрудник.

Information about the author

A. V. Vicentiy — Candidate of Science (Tech.), Senior Research Fellow.

Статья поступила в редакцию 21.11.2025; одобрена после рецензирования 24.11.2025; принята к публикации 25.11.2025.
The article was submitted 21.11.2025; approved after reviewing 24.11.2025; accepted for publication 25.11.2025.

Научная статья
УДК 528.8, 622
doi:10.37614/2949-1215.2025.16.3.011

ДИСТАНЦИОННЫЙ МЕТОД ОПРЕДЕЛЕНИЯ ВОДОНАСЫЩЕННОСТИ ОБЪЕКТОВ НАЗЕМНОЙ ГОРНОЙ ИНФРАСТРУКТУРЫ

Михаил Владимирович Мелихов ✉
Горный институт Кольского научного центра Российской академии наук, Апатиты, Россия, m.melikhov@ksc.ru✉, https://orcid.org/0000-0001-8283-2799

Аннотация

В статье рассмотрены вопросы, связанные с опытом эксплуатации и диагностикой состояния водонасыщенности горнопромышленных гидротехнических сооружений в Арктике. Представлен метод интеллектуального дистанционного площадного мониторинга и оценки стресс-состояний поверхности техногенных массивов, основанный на изучении степени водонасыщенности среды по данным мультиспектральных инфракрасных оптико-электронных спутниковых систем. Оценка производится на основе критерия, связывающего водонасыщенность среды со склонностью грунтов к развитию опасных экзогенных геологических процессов. Особенностью метода является возможность автоматизированного дешифрования и геопространственного анализа космоснимков с помощью машинного зрения. На примере действующих хвостохранилищ горных предприятий показаны опыт и результаты спутниковой съемки и оценки степени водонасыщенности техногенных грунтов в задачах контроля и управления промышленными рисками.

Ключевые слова:

Арктика, горные предприятия, техногенные массивы, горные отходы, водонасыщенность грунтов, опасные геологические процессы, риски, мониторинг, дистанционное зондирование

Для цитирования:

Мелихов М. В. Дистанционный метод определения водонасыщенности объектов наземной горной инфраструктуры // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 154–161. doi:10.37614/2949-1215.2025.16.3.011.

Original article

REMOTE SENSING METHOD FOR DETERMINING WATER SATURATION OF GROUND-BASED MINING INFRASTRUCTURE FACILITIES

Mikhail V. Melikhov ✉
Mining institute of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia, m.melikhov@ksc.ru✉, https://orcid.org/0000-0001-8283-2799

Abstract

This article examines issues related to the operating experience and diagnostics of water saturation in mining hydraulic structures in the Arctic. It presents a method for intelligent remote area monitoring and assessment of surface stress conditions of man-made structures, based on studying the degree of water saturation of the environment using data from multispectral infrared optical-electronic satellite systems. The assessment is based on a criterion linking the water saturation of the environment with the susceptibility of soils to the development of hazardous exogenous geological processes. A distinctive feature of this method is the ability to automatically decipher and geospatially analyze satellite images using machine vision. Using operating tailings dams at mining enterprises as an example, the article presents the experience and results of satellite imaging and assessment of the degree of water saturation of man-made soils for monitoring and managing industrial risks

Keywords:

the Arctic, mining enterprises, man-made massifs, mining waste, soil water saturation, hazardous geological processes, risks, monitoring, remote sensing

For citation:

Melikhov M. V. Remote sensing method for determining water saturation of ground-based mining infrastructure facilities. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 154–161. doi:10.37614/2949-1215.2025.16.3.011.

Введение

В настоящее время при разработке месторождений в Арктике к актуальным проблемам следует отнести задачи контроля и диагностики построенных в техногенных массивах особо опасных гидротехнических сооружений (хвостохранилищ, отстойников т. д.) и других горных объектов (карьеров, отвалов и т. д.) для своевременного принятия правильных управленческих решений с учетом негативного влияния воды. Такая ситуация обусловлена сложной спецификой вопросов и подходов к комплексному обеспечению их промышленной безопасности при предотвращении катастрофических рисков аварий техногенного характера [1]. В рассматриваемых условиях необходимо оперативно отслеживать состояние водонасыщенности техногенных грунтов и других факторов на достаточно большой площади в экстремальных климатических условиях в реальном времени [2; 3]. К ключевым риск-факторам здесь можно отнести: неконтролируемое изменение и превышение уровня грунтовых вод или перелив воды через дамбу либо защитное ограждение, подтопления территории, геологические явления и пыление отходов, а также плохое состояние инфраструктуры и недостатки производственных систем мониторинга и управления [4–9].

Как правило, диагностика и оценка степени водонасыщенности на рассматриваемых объектах проводятся дискретно с помощью лабораторных испытаний выборочных образцов грунта (включая определение влажности и расчет коэффициента водонасыщения), с помощью полевых методов, таких как визуальные и гидрологические исследования, а также анализа косвенных признаков.

Принимая во внимание вышесказанное, одним из перспективных решений является применение космических и цифровых технологий в области дистанционного зондирования Земли, что в техническом аспекте может реализовываться в составе действующей системы наблюдений благодаря высокой степени интеграции и взаимозаменяемости данных в едином цифровом пространстве [10]. Современные спутниковые системы способны обеспечить постоянный контроль и мониторинг всей территории опасных поднадзорных и бесхозных объектов с некоторыми техническими ограничениями с точки зрения детализации и периодичности съемок в северных широтах. На практике контроль и управление техногенными рисками с использованием спутниковых систем производятся посредством создания и анализа геоинформационной продукции (карт, моделей, геоданных и др.), что позволяет решать различные производственные задачи [11–17]. Здесь важно отметить, что в соответствии с действующим законодательством данные с космических аппаратов имеют юридическую силу, являясь достоверным и проверенным источником информации. При этом следует иметь в виду, что спутниковые наблюдения должны основываться на обоснованном выборе методических приемов и программно-технического инструментария с учетом конкретных задач и местных условий.

Цель и задачи исследования заключались в совершенствовании методов дистанционного площадного мониторинга и оценки состояния водонасыщенности поверхности техногенных массивов с применением автоматизированных спутниковых систем в обеспечении промышленной и экологической безопасности горнотехнических объектов.

Методика

В Горном институте Кольского научного центра Российской академии наук в рамках комплексных междисциплинарных фундаментальных исследований [18] разработан новый метод дистанционной площадной гидрогеомеханической съемки и определения степени водонасыщенности поверхности техногенных массивов с помощью машинного зрения, заключающийся в использовании мультиспектральных оптико-электронных спутниковых систем Sentinel-2 (рис. 1) путем комбинации ближнего инфракрасного (NIR) и коротковолнового инфракрасного (SWIR) диапазонов электромагнитного спектра на основе стандартизованного нормализованного разностного водного индекса Normalized Difference Moisture Index (NDMI).

Космические аппараты (КА) Sentinel-2 L2A оснащены камерой MSI с 12 каналами (RGB + NIR + SWIR), имеющей комбинированное разрешение 10/20 м и производящей съемку в видимом/инфракрасном режимах [19] соответственно. Данные системы проходят ежегодную регламентную проверку и сертификацию на точность и достоверность геоданных. Сбор и обработка космоснимков производится с помощью программ SNAP, Sentinel Hub EO Browser, ArcGIS / Esri / QGIS и Google Earth.

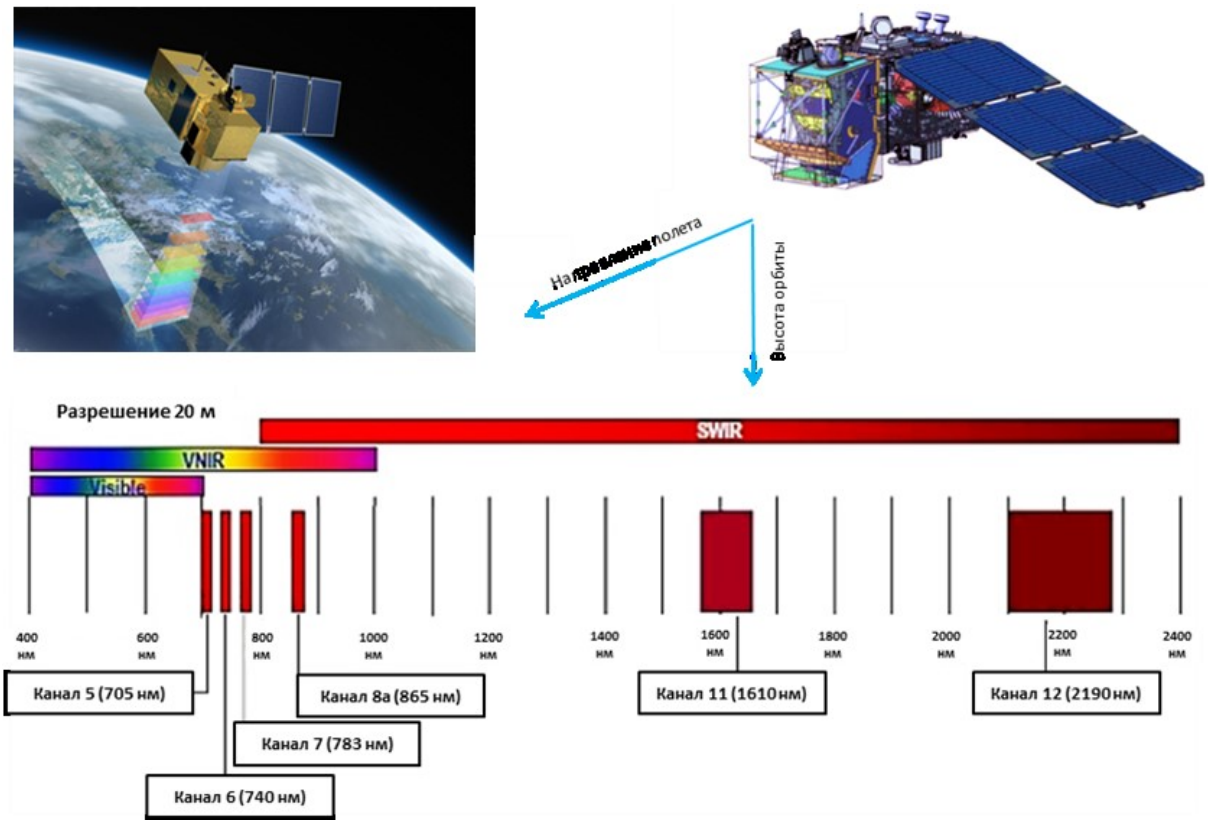


Рис. 1. Общий вид и спектральные характеристики съемочной аппаратуры космических аппаратов Sentinel-2 [19]

В решении поставленной задачи использовался критерий NDMI, позволяющий определить содержание воды и стресс-состояние исследуемого объекта. Расчет показателя NDMI осуществлялся с помощью ПО с открытым исходным кодом. Справочный диапазон значений NDMI составляет от -1 до 1 ед., где отрицательные значения (NDMI = 0 ÷ -1 ед.) соответствуют неводонасыщенным (сухим) средам и, наоборот, положительные значения (NDMI = 0 ÷ 1 ед.) соответствуют водонасыщенным средам. В прикладном значении данный показатель может характеризовать степень водонасыщенности поверхности грунтов и их склонность к развитию и проявлению опасных явлений (фильтрации, суффозии, оплывины, оползни, просадки, выветривание, пыление и т. д.).

Формула расчета NDMI в зависимости от спутниковых систем [20]:

$$\text{Sentinel-2: NDMI} = (B08 - B11) / (B08 + B11); \quad (1)$$

$$\text{Landsat 4-5: TM NDMI} = (B04 - B05) / (B04 + B05); \quad (2)$$

$$\text{Landsat 7: ETM + NDMI} = (B04 - B05) / (B04 + B05); \quad (3)$$

$$\text{Landsat 8: NDMI} = (B05 - B06) / (B05 + B06); \quad (4)$$

$$\text{MODIS: NDMI} = (B02 - B06) / (B02 + B06), \quad (5)$$

где Bn — стандартные обозначения спектральных каналов электромагнитного спектра съемочной аппаратуры.

В процессе спутниковых наблюдений осуществляется идентификация и картирование выявленных зон с разной степенью водонасыщенности на поверхности техногенных массивов на основе их отражательной способности по границам отличенных сред в зависимости от содержания воды в грунтах с учетом реальных местных особенностей. Выполняется комплексный анализ исходных

данных инженерно-геологических и гидрогеологических изысканий, который включает визуальные, полевые и лабораторные результаты исследований, а также результаты камеральной обработки гидрогеологических данных по содержанию и распределению воды в грунтах. Дешифрование и геопространственный анализ спутниковых данных выполнен с использованием визуализированных и автоматизированных методов на основе разработанной классификации грунтов по степени водонасыщенности, приведенной в таблице. Производится автоматическое построение и визуализация изображений в выбранной области по заданным параметрам съемки посредством использования встроенных программных алгоритмов и инструментария. Выполняется геометризация и оконтуривание исследуемого объекта на основе автоматических расчетов и статистического анализа показателя NDMI с определением диапазона его значений в пределах выделенных границ площади территории в указанный период времени. Проводится обобщенный интегрированный анализ и проверка спутниковых данных посредством их сопоставления с результатами выборочного лабораторного отбора проб грунтов с определением точек их географического местоположения и с другими натурными данными, а также совмещения оптических и инфракрасных космоснимков с целью повышения их надежности и достоверности.

Классификация грунтов по водонасыщенности на основе индекса NDMI

№	Значения NDMI	Типы грунтов
1	0,2* ÷ 1	Водонасыщенные
2	-0,2* ÷ 0,2*	Средней водонасыщенности
3	-0,2* ÷ -1	Неводонасыщенные (сухие)

* Условный эмпирический показатель (для условий апатит-нефелиновых месторождений NDMI = -0,3 ÷ 0,96; где при NDMI >0,24 и <-0,26 существуют риски развития гидрогеомеханических фильтрационно-деформационных и геологических экзогенных процессов соответственно).

Результаты и обсуждение

Практическая реализация подходов и методов показана на примере хвостохранилищ апатит-нефелиновых горных предприятий на территории Мурманской области.

Мультиплощадная дистанционная спутниковая съемка хвостохранилищ АНОФ-2 и АНОФ-3 АО «Апатит» (рис. 2) выполнена с помощью КА Sentinel-2 L2A (с учетом атмосферной коррекции данных) с целью изучения степени водонасыщения техногенных намывных откосов и выявления признаков возможных поверхностных проявлений экзогенных геологических процессов в обеспечении промышленной безопасности гидротехнических сооружений. Площадь обследования составила: АНОФ-2 и АНОФ-3 — 5 350 000 и 8 700 000 м² соответственно. В общей сложности собрано и проанализировано 264 космоснимка среднего и высокого пространственного разрешения в режиме облачности до 10 %.

По результатам спутниковых наблюдений посредством интерпретации космоснимков на основе показателя NDMI и обобщенного сравнительного интегрированного анализа разносторонних данных были установлены закономерности в динамике изменения площади и состояния техногенных намывных откосов хвостохранилищ по степени водонасыщенности. Были определены точные периоды сезонного обнажения откосов и наиболее стрессовые ситуации на рассматриваемых объектах. Сезонное обнажение техногенных намывных откосов хвостохранилищ происходит в мае-сентябре, а в другие (зимние) месяцы они находятся в заснеженном состоянии. Период снеготаяния характеризуется резким и интенсивным обнажением большей части поверхности откосов хвостохранилищ со снижением их стабильного и устойчивого состояния. Максимальные значения показателя площади обнаженных откосов хвостохранилищ относятся ко второй половине летнего периода (июль-август). Примечательно, что наиболее негативное воздействие экзогенных процессов на близлежащие населенные пункты происходит непосредственно в это время года. На основе данных мониторинга в отдельных секторах исследуемых сооружений выявлены и идентифицированы признаки и механизмы формирования локальных зон водного стресса

(зон пониженного или повышенного стресса, т. е. сильно или недостаточно увлажненных зон соответственно). Проявление первых способствует развитию опасных фильтрационно-деформационных (суффозий, просадок, оплывин и т. д.), а вторых — экзогенных (в частности, пыления частиц минеральных отходов) процессов и явлений.

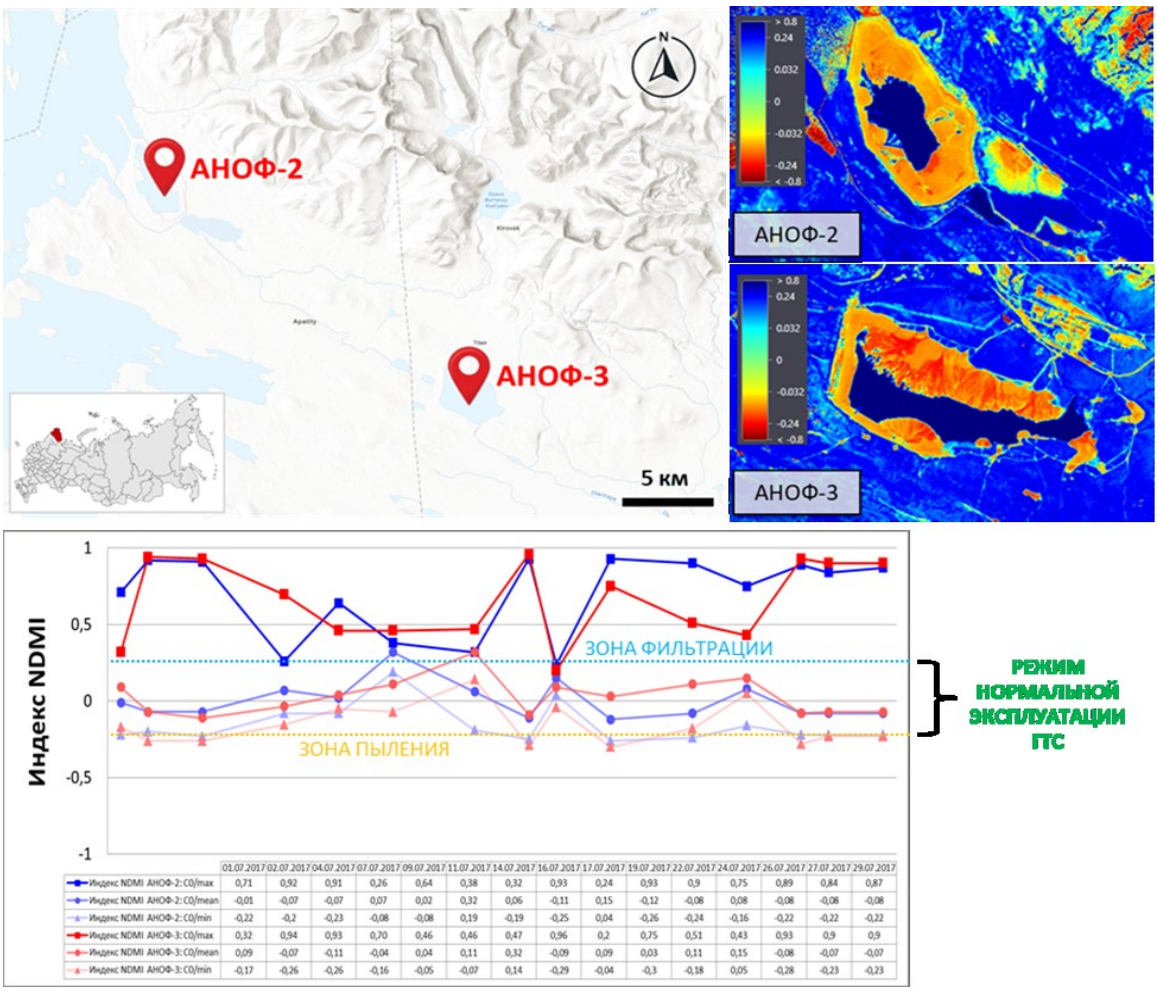


Рис. 2. Результаты мониторинга и оценки степени водонасыщенности техногенных массивов на примере хвостохранилищ апатит-нефелиновых горных предприятий с помощью спутниковых систем Sentinel-2 на основе индекса NDMI

Динамика изменения состояния техногенных намывных откосов по степени водонасыщенности на рассматриваемых хвостохранилищах на основе показателя NDMI представлена на рис. 2. и отображает закономерности сезонных колебаний водонасыщенности грунтов. Результаты оценки степени водонасыщенности техногенных намывных откосов хвостохранилищ имеют высокую сходимость и корреляцию между собой по выбранному критерию, включая данные выборочного отбора проб грунтов.

Заключение

В рамках исследования разработаны новый подход и метод к дистанционной интеллектуальной площадной съемке различных техногенных массивов (карьеров, хвостохранилищ, отвалов, дамб, плотин и др.) с использованием мультиспектральных инфракрасных оптико-электронных спутниковых систем путем автоматизированного дешифрования и геопространственного анализа космоснимков

с помощью машинного зрения на основе критерия, характеризующего стресс-состояние поверхности исследуемых объектов по содержанию воды. Приведена классификация грунтов по водонасыщенности на основе стандартизованного нормализованного разностного водного индекса для оперативной площадной оценки состояния поверхности техногенных грунтов. На примере действующих горных предприятий в Арктической зоне РФ показаны опыт и результаты космического мониторинга хвостохранилищ в задачах контроля и диагностики с целью выявления и идентификации зон с разной степенью водонасыщенности грунтов. Подход может быть успешно применен на других промышленных объектах. Результаты спутниковых наблюдений могут способствовать повышению эффективности защитных мероприятий и рациональному использованию ресурсов в принятии правильных и эффективных управленческих решений на горном предприятии в обеспечении нормального режима эксплуатации инженерных сооружений при борьбе с опасными экзогенными геологическими процессами и явлениями.

Список источников

1. Калашник А. И. Комплексные исследования и мониторинг хвостохранилищ горнопромышленных предприятий Кольского региона // Горный журнал. 2020. № 9. С. 101–106. doi: 10.17580/gzh.2020.09.15.
2. Melikhov M. V., Kalashnik A. I., Ostapenko S. P., Lebedik E. Yu. Integrated approach to remote monitoring of waterworks facilities in the mining industry using space and digital technologies // Journal of Mining Science. 2025. Vol. 61 (1 P. 155–164. https://doi.org/10.1134/S1062739125010168.
3. Kalashnik N. A. Influence of water filtration rate on the functionality of the mining tailings dam // Journal of Physics Conference Series. 2022. Vol. 2388, No. 1. Article 012149. doi:10.1088/1742- 6596/2388/1/012149.
4. WMTF: World Mine Tailings Failures—from 1915 [Электронный ресурс]. URL: https://worldminetailingsfailures.org/ (дата обращения: 15.11.2025).
5. Zare M., Nasategay F., Gomez J. A., Far M. A., Sattarvand J. Review of tailings dam safety monitoring guidelines and systems // Minerals. 2024. 14 (6). 551. https://doi.org/10.3390/min14060551.
6. Franks D., Stringer M., Torres-Cruz L., Baker E. Tailings facility disclosures reveal stability risks // Scientific Reports. 2021. March. 11 (1). doi:10.1038/s41598-021-84897-0.
7. Adamo N., Al-Ansari N., Sissakian V., Laue J., Knutsson S. Dam safety: monitoring of tailings dams and safety reviews // Journal of Earth Sciences and Geotechnical Engineering. 2021. Vol. 11, No. 1. P. 249–289. https://doi.org/10.47260/jesge/1117.
8. Clarkson L., Williams D. Critical review of tailings dam monitoring best practice // International Journal of Mining, Reclamation and Environment. 2020. Vol. 34, Iss. 2. P. 119–148. doi:10.1080/17480930.2019.1625172.
9. Амосов П. В., Бакланов А. А., Горячев А. А., Конина О. Т., Красавцева Е. А., Макаров Д. В., Маслобоев В. А., Ригина О. Ю., Светлов А. В. Пыление хвостов обогащения апатит-нефелиновых руд: экологическая проблема и пути ее решения. Апатиты: ФИЦ КНЦ РАН, 2023. 170 с. doi:10.37614/978.5.91137.505.8.
10. Мелихов М. В. Особенности геоинформационного космического мониторинга горнопромышленных природно-технических систем // Горный информационно-аналитический бюллетень (научно-технический журнал). 2022. № 12–1. С. 29–41. doi: 10.25018/0236_1493_2022_121_0_29.
11. Girija R., Mayappan S. Mapping of mineral resources and lithological units: a review of remote sensing techniques // International Journal of Image and Data Fusion. 2019. Vol. 10: 2. P. 79–106. doi: 10.1080/19479832.2019.1589585.
12. Loginov D. S. Web technologies in cartographic support of geological exploration // Proceedings of the 30th International Cartographic Conference (ICC 2021), 14–18 december 2021, ICA, 2021, Vol. 4 (68). doi: 10.5194/ica-proc-4-68-2021.
13. Song W., Song W., Gu H., Li F. Progress in the remote sensing monitoring of the ecological environment in mining areas // Environmental Research and Public Health. 2020. 17 (6). P. 1846. https://doi.org/10.3390/ijerph17061846.
14. Werner T., Bebbington A., Gregory G. Assessing impacts of mining: Recent contributions from GIS and remote sensing // The Extractive Industries and Society. 2019. Vol. 6, Iss. 3. P. 993–1012. https://doi.org/10.1016/j.exis.2019.06.011.
15. Zheng M., Deng K., Fan H., Du S. Monitoring and analysis of surface deformation in mining area based on InSAR and GRACE // Remote Sensing. 2018. Vol. 10, No. 9. P. 1392. doi:10.3390/rs10091392.
16. Мелихов М. В. Мультиплощадной космический мониторинг хранилищ отходов горного производства в Арктике // Горная промышленность. 2024. № 5S. С. 21–27. https://doi.org/10.30686/1609-9192-2024-5S-21-27.
17. Мелихов М. В. Геоинформационное обеспечение складирования горнопромышленных отходов на основе космических и цифровых технологий // Материалы V Всероссийской научно-технической конференции

«Цифровые технологии в горном деле» (г. Апатиты, 13–16 июня 2023 г.). Апатиты: ФИЦ КНЦ РАН, 2023. С. 32–33. doi:10.37614/978.5.91137.491.4.

18. Калашник А. И., Максимов Д. А., Калашник Н. А., Дьяков А. Ю., Запорожец Д. В., Мелихов М. В. Многоуровневые комплексные исследования и мониторинг хвостохранилищ горнодобывающих предприятий Северо-Западной части Российского сектора Арктики. Апатиты: КНЦ РАН, 2022. 313 с. doi:10.37614/978.5.91137.465.5.
19. Sentinel-2 mission guide [Электронный ресурс]. URL: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2 (дата обращения: 15.11.2025).
20. Normalized Difference Moisture Index (NDMI) [Электронный ресурс]. URL: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/ndmi/ (дата обращения: 15.11.2025).

References

1. Kalashnik A. I. Comprehensive research and monitoring of tailings storage facilities of mining enterprises in the Kola region. *Mining Journal*, 2020, no. 9, pp. 101–106. (In Russ.). doi: 10.17580/gzh.2020.09.15.
2. Melikhov M. V., Kalashnik A. I., Ostapenko S. P., Lebedik E. Yu. Integrated approach to remote monitoring of waterworks facilities in the mining industry using space and digital technologies. *Journal of Mining Science*, 2025, vol. 61 (1), pp. 155–164. https://doi.org/10.1134/S1062739125010168.
3. Kalashnik N. A. Influence of water filtration rate on the functionality of the mining tailings dam. *Journal of Physics Conference Series*, 2022, vol. 2388, no. 1, article 012149. doi:10.1088/1742- 6596/2388/1/012149.
4. WMTF: World Mine Tailings Failures—from 1915. Available at: https://worldminetailingsfailures.org/ (accessed 15.11.2025).
5. Zare M., Nasategay F., Gomez J. A., Far M. A., Sattarvand J. Review of tailings dam safety monitoring guidelines and systems. *Minerals,* 2024, 14 (6), 551. https://doi.org/10.3390/min14060551.
6. Franks D., Stringer M., Torres-Cruz L., Baker E. Tailings facility disclosures reveal stability risks. *Scientific Reports*, 2021, march, 11 (1). doi:10.1038/s41598-021-84897-0.
7. Adamo N., Al-Ansari N., Sissakian V., Laue J., Knutsson S. Dam safety: monitoring of tailings dams and safety reviews. *Journal of Earth Sciences and Geotechnical Engineering*, 2021, vol. 11, no. 1, pp. 249–289. https://doi.org/10.47260/jesge/1117.
8. Clarkson L., Williams D. Critical review of tailings dam monitoring best practice. *International Journal of Mining, Reclamation and Environment*, 2020, vol. 34, issue 2, pp. 119–148. doi:10.1080/17480930.2019.1625172.
9. Amosov P. V., Baklanov A. A., Goryachev A. A., Konina O. T., Krasavtseva E. A., Makarov D. V., Masloboev V. A., Rigina O. Yu., Svetlov A. V. *Dusting of apatite-nepheline ore beneficiation tailings: an environmental problem and ways to solve it*. Apatity, FRC KSC RAS, 2023, 170 p. (In Russ.). doi:10.37614/978.5.91137.505.8.
10. Melikhov M. V. Features of geoinformation space monitoring of mining natural-technical systems. *Mining information and analytical bulletin (scientific and technical journal)*, 2022, no. 12–1, pp. 29–41. (In Russ.). doi:10.25018/0236_1493_2022_121_0_29.
11. Girija R., Mayappan S. Mapping of mineral resources and lithological units: a review of remote sensing techniques. *International Journal of Image and Data Fusion*, 2019, vol. 10: 2, pp. 79–106. doi:10.1080/19479832.2019.1589585.
12. Loginov D. S. Web technologies in cartographic support of geological exploration. *Proceedings of the 30th International Cartographic Conference (ICC 2021)*, Florence, Italy: The International Cartographic Association (ICA), 2021, vol. 4 (68). doi: 10.5194/ica-proc-4-68-2021.
13. Song W., Song W., Gu H., Li F. Progress in the remote sensing monitoring of the ecological environment in mining areas. *Environmental Research and Public Health*, 2020, 17 (6), p. 1846. https://doi.org/10.3390/ijerph17061846.
14. Werner T., Bebbington A., Gregory G. Assessing impacts of mining: Recent contributions from GIS and remote sensing. *The Extractive Industries and Society*, 2019, vol. 6, issue 3, pp. 993–1012. https://doi.org/10.1016/j.exis.2019.06.011.
15. Zheng M., Deng K., Fan H., Du S. Monitoring and analysis of surface deformation in mining area based on InSAR and GRACE. *Remote Sensing*, 2018, vol. 10, no. 9, pp. 1392. doi:10.3390/rs10091392.
16. Melikhov M. V. Multi-area space monitoring of mining waste storage facilities in the Arctic. *Mining Industry*, 2024, no. 5S, pp. 21–27. (In Russ.). https://doi.org/10.30686/1609-9192-2024-5S-21-27.
17. Melikhov M. V. Geoinformation support for mining waste storage based on space and digital technologies. *Proceedings of the V All-Russian Scientific and Technical Conference "Digital Technologies in Mining"*. Apatity, Federal Research Center of the Kola Science Center of the Russian Academy of Sciences, 2023, pp. 32–33. (In Russ.). doi:10.37614/978.5.91137.491.4.

18. Kalashnik A. I., Maksimov D. A., Kalashnik N. A., Dyakov A. Yu., Zaporozhets D. V., Melikhov M. V. *Multilevel integrated studies and monitoring of tailings ponds of mining enterprises in the Northwestern part of the Russian sector of the Arctic*. Apatity, KSC RAS, 2022, 313 p. (In Russ.). doi: 10.37614/978.5.91137.465.5.
19. Sentinel-2 mission guide. Available at: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2 (accessed 15.11.2025).
20. Normalized Difference Moisture Index (NDMI). Available at: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/ndmi/ (accessed 15.11.2025).

Информация об авторе

М. В. Мелихов — кандидат технических наук, старший научный сотрудник.

Information about the author

M. V. Melikhov — Candidate of Science (Tech.), Senior Research Fellow.

Статья поступила в редакцию 28.11.2025; одобрена после рецензирования 01.12.2025; принята к публикации 05.12.2025.
The article was submitted 28.11.2025; approved after reviewing 01.12.2025; accepted for publication 05.12.2025.

Научная статья
УДК 004.9
doi:10.37614/2949-1215.2025.16.3.012

ИДЕНТИФИКАЦИЯ И КЛАССИФИКАЦИЯ ОПАСНЫХ ОБЪЕКТОВ

Сергей Юрьевич Яковлев[1, 3], Алексей Сергеевич Шемякин[2]✉
[1, 2]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[3]Мурманский арктический университет, Апатиты, Россия
[1]s.yakovlev@ksc.ru, https://orcid.org/0000-0001-6433-2096
[2]a.shemyakin@ksc.ru✉, https://orcid.org/0000-0001-5308-5456

Аннотация
Рассматриваются информационные аспекты идентификации и классификации различных видов опасных объектов. Принимаются во внимание исторические изменения, а также освещается текущее содержание процессов выделения и категорирования объектов. Намечаются направления и методы автоматизации.

Ключевые слова:
опасные объекты, идентификация и классификация, оценка риска, информационная поддержка управления безопасностью

Благодарности:
работа выполнена в рамках темы научно-исследовательской работы FMEZ-2025-0054 «Методы и информационные технологии мониторинга и управления региональными критическими инфраструктурами Арктической зоны Российской Федерации».

Для цитирования:
Яковлев С. Ю., Шемякин А. С. Идентификация и классификация опасных объектов // Труды Кольского научного центра РАН. Серия: Технические науки. 2025. Т. 16, № 3. С. 162–176. doi:10.37614/2949-1215.2025.16.3.012.

Original article

IDENTIFICATION AND CLASSIFICATION OF HAZARDOUS OBJECTS

Sergey Yu. Yakovlev[1, 3], Alexey S. Shemyakin[2]✉
[1, 2]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[3]Murmansk Arctic University, Apatity, Russia
[1]s.yakovlev@ksc.ru, https://orcid.org/0000-0001-6433-2096
[2]a.shemyakin@ksc.ru✉, https://orcid.org/0000-0001-5308-5456

Abstract
This article examines information aspects of identifying and classifying various types of hazardous objects. Historical changes are considered, and the current processes of identifying and categorizing objects are discussed. Directions and methods for automation are outlined.

Keywords:
hazardous objects, identification and classification, risk assessment, information support for safety management

Acknowledgments:
The work was carried out within the framework of the research topic FMEZ-2025-0054 “Methods and information technologies for monitoring and managing regional critical infrastructures of the Arctic zone of the Russian Federation”.

For citation:
Yakovlev S. Yu., Shemyakin A. S. Identification and classification of hazardous objects. *Trudy Kol'skogo nauchnogo centra RAN. Seriya: Tekhnicheskie nauki* [Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences], 2025, Vol. 16, No. 3, pp. 162–176. doi:10.37614/2949-1215.2025.16.3.012.

Введение

Проблема идентификации и классификации опасностей — постоянная для теории безопасности и риска, требующая регулярных обновлений и адаптаций в связи с изменениями в экономике регионов,

обнаружением новых видов возможных аварий и инцидентов, а также вследствие появления оригинальных технологий мониторинга и обработки данных. Отметим междисциплинарный характер проблемы.

В нормативной базе и литературе упоминаются различные виды опасных объектов. Начнем рассмотрение с самой «многочисленной» группы — опасные производственные объекты (ОПО), а затем перейдем к другим видам. Вначале будут описываться идентификационные признаки, а затем — классификационные.

Опасные производственные объекты

Ранее действующие правовые нормы идентификации ОПО были подробно изложены в работе [1]. Современный федеральный закон №116-ФЗ [2] сохраняет большинство положений относительно опасных производственных объектов. Согласно данному закону, ОПО признаются предприятия, цеха, технологические линии, территории либо иные производства, перечисленные в приложениях 1 и 2 к указанному федеральному закону. Приложения включают следующие типы объектов:

предприятия, работающие с обращением опасных веществ разных классов опасности, характеристики которых описаны в приложении 1 и приложении 2 закона №116-ФЗ;

объекты, эксплуатирующие технику, функционирующую под высоким рабочим давлением (превышающим 0,07 мегапаскалей);

производства, использующие подъемные устройства и краны различного типа;

компании, занимающиеся обработкой и переработкой жидких металлов черных и цветных сплавов;

организации, ведущие добычу полезных ископаемых и обогащение рудных материалов;

предприятия, занятые хранением и переработкой сельскохозяйственной продукции растительного происхождения;

заводы и лаборатории, задействованные в обращении химоружия и специальных химических продуктов;

нефтяные и газовые скважины, установки по добыче нефти, природного газа и газовых конденсатов;

инфраструктуры газоснабжения и потребления газа.

Таким образом, в числе признаков, выделяющих ОПО, фигурируют некоторые процессы, которые признаются опасными (опасные процессы). Рассмотрим эти признаки и процессы подробнее.

Перечислим ключевые направления формирования критериев отнесения процессов и объектов к ОПО согласно действующему законодательству:

1. Обращение с опасными веществами: определяется широкий спектр операций (производство, применение, преобразование, накопление, складирование, транспортировка, утилизация), осуществляемых с определенными категориями веществ, включая воспламеняющиеся, окислительные, взрывоопасные, пожароопасные, токсичные вещества, способные нанести ущерб здоровью населения и природной среде. Перечень включает конкретные наименования, такие как аммиак и метилизоцианат, с указанием качественных и количественных показателей, достижение которых влечет признание процесса или объекта опасным.

2. Использование оборудования высокого давления: сюда входят операции эксплуатации устройств, работающих под повышенным давлением пара, газов, жидкости и других сред. Законодательством установлены четкие требования и ограничения по характеристикам используемых веществ, соблюдение которых определяет потенциальную угрозу и классифицирует объект как источник повышенной опасности.

3. Эксплуатация подъемных сооружений: особое внимание уделяется процессам функционирования стационарных грузоподъемных машин, за исключением лифтов общего пользования и платформ для перемещения маломобильных граждан. Дополнительные риски связаны с использованием эскалаторов метрополитенов, канатных дорог и фуникулеров, эксплуатация которых требует особого подхода к обеспечению безопасности персонала и пассажиров.

4. Процессы с расплавами металлов включают в себя возможные виды процессов/работ (получение, транспортировка, использование) и характеристики расплавов (качественные и/или количественные), при которых соответствующие процессы (и объекты) признаются опасными.

5. Процессы ведения горных работ содержат перечень исключений (не входящих в ОПО) из видов работ, а также работы по обогащению полезных ископаемых.

6. Операции, связанные с заготовкой, хранением и переработкой растительной продукции, характеризуются образованием потенциально взрывоопасных пылевых облаков, подверженных риску возгорания и горения. Отдельно выделяются процессы хранения зерновых культур, продуктов мукомольного производства и кормовых составов, имеющих повышенную склонность к саморазогреву и самонагревательному возгоранию.

7. Деятельность, связанная с химическим вооружением, охватывает комплекс мероприятий по размещению, утилизации боеприпасов и обращение с особыми химическими материалами и компонентами специального назначения.

8. Работы по разведочному бурению и промышленному извлечению углеводородного сырья (нефть, природный газ, газоконденсатные месторождения) подлежат строгому контролю и регламентации.

9. Система распределения и потребления газа подразумевает наличие специализированных инфраструктурных элементов, признанных источниками повышенного риска (газораспределительные пункты, магистральные трубопроводы, распределительная сеть), чьи характеристики определяются установленными качественными и количественными параметрами.

Отметим еще раз, что для процессов и объектов в 116-ФЗ в ряде случаев указаны исключения, и соответствующие объекты не относятся к ОПО.

Попробуем обобщить указанные показатели идентификации ОПО (рис. 1). Выделены направления/виды деятельности/виды работ — опасные процессы. Эти процессы, разновидности процессов считаются опасными: при наличии определенных опасных компонентов (веществ, оборудования, механизмов, иной специфики); при наличии определенных качественных и количественных характеристик (ограничений, атрибутов) процессов или компонентов.

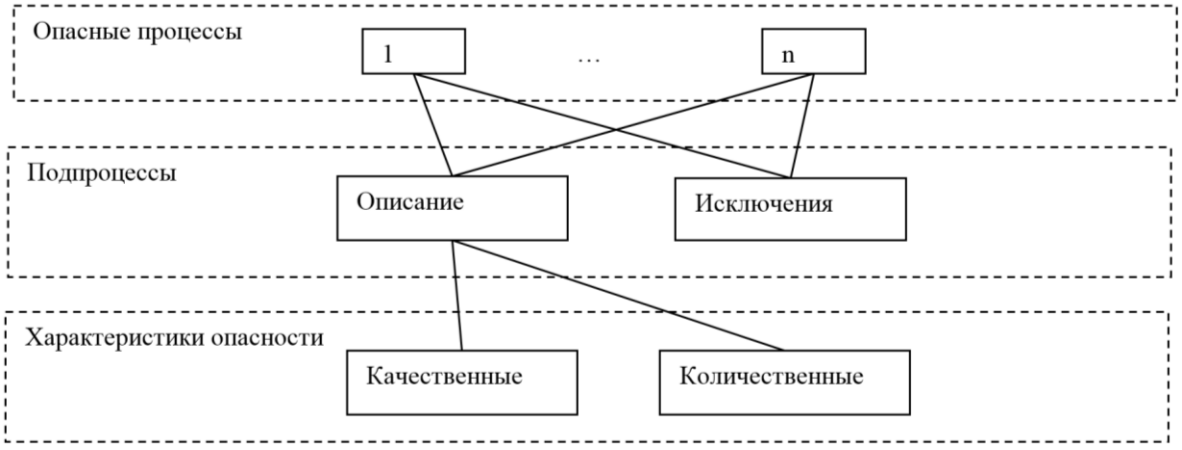


Рис. 1. Схема идентификации ОПО

Перейдем к вопросам категорирования ОПО.

Приложение 2 к 116-ФЗ посвящено классификации (категорированию) объектов по уровням опасности. Выделяются 4 возможных класса опасности: от I (чрезвычайно высокая опасность) до IV (низкая). Рассмотрим, как устанавливаются классы для перечисленных ранее опасных процессов.

1. Процессы с опасными веществами. Возможно присвоение от I до IV классов в зависимости от количества опасных веществ определенных видов, наименований веществ, условий использования (таблицы 1 и 2 приложения 2). Подсчет суммарного количества опасного вещества регулируется рядом положений 116-ФЗ.

2. Процессы с оборудованием под давлением. Присваивается III или IV класс опасности в зависимости от специфики (теплоснабжение) и характеристик использования (давление, температура).

3. Процессы с грузоподъемными механизмами. Возможен III (для подвесных канатных дорог) или IV (для других механизмов) класс опасности.

4. Процессы с расплавами металлов. Устанавливаются II или III классы в зависимости от максимального количества расплава.

5. Процессы ведения горных работ. Возможно установление от I до IV классов в зависимости от специфики и условий (возможности взрывов, выбросов, горных ударов, прорывов воды) подземных работ, а также от видов и объемов открытых работ.

6. Процессы хранения или переработки растительного сырья. Возможен III (для определенных объектов и производств) или IV (для иных вариантов) класс.

7. Для процессов с химическим оружием устанавливается I класс опасности.

8. Процессы добычи нефти и газа. Возможны II, III или IV классы в зависимости от возможных объемов выбросов сернистого водорода.

9. Процессы газораспределения и газопотребления. Возможно установление II или III класса для объектов транспортировки газа в зависимости от его давления.

Общая схема категорирования ОПО приведена на рис. 2.

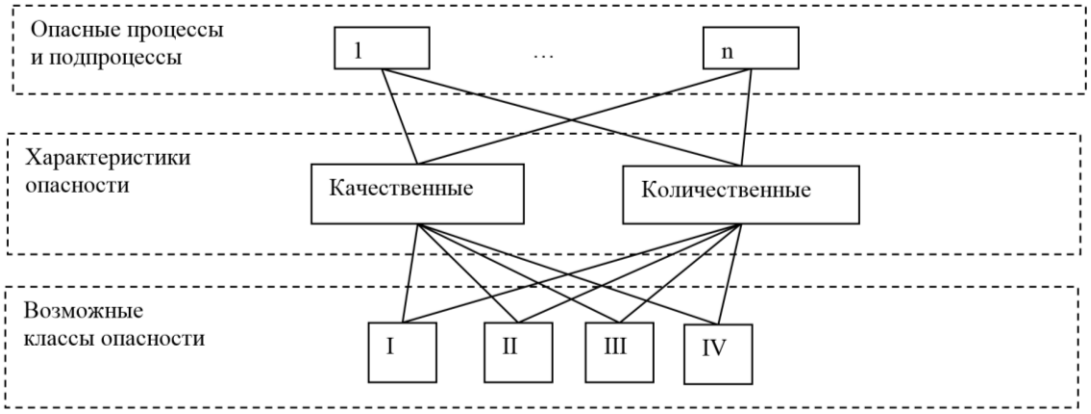


Рис. 2. Схема категорирования ОПО

В случае возможности установления для одного ОПО разных классов опасности устанавливается наиболее высокий класс. Для ОПО II, III или IV классов опасности возможно повышение класса в случае расположения ОПО на особых территориях или акваториях.

Констатируется, что (п. 3 статьи 2 закона 116-ФЗ) разделение объектов на классы проведено в зависимости от уровня потенциальной опасности аварий на них для жизненно важных интересов личности и общества, что не напрямую соответствует упомянутым критериям (признакам) в приложениях 1 и 2 к 116-ФЗ.

Формально присвоение класса опасности ОПО осуществляется при его регистрации в государственном реестре. Важные дополнения к приведенным схемам содержатся в приказе Ростехнадзора об утверждении Требований к регистрации ОПО [3]. В приложении 1 к Требованиям содержатся типовые наименования ОПО, признаки опасности, границы объектов. Перечислено 19 опасных видов/направлений деятельности (опасных процессов), в рамках которых выделяются типовые наименования опасных объектов:

1. Угольная, сланцевая и торфяная промышленность.
2. Горнорудная и нерудная промышленность.
3. Оборот взрывчатых веществ.
4. Нефтегазодобывающий комплекс.
5. Магистральный трубопроводный транспорт.

6. Геолого-разведочные и геофизические работы при разработке месторождений.
7. Химия, нефтехимия, нефтепереработка и другие взрывопожароопасные и вредные производства.
8. Нефтепродуктообеспечение.
9. Системы водоподготовки.
10. Пищевая и масложировая промышленность
11. Газоснабжение.
12. Тепло- и электроэнергетика, оборудование под давлением или при температуре.
13. Металлургическая промышленность.
14. Производство черных и цветных металлов.
15. Стационарные грузоподъемные механизмы, эскалаторы, канатные дороги, фуникулеры.
16. Хранение и переработка растительного сырья.
17. Транспортировка опасных веществ.
18. Добыча минеральных вод.
19. Спецхимия.

Некоторые из перечисленных процессов соответствуют ранее упомянутым (9 наименований), другие более соответствуют отраслям промышленности (ведомственные направления).

Статья [4] посвящена исследованию особенностей идентификации и постановки на учет ОПО применительно к угольным предприятиям. Авторы подчеркивают важность детализированного уточнения отдельных категорий, таких как «углеобработка», «шахтное производство», «карьеры по добыче угля», «фабрики обогащения» и «обогащающие технологии». Отмечается, что регистрация второстепенных структурных подразделений шахт, карьеров и обогатительных предприятий, таких как административные здания, помещения бытового обслуживания, складские комплексы и прочие вспомогательные постройки, является излишней мерой.

Процесс идентификации предполагает выявление всех признаков потенциальной угрозы на предприятии, оценку их качественных и количественных параметров, а также учет технологических процессов и технических установок, соответствующих критериям, указанным в приложении 1 федерального закона № 116-ФЗ. Такой анализ позволяет объективно определить принадлежность объекта к классу ОПО. Анализируя нормативно-правовую базу в сфере промышленной безопасности, авторы указывают, что категорию ОПО определяют как используемые на предприятиях технические средства, так и сами здания и инженерные конструкции.

Регистрация и идентификация шахт, карьеров и обогатительных производств должна основываться на следующих критериях: фиксированный статус производственной единицы (предприятие, отдельные подразделения, рабочие зоны и площади); осуществление конкретных видов деятельности и производственных процессов, предусмотренных в приложении 1 федерального закона № 116-ФЗ. Исходя из предложенного методологического подхода, возможны разные сценарии реализации процедуры идентификации и дальнейшей регистрации ОПО в угольной отрасли.

Потенциально опасные объекты

Наряду с ОПО, в нормативной базе фигурируют потенциально опасные объекты (ПОО). Сошлемся на определение из закона 68-ФЗ [5]: «ПОО — это объект, на котором расположены здания и сооружения повышенного уровня ответственности, либо объект, на котором возможно одновременное пребывание более пяти тысяч человек».

Согласно материалам исследования, приведенным в публикации [6], решение о включении организаций в список ПОО принимается исходя из определенных условий, устанавливаемых соответствующим документом правительства № 1226 [7]. Однако постановление № 1226 ограниченно в своем содержании: оно устанавливает лишь общие принципы формирования критериев оценки степени опасности объектов, непосредственно же показатели (конкретные названия индикаторов и пороговые значения) отсутствуют. Разработку детальных методик расчета критериев поручили Министерству чрезвычайных ситуаций Российской Федерации.

Документ [5] вводит ряд важных дефиниций, среди которых выделяются понятия «здание», «здания и сооружения с повышенной ответственностью», «уровень опасности потенциально опасных объектов», «объект», «организация, подлежащая включению в реестр потенциально опасных объектов», «строительное сооружение». Под зданиями и сооружениями с повышенной степенью ответственности понимаются строения, отнесенные в соответствии с нормами Градостроительного кодекса Российской Федерации [8]: к особо опасным и технически сложным сооружениям (группа 1), уникальным строительным объектам (группа 2).

К группе 1 отнесены:

1. Объекты использования атомной энергии.
2. Гидротехнические сооружения.
3. Сооружения связи.
4. Объекты электросетевого хозяйства.
5. Объекты космической инфраструктуры.
6. Объекты воздушного транспорта.
7. Объекты железнодорожного транспорта.
8. Объекты внеуличного транспорта.
9. Объекты морского порта.
10. Тепловые электростанции и подвесные канатные дороги.
11. Некоторые виды ОПО:

a) объекты I и II классов опасности, связанные с оборотом опасных веществ;

b) объекты, связанные с расплавами черных и цветных металлов (максимальное количество 500 кг и более);

c) объекты ведения горных работ (с исключениями).

К группе 2 относятся объекты капитального строительства, за исключением указанных для (1), обладающие хотя бы одной из количественных характеристик по некоторому параметру (высота, длина консоли или пролета, заглубление подземной части).

При этом отметим, что для ряда объектов указаны исключения (по которым объекты не относятся к ПОО); некоторые виды (группы) объектов регулируются собственным законодательством; в ряде случаев указаны количественные характеристики отнесения к ПОО.

Признаки отнесения к ПОО приведены в общем виде на рис. 3.

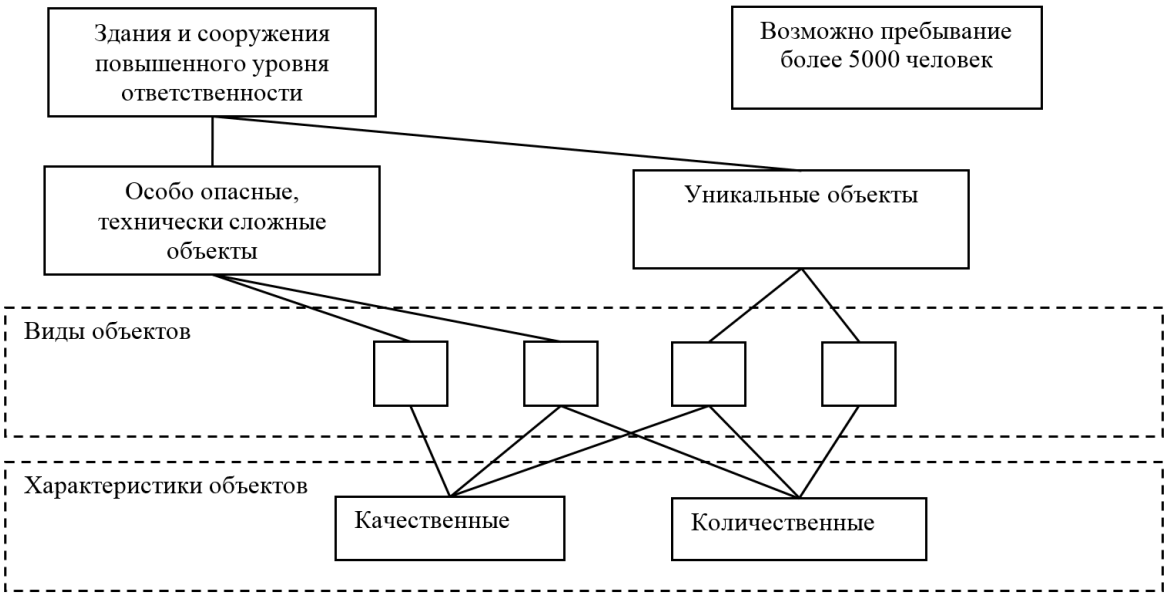


Рис. 3. Признаки отнесения к ПОО

В зависимости от характера (уровня) возможной чрезвычайной ситуации (федеральный, межрегиональный, региональный, межмуниципальный, муниципальный, локальный и ниже) выделяются ПОО шести категорий (уровней) опасности (1-я категория опасности — особо высокий уровень опасности, 2-я — чрезвычайно высокий, 3-я — высокий, 4-я — повышенный, 5-я — средний, 6-я — низкий). Уровни ЧС регулируются постановлением [9] и определяются показателями ущерба (см. далее).

В соответствии с [7] разработан ряд критериев [8; 10]. Так, в [8] для ряда объектов, регулируемых Ростехнадзором, уточнены признаки отнесения к ПОО и категории опасности ПОО: 1) гидротехнические сооружения (с исключениями) I или II классов опасности (установленных в соответствии с законодательством о безопасности гидротехнических сооружений); 2) объекты электросетевого хозяйства, напряжение 330 киловольт и более; 3) тепловые электростанции, мощность 150 мегаватт и выше; 4) некоторые ОПО (см. пункт 11 выше).

В [10] для объектов использования атомной энергии, регулируемых Роспотребнадзором, уточнены качественные и количественные показатели отнесения к ПОО и категорирования ПОО. Отметим, что количественные показатели категорирования соответствуют показателям классификации чрезвычайной ситуации (ЧС) природного и техногенного характера [9] и рассмотрены ниже.

Обобщенная схема категорирования ПОО приведена на рис.4.

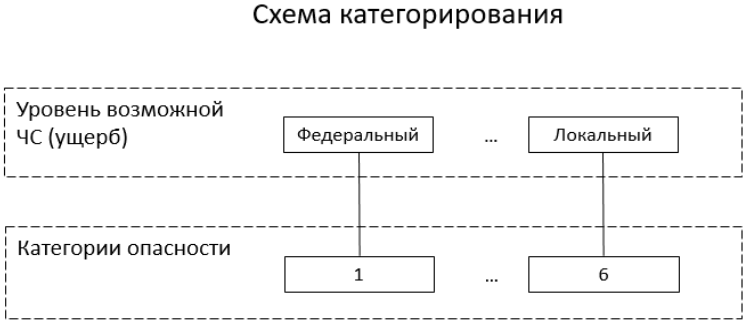


Рис.4. Схема категорирования ПОО

В исследовании [11] затронут важный аспект организации системы безопасности на ПОО. Для эффективного планирования защитных мероприятий предлагается оценивать безопасность объекта одновременно по двум направлениям: 1) степень вероятности возникновения ЧС и риск превращения объекта в источник аварии или инцидента; 2) уровень готовности к минимизации ущерба и снижению негативного воздействия ЧС на население, соседствующие объекты инфраструктуры и окружающую среду.

Объект рассматривается двояко: как потенциальный источник опасности и как структура, нуждающаяся в защите. Соответственно, обеспечение комплексной безопасности должно включать два ключевых компонента: 1) управление уровнем потенциальной угрозы, исходящей от самого объекта; 2) обеспечение необходимой защищенности объекта, включая меры противодействия терроризму и техническим сбоям.

Авторами отмечается, что присвоение категориям объектов конкретных уровней опасности и защищенности носит достаточно произвольный характер. Количество градаций, выбор критериев и нормативные значения устанавливаются человеком на основе текущих нужд и доступной базы данных для оценки каждого показателя. Поэтому система категорирования объектов должна гибко адаптироваться к меняющимся условиям и потребностям. Для повышения точности оценки рекомендуется учесть три основных параметра: уровень физической охраны объекта; условия внутренней технологической защищенности; качество информационной безопасности.

Такой многосторонний подход обеспечивает комплексное выполнение целей национальной политики в области защиты объектов от угроз разного характера — террористического, техногенного и природного происхождения.

Критически важные объекты

Наряду с ОПО и ПОО, в нормативной базе фигурируют критически важные объекты (КВО). Сошлемся также на определение из закона 68-ФЗ [3]: «КВО — это объект, нарушение или прекращение функционирования которого приведет к потере управления экономикой Российской Федерации, субъекта Российской Федерации или административно-территориальной единицы (АТЕ) субъекта Российской Федерации, ее необратимому негативному изменению (разрушению) либо существенному снижению безопасности жизнедеятельности населения». На рис. 5 приведена соответствующая схема идентификации КВО.

Рис. 5. Признаки отнесения к КВО

Таким образом, по сравнению с ОПО и ПОО, основным признаком отнесения объекта к КВО служит не область деятельности (опасный процесс), а возможный результат (ущерб) при нарушении функционирования (например, при аварии) объекта. Поэтому для отнесения объекта к КВО необходимы, главным образом, не качественные, а количественные характеристики опасности.

Исследование [12] подчеркивает, что включение компаний в перечень КВО осуществляется согласно установленным критериям постановления № 1225 [13]. Вопросы идентификации и классификации КВО также детально рассмотрены ранее [14].

Постановление [13] предусматривает обязанность государственных органов и ведомств разрабатывать необходимые нормативные акты, определяющие критерии отбора объектов, а Министерство по чрезвычайным ситуациям осуществляет контроль над процессом формирования методологии. Среди базовых определений, представленных в документе [13]: понятие «категория значимости КВО», терминология «объект» и характеристика объектов, относящихся к числу КВО. Выделены три категории значимости объектов: 1) федеральный уровень значимости: возможные последствия катастроф приводят к утрате управляемости экономикой сразу в двух и более регионах России; 2) региональный уровень значимости: инциденты способны вызвать нарушение управленческих функций в рамках одного конкретного субъекта Федерации; 3) муниципальный уровень значимости: катастрофы вызывают потерю экономического контроля в границах отдельного муниципального образования.

Подлежать такому ранжированию могут любые объекты независимо от формы собственности, если возможная авария способна повлиять на экономику государства, региона или муниципалитета. Критерии отбора подразумевают формирование перечня специфичных индикаторов и порогов, на основе которых объекты распределяют по категориям значимости.

Дополнительные рекомендации представлены в работе [15], где категории значимости расширяются путем введения численных показателей, аналогичных классификации ЧС [9]. Согласно этому документу, объекты федерального уровня значимости характеризуются социальным ущербом свыше 500 пострадавших или общим экономическим и экологическим ущербом более 1,2 млрд рублей; к региональным КВО принадлежат объекты, на которых число пострадавших

колеблется от 50 до 500 человек, а общий экономический ущерб находится в диапазоне от 12 млн до 1,2 млрд рублей; муниципальные КВО охватывают случаи, где число пострадавших не превышает 50 человек, а ущерб ограничен суммой менее 12 млн рублей.

Таким образом, в [15] уточняются и дополняются количественные ориентиры, отражаемые в тексте постановления [13]. При этом в самом постановлении [9] предусмотрены аналогичные характеристики социальных, экономических и экологических потерь, однако отсутствуют аналоги межрегиональных, межмуниципальных и локальных кризисов, наблюдающихся в категории значимых объектов.

Альтернативный подход представлен в [16], где значимость объектов дополнительно оценивается временным фактором — продолжительность дисфункции объекта после происшествия должна превышать 24 часа. Это дополнительное условие вводится наряду с территориальной привязкой (количество вовлеченных регионов или муниципальных образований).

Принципы категорирования КВО обобщены на рис. 6.

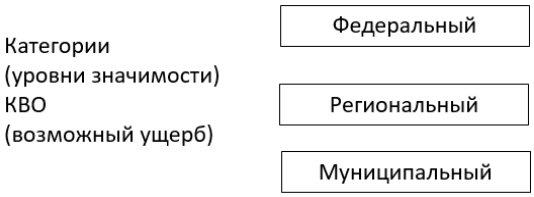


Рис. 6. Схема категорирования КВО

В статье [17] рассматриваются вопросы классификации КВО. Предложены определения основных понятий, группы критериев отнесения к КВО, классификации уровней критической важности (значимости) КВО. Существенное внимание уделено планирующему документу — паспорту безопасности КВО (см. также [18]), который считается основным документом по безопасности КВО, содержащим информацию, характеризующую важность (значимость) объекта. Анализируются возможности отнесения объекта одновременно к КВО и ПОО.

Опасные объекты и критические инфраструктуры

Вопросы идентификации и категорирования опасных объектов касаются проблем взаимодействия этих объектов с федеральными и региональными опасными надсистемами, в частности с критическими инфраструктурами. В работе [14] были намечены некоторые направления информационного обеспечения учета такого взаимодействия. В [19] отмечалась целесообразность создания и использования онтологий, единой проблемно-ориентированной информационно-аналитической среды, подчеркивалась важность учета вышестоящих структур (надсистем) и критических функций. В работе [20] предложен подход к системной оценке потенциальных рисков взаимодействия объектов с инфраструктурами. В [21] представлен информационный проект учета взаимодействующих надсистем, внутренней структуры объекта, всевозможных ЧС.

В качестве примера перечислим некоторые критические инфраструктуры (КИ) Арктической зоны Российской Федерации (АЗРФ): 1) транспортные; 2) энергетические; 3) космические; 4) горнодобывающие; 5) гидротехнические; 6) оборот опасных веществ; 7) металлургические; 8) нефтегазовые; 9) химические; 10) связь; 11) трубопроводные; 12) продовольственные.

Также в качестве примера можно привести региональные КИ (Мурманской области).

Возможные КИ АЗРФ (Мурманской области): 1) транспортные (воздушные, внеуличные, морские, железнодорожные); 2) энергетические (атомные, электросетевые, ГЭС); 3) горнодобывающие (подземные горные работы, открытые горные работы, работы по обогащению).

Каждая из КИ имеет свои показатели жизнедеятельности и опасности [22]. Так, функциональные возможности транспортных структур могут быть охарактеризованы темпами доставки и стоимостью «грузов».

Каждой из КИ в рамках Федерации, округа, региона, АТЕ может быть приписана какая-то важность *p*, отражающая значение КИ в определенных административно-территориальных рамках (сумма важностей равна единице). На наш взгляд, эти важности могут меняться в зависимости от текущей обстановки (мирное или военное время, чрезвычайные ситуации, санкции и другие ограничения).

Перечень региональных и федеральных КИ не утвержден законодательно. В нормативных документах упоминание о КИ можно найти, анализируя списки опасных процессов/видов работ (см., например, [2; 3]).

Особенности автоматизации

Анализ позволяет выявить, что ОПО, ПОО, КВО — опасные объекты в порядке возрастания их важности для экономики региона, страны. Отметим, что при решении вопросов идентификации бóльшую роль играют качественные характеристики, а при категорировании — количественные. Выше отмечалось, что по мере роста важности объекта для его идентификации также необходимо больше количественных показателей (ущерб).

Решение вопросов идентификации и категорирования требует всестороннего исследования вариантов функционирования и законодательных особенностей. Некоторые объекты могут быть нормативно отнесены к нескольким видам одновременно, что может привести к избыточным расходам.

Каждый из опасных объектов функционирует в рамках одной или нескольких КИ, потому необходима оценка взаимного влияния объектов и КИ. В [15] содержатся положения, которые могут быть положены в основу автоматизации процессов идентификации и категорирования. Полезными в приложении 1 в [3] представляются типовые наименования опасных объектов, охватывающие широкий круг промышленных организаций и привязанные к 19 опасным видам деятельности.

При автоматизации необходимо сохранить сложившиеся многолетние традиции и наработки по классификации и учету опасностей и нарастить их новыми возможностями. В связи с этим начнем с формирования перечня (точнее, дерева) опасных процессов (и подпроцессов), опирающегося на нормы управления безопасностью в РФ (см. выше 19 наименований). Основные разделы перечня (приложение 1 к [3]) представлены на рис. 7.

Рис. 7. Типовые опасные процессы и объекты

Сопоставление характеристик какого-либо объекта с этим перечнем позволит выявить наличие или отсутствие соприкосновений и уточнить место объекта в общей картине: либо в явном виде найти зарегистрированный объект, либо наметить его расположение, либо констатировать отсутствие направления в дереве опасных процессов.

Провести такую процедуру возможно, если проверяемый объект обладает рядом характеристик: наименование, область деятельности, ведомственная и территориальная принадлежность.

К числу параметров, определяющих роль и значение (важность, опасность) объекта, могут относиться все взаимосвязи этого объекта с другими объектами и системами, в том числе не учитываемыми (или частично учитываемыми) в законодательстве по промышленно-экологической безопасности. Это могут быть надсистемы и подсистемы, взаимодействующие системы: ведомственные, территориальные, проектные, проблемные, КИ. Выявить эти (дополнительные) взаимосвязи предполагается с привлечением средств искусственного интеллекта (ИИ), экспертно-аналитических систем. Предложения по учету взаимодействия с КИ (регионального уровня) представлены на рис. 8.

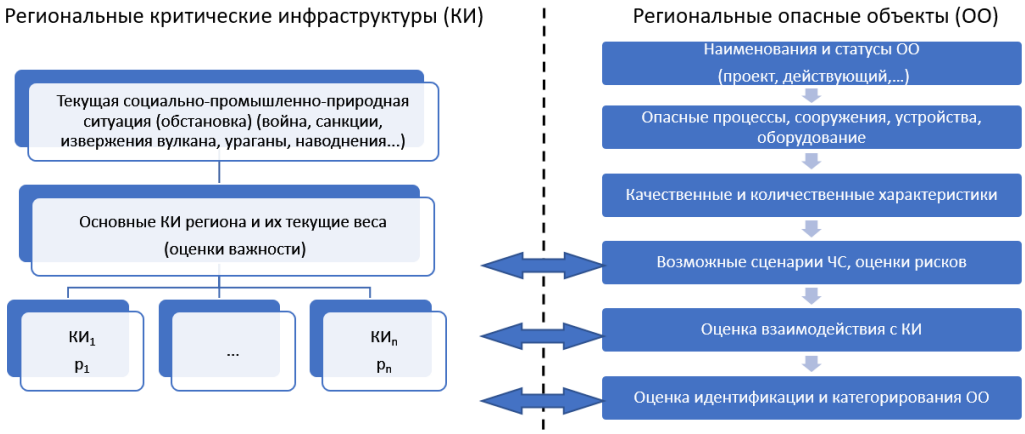


Рис. 8. Учет взаимодействия с КИ

Заключение

Рассмотрены концептуальные основы выявления и категорирования опасных объектов различных видов. Представлены обобщенные информационные схемы отнесения к конкретным видам и определения уровня значимости в рамках этих видов. Сформулированы основные проблемы указанных процедур и их желательные свойства. Рассмотрен вопрос взаимодействия опасных объектов и соответствующих критических инфраструктур. Представлены соображения по автоматизации описанных методик.

Список источников

1. Яковлев С. Ю. Идентификация опасных производственных объектов // Идентификация систем и задачи управления (SICPRO'2000) // Труды международной конференции (Москва, 26–28 сентября 2000 г., ИПУ РАН). М.: изд. ИПУ РАН, ISBN 5-201-09605-0, выпуск на CD-ROM. С. 890–897.
2. Федеральный закон от 21.07.1997 № 116-ФЗ (ред. от 14.11.2023) «О промышленной безопасности опасных производственных объектов» // Собрание законодательства РФ. 1997. № 30. Ст. 3588.
3. Требования к регистрации объектов в государственном реестре опасных производственных объектов и ведению государственного реестра опасных производственных объектов, утвержденные приказом Ростехнадзора от 30 ноября 2020 г. № 471 (зарегистрирован Минюстом России 18 декабря 2020 г. № 61590).
4. Подображин С. Н., Стульская Т. В. Особенности идентификации и регистрации опасных производственных объектов угольной промышленности // Безопасность труда в промышленности. 2020. № 4. С. 68–73. doi:10.24000/0409-2961-2020-4-68-73.
5. Федеральный закон от 21.12.1994 № 68-ФЗ (ред. от 30.01.2024) «О защите населения и территорий от чрезвычайных ситуаций природного и техногенного характера» // Собрание законодательства РФ. 1994. № 35. Ст. 3648.
6. Яковлев С. Ю., Шемякин А. С. Изменения в информационной технологии разработки паспортов безопасности потенциально опасных объектов // Экология промышленного производства. 2023. № 2. С. 48–52. doi:10.52190/2073-2589_2023_2_48.

7. Постановление Правительства РФ от 14.08.2020 № 1226 «Об утверждении Правил разработки критериев отнесения объектов всех форм собственности к потенциально опасным объектам» // Собрание законодательства РФ. 24.08.2020. № 34. Ст. 5456.
8. Градостроительный кодекс Российской Федерации от 29.12.2004 № 190-ФЗ (ред. от 19.12.2022) // Собрание законодательства РФ. 03.01.2005. № 1 (часть 1). Ст. 16.
9. О классификации чрезвычайных ситуаций природного и техногенного характера: Постановление Правительства РФ от 21.05.2007 № 304 // ГАРАНТ: сайт. URL: https://base.garant.ru/12153609/ (дата обращения: 23.01.2023).
10. Приказ Роспотребнадзора от 26.05.2022 № 278 «Об утверждении критериев отнесения объектов всех форм собственности, правообладателями которых являются Федеральная служба по надзору в сфере защиты прав потребителей и благополучия человека или организации, в отношении которых Федеральная служба по надзору в сфере защиты прав потребителей и благополучия человека осуществляет координацию и регулирование деятельности, к потенциально опасным объектам» (зарегистрировано в Минюсте России 10.08.2022 № 69585).
11. Кубасов И. А. Управление безопасностью потенциально опасных объектов // Надежность и качество сложных систем. 2020. № 4 (32). С. 42–49. doi 10.21685/2307-4205-2020-4-5.
12. Яковлев С. Ю., Шемякин А. С. Новый планирующий документ — паспорт безопасности критически важного объекта // Проблемы безопасности и чрезвычайных ситуаций. 2024. № 5. С. 6–10. doi:10.36535/0869-4176-2024-05-1.
13. Постановление Правительства РФ от 14.08.2020 № 1225 «Об утверждении Правил разработки критериев отнесения объектов всех форм собственности к критически важным объектам» // Собрание законодательства РФ. 24.08.2020. № 34. Ст. 5455.
14. Яковлев С. Ю., Шемякин А. С., Олейник А. Г. Регулирование техногенно-экологической безопасности критически важных объектов инфраструктуры: обновление нормативной базы // Труды Кольского научного центра РАН. Серия: Технические науки. 2022. Т. 13, № 2. С. 93–102. doi:10.37614/2949-1215.2022.13.2.008.
15. Об утверждении критериев отнесения объектов всех форм собственности, правообладателями которых являются Федеральная служба по надзору в сфере защиты прав потребителей и благополучия человека или организации, в отношении которых Федеральная служба по надзору в сфере защиты прав потребителей и благополучия человека осуществляет координацию и регулирование деятельности, к критически важным объектам: Приказ Роспотребнадзора от 26.05.2022 № 277 // Официальный интернет-портал правовой информации. URL: http://publication.pravo.gov.ru/Document/View/0001202208100012 (дата обращения: 23.01.2023).
16. Об утверждении критериев отнесения объектов всех форм собственности, правообладателями которых являются Федеральное агентство водных ресурсов или организации, в отношении которых Федеральное агентство водных ресурсов осуществляет координацию и регулирование деятельности, к критически важным объектам: Приказ Федерального агентства водных ресурсов от 22.11.2021 № 301 // Официальный интернет-портал правовой информации. URL: http://publication.pravo.gov.ru/Document/View/0001202202240021 (дата обращения: 23.01.2023).
17. Анюгина М. И. Некоторые подходы к классификации и разработке паспортов безопасности критически важных объектов // Технологии гражданской безопасности. 2020. Т. 17, № 2 (64). С. 47–51. doi:10.54234/CST.19968493.2020.17.2.64.8.47.
18. Яковлев С. Ю., Шемякин А. С. Новый планирующий документ — паспорт безопасности критически важного объекта // Проблемы безопасности и чрезвычайных ситуаций. 2024. № 5. С. 6–10. doi:10.36535/0869-4176-2024-05-1. EDN LBZGIY.
19. Яковлев С. Ю., Шемякин А. С., Олейник А. Г. Методические основы информационной технологии категорирования значимости и оценки безопасности объектов инфраструктуры арктических регионов // Предупреждение и ликвидация чрезвычайных ситуаций в Арктической зоне Российской Федерации: материалы научно-практической конференции, 4–7 апреля 2023 года / [ответственный редактор Е. А. Румянцева]. Мурманск: МАГУ, 2023. С. 355–357.
20. Яковлев С. Ю., Олейник А. Г., Шемякин А. С. Возможные подходы к оценке безопасности промышленно-природных объектов и критических инфраструктур в Арктической зоне РФ // III Юдахинские чтения : сборник научных материалов, Архангельск, 25–28 июня 2024 года. Архангельск: КИРА, 2024. С. 88–92. EDN LAXINX.
21. Яковлев С. Ю., Шемякин А. С. Обеспечение промышленно-экологической безопасности организаций при планировании действий в чрезвычайных ситуациях // Экология промышленного производства. 2025. № 4. (В печати).

22. Цыгичко В. Н., Черешкин Д. С., Смолян Г. Л. Безопасность критических инфраструктур. Изд. 2-е, стереотип. М.: ЛЕНАНД, 2021. 200 с.

References

1. Yakovlev S. Yu Identifikacija opasnyh proizvodstvennyh ob"ektov [Identification of hazardous production facilities] Identifikacija sistem i zadachi upravlenija (SICPRO'2000). *Trudy mezhdunarodnoj konferencii (Moskva, 26–28 sentjabrja 2000 g., IPU RAN)*. Moscow, izd. IPU RAN, ISBN 5-201-09605-0, vypusk na CD-ROM, pp. 890–897. (In Russ.).
2. Federal'nyj zakon ot 21.07.1997 No. 116-FZ (red. ot 14.11.2023) “O promyshlennoj bezopasnosti opasnyh proizvodstvennyh ob"ektov” [Federal Law No. 116-FZ dated 07/21/1997 (as amended on 11/14/2023) “On Industrial Safety of Hazardous Production Facilities”]. *Sobranie zakonodatel'stva RF* [Collection of Legislation of the Russian Federation], 1997, no. 30, article. 3588. (In Russ.).
3. Trebovanija k registracii ob"ektov v gosudarstvennom reestre opasnyh proizvodstvennyh ob"ektov i vedeniju gosudarstvennogo rejestra opasnyh proizvodstvennyh ob"ektov, utverzhdennye prikazom Rostehnadzora ot 30 nojabria 2020 g. No. 471 (zaregistrirovan Minjustom Rossii 18 dekabrja 2020 g. No. 61590) [Requirements for registration of facilities in the state Register of Hazardous production facilities and maintenance of the state register of hazardous production facilities, approved by Rostechnadzor Order No. 471 dated November 30, 2020 (registered by the Ministry of Justice of Russia on December 18, 2020 No. 61590)]. (In Russ.).
4. Podobrazhin S. N., Stulskaya T. V. Osobennosti identifikatsii i registratsii opasnikh proizvodstvennikh obektov ugolnoi promishlennosti [Features of identification and registration of hazardous production facilities in the coal industry]. *Bezopasnost truda v promishlennosti* [Occupational safety in industry], 2020, no. 4, pp. 68–73. doi:10.24000/0409-2961-2020-4-68-73.
5. Federal'nyj zakon ot 21.12.1994 No. 68-FZ (red. ot 30.01.2024) “O zashchite naselenija i territorij ot chrezvychajnyh situacij prirodnogo i tekhnogennogo haraktera” [Federal Law No. 68-FZ dated 12/21/1994 (as amended on 01/30/2024) “On the protection of the population and environment from natural and man-made emergencies”]. *Sobranie zakonodatel'stva RF* [Collection of Legislation of the Russian Federation], 1994, no. 35, article 3648. (In Russ.).
6. Yakovlev S. Yu, Shemyakin A. S. Izmenenija v informacionnoj tehnologii razrabotki pasportov bezopasnosti potencial'no opasnyh ob"ektov [Changes in information technology for developing safety passports for potentially dangerous objects]. *Jekologija promyshlennogo proizvodstva* [Ecology of industrial production], 2023, no. 2, pp. 48–52. (In Russ.). doi: 10.52190/2073-2589_2023_2_48.
7. Postanovlenie Pravitelstva RF ot 14.08.2020 No. 1226 “Ob utverzhdenii Pravil razrabotki kriteriev otneseniya obektov vsekh form sobstvennosti k potentsialno opasnim obektam” [Decree of the Government of the Russian Federation dated 08/14/2020 No. 1226 “On Approval of the Rules for Developing Criteria for Classifying Objects of all Forms of ownership as potentially dangerous objects”]. *Sobranie zakonodatelstva RF* [Collection of Legislation of the Russian Federation], 24.08.2020, no. 34, article 5456. (In Russ.).
8. Gradostroitelnii kodeks Rossiiskoi Federatsii ot 29.12.2004 No. 190-FZ (red. ot 19.12.2022) [Urban Planning Code of the Russian Federation No. 190-FZ dated December 29, 2004 (as amended on December 19, 2022)]. *Sobranie zakonodatelstva RF* [Collection of legislation of the Russian Federation], 03.01.2005, No 1 (chast 1), article 16. (In Russ.).
9. O klassifikacii chrezvychajnyh situacij prirodnogo i tekhnogennogo haraktera: Postanovlenie Pravitel'stva RF ot 21.05.2007 No. 304 [On the classification of natural and man-made emergencies: Decree of the Government of the Russian Federation dated 05/21/2007 No. 304]. (In Russ.). Available at: https://base.garant.ru/12153609 (accessed 23.01.2023).
10. Prikaz Rospotrebnadzora ot 26.05.2022 No. 278 “Ob utverzhdenii kriteriev otnesenija ob"ektov vseh form sobstvennosti, pravoobladateljami kotoryh javljajutsja Federal'naja sluzhba po nadzoru v sfere zashity prav potrebitelej i blagopoluchija cheloveka ili organizacii, v otnoshenii kotoryh Federal'naja sluzhba po nadzoru v sfere zashity prav potrebitelej i blagopoluchija cheloveka osushhestvljaet koordinaciju i regulirovanie dejatel'nosti, k potencial'no opasnym ob"ektam”(zaregistrirovano v Minjuste Rossii 10.08.2022 No. 69585) [Rospotrebnadzor Order No. 278 dated 05/26/2022 "On Approval of Criteria for Classifying Objects of All Forms of Ownership Owned by the Federal Service for Supervision of Consumer Rights Protection and Human Welfare or Organizations in Respect of which the Federal Service for Supervision of Consumer Rights Protection and Human Welfare Coordinates and Regulates Activities as Potentially Dangerous Objects" (registered with the Ministry of Justice of Russia on 08/10/2022 No. 69585)]. (In Russ.).

11. Kubasov I. A. Upravlenie bezopasnostyu potentsialno opasnikh obektov [Safety management of potentially dangerous facilities]. *Nadezhnost i kachestvo slozhnikh sistem* [Reliability and quality of complex systems], 2020, no. 4 (32), pp. 42–49. (In Russ.). doi 10.21685/2307-4205-2020-4-5.
12. Yakovlev S. Yu., Shemyakin A. S. Nowyj planirujushchij dokument — pasport bezopasnosti kriticheski vaznogo ob"ekta [New planning document—passport of safety of a critically important object]. *Problemy bezopasnosti i chrezvychajnyh situacij* [Security and emergency situations issues], 2024, no. 5, pp. 6–10. (In Russ.). doi 10.36535/0869-4176_2024_5_1.
13. Postanovlenie Pravitelstva RF ot 14.08.2020 No. 1225 “Ob utverzhdenii Pravil razrabotki kriteriev otneseniya obektov vsekh form sobstvennosti k kriticheski vazhnim obektam” [Decree of the Government of the Russian Federation dated 08/14/2020 No. 1225 “On Approval of the Rules for Developing Criteria for Classifying Objects of all Forms of Ownership as Critically important objects”]. *Sobranie zakonodatelstva RF* [Collection of Legislation of the Russian Federation], 24.08.2020, No 34, article 5455.
14. Yakovlev S. Yu., Shemyakin A. S., Oleynik A. G. Regulirovanie tekhnogenno-jekologicheskoj bezopasnosti kriticheski vaznyh ob"ektov infrastruktury: obnovlenie normativnoj bazy [Regulation of technogenic-ecological safety of critically important infrastructure facilities: updating regulatory framework]. *Trudy Kol'skogo nauchnogo centra RAN. Serija: Tehnicheskie nauki* [Proceedings of the KSC RAS. Information Technology], 2022, vol. 13, no. 2, pp. 93–102. (In Russ.). doi:10.37614/2949-1215.2022.13.2.008.
15. Ob utverzhdenii kriteriev otnosenija ob"ektov vseh form sobstvennosti, pravoobladateljami kotoryh javljajutsja Federal'naja sluzhba po nadzoru v sfere zashity prav potrebitelej i blagopoluchija cheloveka ili organizacii, v otnoshenii kotoryh Federal'naja sluzhba po nadzoru v sfere zashity prav potrebitelej i blagopoluchija cheloveka osushhestvljaet koordinaciju i regulirovanie dejatel'nosti, k kriticheski vaznym ob"ektam: Prikaz Rospotrebnadzora ot 26.05.2022 No. 277 [On approval of criteria for classifying objects of all forms of ownership owned by the Federal Service for Supervision of Consumer Rights Protection and Human Well-being or Organizations for which the Federal Service for Supervision of Consumer Rights Protection and Human Well-being coordinates and regulates activities as critically Important objects: Rospotrebnadzor Order No. 277 dated 05.26.2022]. (In Russ.). Available at: http://publication.pravo.gov.ru/Document/View/00012022 08100012 (accessed 23.01.2023).
16. Ob utverzhdenii kriteriev otnosenija ob"ektov vseh form sobstvennosti, pravoobladateljami kotoryh javljajutsja Federal'noe agentstvo vodnyh resursov ili organizacii, v otnoshenii kotoryh Federal'noe agentstvo vodnyh resursov osushhestvljaet koordinaciju i regulirovanie dejatel'nosti, k kriticheski vaznym ob"ektam: Prikaz Federal'nogo agentstva vodnyh resursov ot 22.11.2021 No. 301 [On approval of criteria for classifying objects of all forms of ownership owned by the Federal Agency for Water Resources or organizations for which the Federal Agency for Water Resources coordinates and regulates activities as critically Important objects: Order No. 301 of the Federal Agency for Water Resources dated 11/22/2021]. (In Russ.). Available at: http://publication.pravo.gov.ru/Document/View/0001202202240021 (accessed 23.01.2023).
17. Anyugina M. I. Nekotorie podkhodi k klassifikatsii i razrabotke pasportov bezopasnosti kriticheski vazhnikh obektov [Some approaches to the classification and development of safety data sheets critical facilities]. *Tekhnologii grazhdanskoi bezopasnosti* [Civil security technologies], 2020, vol. 17, no. 2 (64), pp. 47–51. (In Russ.). doi:10.54234/CST.19968493.2020.17.2.64.8.47.
18. Yakovlev S. Yu., Shemyakin A. S. Novii planiruyushchii dokument — pasport bezopasnosti kriticheski vazhnogo obekta [New planning document—Safety data sheet for a critically important facility]. *Problemi bezopasnosti i chrezvichainikh situatsii* [Safety and Emergency Situations], 2024, no. 5, pp. 6–10. (In Russ.). doi:10.36535/0869-4176-2024-05-1. EDN LBZGIY.
19. Yakovlev S. Yu., Shemyakin A. S., Oleinik A. G. Metodicheskie osnovi informatsionnoi tekhnologii kategorirovaniya znachimosti i otsenki bezopasnosti obektov infrastrukturi arkticheskikh regionov [Methodological foundations for information technology for categorizing the importance and safety assessment of infrastructure facilities in the Arctic]. *Preduprezhdenie i likvidatsiya chrezvichainikh situatsii v Arkticheskoi zone Rossiiskoi Federatsii: materiali nauchno-prakticheskoi konferentsii, 4–7 aprelya 2023 goda* regions [Prevention and elimination of emergency situations in the Arctic zone of the Russian Federation: proceedings of the scientific and practical conference, April 4–7, 2023]. Murmansk, MAGU, 2023, pp. 355–357. (In Russ.).
20. Yakovlev S. Yu., Oleinik A. G., Shemyakin A. S. Vozmozhnie podkhodi k otsenke bezopasnosti promishlenno-prirodnikh obektov i kriticheskikh infrastruktur v Arkticheskoi zone RF [Possible approaches to assessing the safety of industrial and natural facilities and critical infrastructures in the Arctic zone of the Russian Federation].

III Yudakhinskie chteniya : sbornik nauchnikh materialov, Arkhangelsk, 25–28 iyunya 2024 goda [III Yudakhinsky readings: proceedings, Arkhangelsk, June 25–28, 2024]. Arkhangelsk, KIRA, 2024, pp. 88–92. (In Russ.). EDN LAXINX.

21. Yakovlev S. Yu., Shemyakin A. S. Obespechenie promishlenno-ekologicheskoi bezopasnosti organizatsii pri planirovanii deistvii v chrezvichainikh situatsiyakh [Ensuring industrial and environmental safety of organizations when planning actions in emergency situations]. *Ekologiya promishlennogo proizvodstva* [Ecology of industrial production], 2025, no. 4. (Publishing).
22. Tsigichko V. N., Chereshkin D. S., Smolyan G. L. *Bezopasnost kriticheskikh infrastruktur* [Security of critical infrastructures]. Moscow, LENAND, 2021, 200 p. (In Russ.).

Информация об авторах

С. Ю. Яковлев — кандидат технических наук, старший научный сотрудник;
А. С. Шемякин — младший научный сотрудник.

Information about the authors

S. Yu. Yakovlev — Candidate of Science (Tech.), Senior Researcher;
A. S. Shemyakin — Junior Researcher.

Статья поступила в редакцию 10.11.2025; одобрена после рецензирования 18.11.2025; принята к публикации 20.11.2025.
The article was submitted 10.11.2025; approved after reviewing 18.11.2025; accepted for publication 20.11.2025.


РИО
КНЦ