

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
МЕЖДУНАРОДНАЯ АКАДЕМИЯ ИНФОРМАТИЗАЦИИ
СОЮЗ МАШИНОСТРОИТЕЛЕЙ РОССИЙСКОЙ ФЕДЕРАЦИИ
МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И СВЯЗИ РЯЗАНСКОЙ ОБЛАСТИ
РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
ИМЕНИ В.Ф. УТКИНА

СОВРЕМЕННЫЕ ТЕХНОЛОГИИ В НАУКЕ И ОБРАЗОВАНИИ

СТНО-2026

**IX МЕЖДУНАРОДНЫЙ НАУЧНО-
ТЕХНИЧЕСКИЙ ФОРУМ**

Сборник трудов

Том 4

Рязань
2026

УДК 004 + 001.1 + 681.2+ 681.2+ 681.3+681.5
С 568

Современные технологии в науке и образовании – СТНО-2026 [текст]: сб. тр. IX междунар. науч.-техн. форума: в 11 т. Т.4./ под общ. ред. О.В. Миловзорова. – Рязань: Рязан. гос. радиотехн. ун-т, 2026.

ISBN 978-5-7722-0443-6

Т.4: –  с.,: ил.

ISBN 978-5-7722-0447-4

Сборник включает труды участников IX Международного научно-технического форума «Современные технологии в науке и образовании» СТНО-2026.

В сборнике освещаются вопросы математического моделирования, новых технологий в радиотехнике, телекоммуникациях, электротехнике и радиоэлектронике, вопросы полупроводниковой наноэлектроники, приборостроения, лазерной, микроволновой техники, силовой промышленной электроники, новые технологии в измерительной технике и системах, биомедицинских системах, алгоритмическое и программное обеспечение вычислительной техники, вычислительных сетей и комплексов, вопросы разработки и применения элементов искусственного интеллекта в информационных системах и образовании, различные аспекты автоматизированного проектирования, обработки изображений и управления в технических системах, перспективные технологии в машиностроительном и нефтехимическом производствах, новые технологии и методики в высшем образовании, в т.ч. вопросы гуманитарной и физико-математической подготовки студентов, обучения их иностранным языкам, перспективные технологии электронного обучения, в том числе, дистанционного, вопросы экономики, управления предприятиями и персоналом, менеджмента, а также вопросы гуманитарной сферы.

Авторская позиция и стилистические особенности сохранены.

УДК 004 + 001.1 + 681.2+ 681.2+ 681.3+681.5

ISBN 978-5-7722-0443-6

ISBN 978-5-7722-0447-4

© Рязанский государственный
радиотехнический университет, 2026

ИНФОРМАЦИЯ О IX МЕЖДУНАРОДНОМ ФОРУМЕ «СОВРЕМЕННЫЕ ТЕХНОЛОГИИ В НАУКЕ И ОБРАЗОВАНИИ» СТНО-2026

IX Международный научно-технический форум «Современные технологии в науке и образовании» СТНО-2026 состоялся 03.03.2026-05.03.2026 в г. Рязань в Рязанском государственном радиотехническом университете имени В.Ф. Уткина.

В рамках форума «Современные технологии в науке и образовании» СТНО-2026 состоялась работа четырех Международных научно-технических конференций:

«Современные технологии в науке и образовании. Радиотехника и электроника», секции

- Радиотехнические системы;
- Радиотехнические устройства;
- Телекоммуникационные системы и устройства;
- Цифровые информационные технологии реального времени;
- Промышленная силовая электроника, электроэнергетика и электроснабжение;
- Физика полупроводников, микро- и нанoeлектроника;
- Микроволновая, оптическая и квантовая электроника;
- Актуальные задачи химических технологий;

«Современные технологии в науке и образовании. Вычислительная техника, математическое моделирование и автоматизированные системы», секции

- Алгоритмическое и программное обеспечение вычислительных систем и сетей;
- Искусственный интеллект и информационные технологии;
- ЭВМ и системы;
- Модели искусственного интеллекта в САПР;
- Информационные технологии в конструировании электронных средств;
- Интеллектуальные системы и технологии в образовании;
- Информационные системы и защита информации;
- Математические методы и модели в научных исследованиях и прикладных проблемах. К 90-летию И.П. Карасева - основателя теории управления "в малом";
- Обработка данных, изображений и управление в технических системах;
- Геоинформационные и космические технологии;
- Автоматизация производственно-технологических процессов в приборо- и машиностроении;
- Информационно-измерительные устройства и системы в технике и медицине.
- Стандартизация и управление качеством;
- Информационные системы и технологии;

«Современные технологии в науке и образовании. Экономика и управление», секции;

- Современные технологии государственного и муниципального управления;
- Экономика, менеджмент и организация производства;
- Бухгалтерский учет, анализ и аудит;
- Экономическая безопасность;

«Современные технологии в науке и образовании. Новые технологии и методы в высшем образовании», секции

- Современные технологии электронного обучения;
- Иностранный язык в техническом вузе;
- Лингвистика и межкультурная коммуникация;
- Направления и формы гуманитаризации высшего образования и гуманитарная подготовка студентов;
- Методы преподавания и организация учебного процесса в вузе;
- Физико-математическая подготовка студентов;

- Особенности военного образования на современном этапе.

Организационный комитет Форума:

Гусев С.И., проректор по научной работе и инновациям, д.т.н., проф.;

Миловзоров О.В., зам. начальника управления организации научных исследований, к.т.н, доц. – координатор, главный редактор сборника трудов Форума;

Миронов Вал.В., профессор кафедры высшей математики, д.ф.-м.н., проф. – ответственный редактор сборника трудов Форума;

Устинова Л.С., начальник отдела информационного обеспечения – отв. за информационную поддержку;

члены оргкомитета:

Аронов Л.В., доцент кафедры радиоуправления и связи, к.т.н., доц.;

Бабаян П.В., зав. кафедрой автоматики и информационных технологий в управлении, к.т.н., доц.;

Бубнов С.А., доцент кафедры вычислительной и прикладной математики, к.ф.-м.н., доц.;

Витязев В.В., зав. кафедрой телекоммуникаций и основ радиотехники, д.т.н., проф.;

Волченков В.А., доцент кафедры телекоммуникаций и основ радиотехники, к.т.н.;

Горлин О.А., доцент кафедры электронных приборов, к.т.н., доц.;

Городничева Е.В., ассистент кафедры автоматизации информационных и технологических процессов;

Гостин А.М., доцент кафедры систем автоматизированного проектирования вычислительных средств, к.т.н., доц.;

Губарев А.В., доцент кафедры информационно-измерительной и биомедицинской техники, к.т.н., доц.;

Гуржин С.Г., доцент кафедры информационно-измерительной и биомедицинской техники, к.т.н., доц.;

Дмитриев В.Т., зав. кафедрой радиоуправления и связи, д.т.н., доц.;

Дмитриева Т.А., доцент кафедры вычислительной и прикладной математики, к.т.н., доц.;

Евдокимова Е.Н., зав. кафедрой экономики, менеджмента и организации производства, д.э.н., проф.;

Еремеев В.В., директор НИИ «Фотон», д.т.н., проф.;

Есенина Н.Е., зав. кафедрой иностранных языков, к.п.н., доц.;

Жулев В.И., зав. кафедрой информационно-измерительной и биомедицинской техники, д.т.н., проф.;

Каширин И.Ю., доктор технических наук, профессор, профессор кафедры вычислительной и прикладной математики, д.т.н., проф.;

Киселёва О.В., доцент кафедры экономической безопасности, анализа и учета, к.э.н., доц.;

Кислицына Т.С., старший преподаватель кафедры радиотехнических систем;

Клейносова Н.П., директор центра дистанционного обучения, к.п.н., доц.;

Клочко В.К., профессор кафедры автоматики и информационных технологий в управлении, д.т.н., проф.;

Коваленко В.В., зав. кафедрой химической технологии, к.т.н., доц.;

Корячко В.П., зав. кафедрой систем автоматизированного проектирования вычислительных средств, д.т.н., проф.;

Костров Б.В., зав. кафедрой электронных вычислительных машин, д.т.н., проф.;

Кошелев В.И., зав. кафедрой радиотехнических систем, д.т.н., проф.;

Крошила С.В., доцент кафедры вычислительной и прикладной математики, к.т.н., доц.;

Круглов С.А., зав. кафедрой промышленной электроники, д.т.н., доц.;

Кузьмин Ю.М., доцент кафедры информационной безопасности, к.т.н., доц.;
Куприна О.Г., доцент кафедры иностранных языков, к.филол.н., доц.;
Ленков М.В., зав. кафедрой автоматизации информационных и технологических процессов, к.т.н., доц.;
Литвинов В.Г., зав. кафедрой микро- и нанoeлектроники, д.ф.-м.н., доц.;
Лукьянова Г.С., зав. кафедрой высшей математики, к.ф.-м.н., доц.;
Маметова Ю.Ф., доцент кафедры иностранных языков, к.п.н., доц.;
Миронов В.В., старший преподаватель кафедры Воздушно-космических сил;
Митрошин А.А., доцент кафедры систем автоматизированного проектирования вычислительных средств, к.т.н., доц.;
Мишустин В.Г., доцент кафедры микро- и нанoeлектроники, к.ф.-м.н., доц.;
Овечкин Г.В., зав. кафедрой вычислительной и прикладной математики, д.т.н., проф.;
Паршин А.Ю., доцент кафедры радиотехнических устройств, к.т.н., доц.;
Паршин Ю.Н., зав. кафедрой радиотехнических устройств, д.т.н., проф.;
Перфильев С.В., зав. кафедрой государственного, муниципального и корпоративного управления, д.э.н., проф.;
Подгорнова Н.А., доцент кафедры государственного, муниципального и корпоративного управления, к.э.н., доц.;
Пржегорлинский В.Н., зав. кафедрой информационной безопасности, к.т.н., доц.;
Пылькин А.Н., профессор кафедры вычислительной и прикладной математики, д.т.н., проф.;
Саблина В.А., доцент кафедры электронных вычислительных машин, к.т.н., доц.;
Сапрыкин А.Н., доцент кафедры систем автоматизированного проектирования вычислительных средств, к.т.н., доц.;
Семенов А.Р., доцент кафедры химической технологии, к.ф.-м.н.;
Скрипкина О.В. доцент кафедры экономической безопасности, анализа и учета, к.э.н., доц.;
Соколов А.С., зав. кафедрой истории, философии и права, д.и.н;
Соловьева И.П., доцент кафедры экономики, менеджмента и организации производства, к.э.н., доц.;
Таганов А.И., профессор кафедры космических технологий, д.т.н., проф.;
Тарасова В.Ю., ассистент кафедры электронных вычислительных машин, магистр;
Торженева Т.В. доцент кафедры экономической безопасности, анализа и учета, к.э.н., доц.;
Харитонов А.Ю., нач. военного учебного центра, полковник, к.т.н., доц;
Холопов С.И., зав. кафедрой автоматизированных систем управления, к.т.н., доц.;
Цыцына М.И., ассистент кафедры космических технологий, магистр;
Чеглакова С.Г., зав. кафедрой экономической безопасности, анализа и учета, д.э.н., проф.;
Челебаев С.В., доцент кафедры автоматизированных систем управления, к.т.н., доц.;
Щевьев А.А., доцент кафедры истории, философии и права, к.п.н., доц.

МЕЖДУНАРОДНАЯ НАУЧНО-ТЕХНИЧЕСКАЯ КОНФЕРЕНЦИЯ «СОВРЕМЕННЫЕ ТЕХНОЛОГИИ В НАУКЕ И ОБРАЗОВАНИИ. ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА, МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ И АВТОМАТИЗИРОВАННЫЕ СИСТЕМЫ»

СЕКЦИЯ «АЛГОРИТМИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ И СЕТЕЙ »

УДК 004; ГРНТИ 20

ОБЗОР МЕТОДОВ ОБНАРУЖЕНИЯ КЛОНОВ КОДА НА ОСНОВЕ ИНВЕРТИРОВАННОГО ИНДЕКСА

Х. Шафает*, А.В. Пруцков***

**Липецкий государственный педагогический университет имени П.П. Семенова-Тян-Шанского,
Российская Федерация, Липецк,*

***Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, mail@prutzkow.com*

Аннотация. Стремительное расширение крупномасштабных репозиторий программного обеспечения привело к значительному увеличению повторного использования существующего исходного кода, что обусловило повсеместное распространение клонов кода. Эффективное обнаружение таких клонов имеет важное значение для сопровождения программного обеспечения, обеспечения качества и защиты интеллектуальной собственности. Традиционные методы обнаружения клонов, основанные на синтаксическом или структурном анализе, часто сталкиваются с проблемами масштабируемости при применении к огромным кодобазам. В ответ на эти вызовы подходы на основе инвертированного индекса, изначально разработанные для информационно-поисковых систем, были адаптированы для обнаружения клонов кода. Статья представляет обзор методов обнаружения клонов кода, использующих инвертированный индекс. Рассматриваются теоретические основы, стратегии индексирования, репрезентативные инструменты, метрики оценки и существующие ограничения. Описываются перспективные направления исследований, включая семантически-осознанное индексирование, адаптивную оптимизацию и безопасные механизмы поиска кода.

Ключевые слова: обнаружение клонов кода; инвертированный индекс; схожесть исходного кода; информационный поиск; программная инженерия.

A REVIEW OF CODE CLONE DETECTION METHODS BASED ON AN INVERTED INDEX

H. Shafayet*, A. Prutzkow,****

**Lipetsk State Pedagogical University, Russia, Lipetsk*

***Ryazan State Radio Engineering University, Russia, Ryazan*

Abstract. The rapid expansion of large-scale software repositories has led to a significant increase in the reuse of existing source code, leading to the ubiquity of code clones. Effective detection of such clones is essential for software maintenance, quality assurance, and intellectual property protection. Traditional clone detection methods based on syntactic or structural analysis often face scalability issues when applied to large codebases. In response to these challenges, inverted index-based approaches, originally developed for information retrieval systems, have been adapted for code clone detection. The paper reviews code clone detection methods using an inverted index. Theoretical foundations, indexing strategies, representative tools, evaluation metrics, and existing limitations are discussed. Promising research directions, including semantically aware indexing, adaptive optimization, and secure code search mechanisms, are described.

Keywords: Code clone detection; inverted index; source code similarity; information retrieval; software engineering.

Введение

Повторное использование программного обеспечения (ПО) стало основной практикой в ее современной разработке, позволяя ускорить и снизить затраты на разработку. Однако масштабное повторное использование часто приводит к дублированию или почти дублированию фрагментов кода, обычно называемых клонами кода. Хотя некоторый уровень дублирования допустим, неконтролируемое клонирование кода увеличивает усилия по сопровождению, усложняет отладку и может привести к появлению уязвимостей безопасности. Автоматизированное обнаружение клонов кода привлекает значительное внимание исследователей в течение последних двух десятилетий [1].

Ранние подходы к обнаружению клонов опирались на прямое сравнение текста или анализ на основе токенов. Хотя они эффективны для небольших проектов, эти методы плохо масштабируются до репозиторий, содержащих миллионы строк кода. Структурные подходы, основанные на абстрактных синтаксических деревьях и графах зависимостей программ, обеспечивают более высокую точность, но требуют дорогостоящих этапов предварительной обработки и сравнения [2]. Для решения проблем масштабируемости исследователи все чаще применяют методы информационного поиска, в частности инвертированные индексы, для эффективного поиска похожих фрагментов кода в больших наборах данных [3].

Связанные работы

Исследования по обнаружению клонов кода можно в общих чертах классифицировать на текстовые, токен-ориентированные, основанные на деревьях и семантические подходы. Текстовые методы выполняют прямое сравнение строк, в то время как токен-ориентированные техники нормализуют исходный код в последовательности токенов для уменьшения синтаксической вариативности. Методы, основанные на деревьях, представляют программы в виде абстрактных синтаксических деревьев и обнаруживают структурное сходство путем сопоставления деревьев или векторизации [4].

Методы, основанные на информационном поиске, представляют фрагменты исходного кода в виде документов и применяют модели индексирования и ранжирования схожести, изначально разработанные для обработки естественного языка [5]. Среди них подходы на основе инвертированного индекса продемонстрировали превосходную масштабируемость и приемлемую точность, что делает их подходящими для обнаружения клонов в крупномасштабных системах [6].

Теоретические основы

Инвертированный индекс — это структура данных, которая отображает термины в документы, в которых они встречаются. При анализе ПО токены, извлеченные из исходного кода, служат терминами, а функции, файлы или классы представляют документы. Эта структура обеспечивает эффективное извлечение кандидатов — фрагментов кода, имеющих общие токены.

Клоны кода обычно делятся на четыре типа: Клоны Типа I идентичны за исключением форматирования; Клоны Типа II отличаются идентификаторами или литералами; Клоны Типа III включают измененные или переупорядоченные операторы; и Клоны Типа IV функционально схожи, но синтаксически различны. Методы на основе инвертированного индекса особенно эффективны для обнаружения клонов Типа I–III.

Методология обнаружения клонов на основе инвертированного индекса

Процесс обнаружения клонов (рис. 1) начинается с парсинга и токенизации исходного кода с использованием языково-специфичных лексеров. Несущественные элементы, та-

кие как комментарии и пробелы, удаляются, и применяются техники нормализации для уменьшения поверхностных различий. Полученные токены вставляются в структуру инвертированного индекса, обычно реализованную с использованием хэш-таблиц или фреймворков поисковых систем [7].

Во время обработки запроса целевой фрагмент кода токенизируется и сопоставляется с индексом для извлечения кандидатов. Для ранжирования кандидатов используются меры схожести, такие как TF-IDF, косинусное сходство или коэффициент Жаккара. Часто применяется этап верификации для устранения ложных срабатываний [8].

ОБРАТНЫЙ ИНДЕКС



Рис. 1. Архитектура инвертированного индекса

Репрезентативные системы

Несколько систем успешно применяли методы на основе инвертированного индекса для обнаружения клонов кода. SourcererCC – один из наиболее заметных инструментов, способный масштабироваться до миллионов строк кода, используя токен-ориентированное индексирование и эффективные эвристики фильтрации. Другие системы, такие как CCAAligner, комбинируют инвертированные индексы с locality-sensitive хэшированием для дальнейшего повышения производительности.

Метрики оценки

Оценка производительности инструментов обнаружения клонов обычно основывается на точности, полноте, F1-мере, а также времени индексирования и выполнения запросов. Эталонные наборы данных, такие как BigCloneBench, обычно используются для обеспечения воспроизводимости и честного сравнения.

Ниже приведено сравнение репрезентативных инструментов обнаружения клонов (см. таблицу).

Таблица. Сравнение репрезентативных инструментов обнаружения клонов

Инструмент	Техника	Типы клонов	Масштабируемость	Точность
SourcererCC	Инвертированный индекс	I–III	Очень высокий	Высокий
CCFinder	На основе токенов	I–II	Средний	Высокий
Deckard	Векторы деревьев	I–III	Средний	Высокий
NiCad	Текст + Парсинг	I–III	Средний	Очень высокий

Проблемы и будущие направления исследований

Несмотря на свои преимущества в масштабируемости, подходы на основе инвертированного индекса сталкиваются с несколькими проблемами. Большинство методов сосредоточены на синтаксическом сходстве и плохо справляются с семантическими клонами. Интеграция семантических векторных представлений (эмбеддингов), полученных из моделей глубокого обучения, представляет собой перспективное направление исследований. Кроме того, адаптивная оптимизация индекса и механизмы поиска с сохранением конфиденциальности остаются открытыми исследовательскими проблемами [9].

Заключение

В данном обзоре рассмотрены методы обнаружения клонов кода на основе инвертированного индекса, освещены их теоретические основы, практические реализации и методологии оценки. Подходы на основе инвертированного индекса обеспечивают эффективный баланс между масштабируемостью и точностью, делая их подходящими для крупномасштабных репозиториях ПО. Будущая работа должна быть сосредоточена на гибридных моделях, сочетающих эффективность индексирования с семантическим пониманием исходного кода [10].

Библиографический список

1. Sajnani H. et al. (2016) SourcererCC: Scaling code clone detection to big code // ICSE.
2. Roy C. K., Cordy J. R. (2008) NiCad: Accurate detection of near-miss intentional clones using flexible pretty-printing and code normalization // ICPC.
3. Jiang L. et al. (2007) Deckard: Scalable and accurate tree-based code clone detection // ICSE.
4. Hammad M., Khomh F. (2020) An empirical study of code clone detection tools and their evaluation metrics // Journal of Systems and Software, (162).
5. Bellon S. et al. (2007) Comparison and evaluation of clone detection tools // IEEE Transactions on Software Engineering, 33(9).
6. Keivanloo I. et al. (2013) Benchmarking clone detection tools with BigCloneEval // ICPC.
7. Kamiya T. et al. (2002) CCFinder: A multilingual token-based code clone detection system // IEEE TSE.
8. Wang S. et al. (2018) Detecting semantically similar code with Tree-Based Convolutional Neural Networks // ICSE.
9. Liu, C., et al. (2019) Learning-based code clone detection: A review // Information and Software Technology, (106).
10. Tóth A. et al. (2020) Towards machine learning-based code clone detection // Software Quality Journal, (28).

УДК 004.89; ГРНТИ 28.23.25

ВЫЧИСЛЕНИЕ СЕМАНТИЧЕСКОГО СХОДСТВА ИНТЕЛЛЕКТУАЛЬНЫМИ СИСТЕМАМИ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

И.Ю. Каширин

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, igor-kashirin@mail.ru*

Аннотация. Приводится обзор основных методов и технологий вычисления семантического сходства словарных конструкций естественного языка (текстов, предложений, словосочетаний и слов). Главное внимание уделяется рассмотрению подходов, связанных с использованием теории представления знаний на примере семантических сетей, онтологических моделей и графов знаний. Рассматриваются методы, использующие таксономические иерархии для организации тезаурусов и словарей. Кратко излагается новая концепция применения десятичных иерархических чисел для индексирования концептов (понятий) для вычисления семантического сходства словарных конструкций. Приводится мате-

матическая формула для таких вычислений, отмечается эффективность применения подхода с использованием иерархических чисел, не уступающая перечисленным в обзорной части подходам. Эффективность нового подхода заключается также в существенном понижении вычислительной сложности в сравнении с современными методами.

Ключевые слова: семантическое сходство, таксономия, иерархические числа, семантические сети, графы знаний, вычислительная сложность.

COMPUTATION OF SEMANTIC SIMILARITY INTELLIGENT SYSTEMS NATURAL LANGUAGE

I.Yu. Kashirin

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, horek.colupaeva@yandex.ru*

The summary. An overview of the main methods and technologies for calculating the semantic similarity of dictionary constructions of natural language (texts, sentences, phrases and words) is given. The main attention is paid to the consideration of approaches related to the use of knowledge representation theory using the example of semantic networks, ontological models and knowledge graphs. Methods using taxonomic hierarchies for organizing thesauri and dictionaries are considered. A new concept of using decimal hierarchical numbers for indexing concepts (concepts) to calculate the semantic similarity of dictionary constructions is briefly described. A mathematical formula for such calculations is provided, and the effectiveness of using the approach based on hierarchical numbers is noted, which is comparable to the approaches listed in the review section. The effectiveness of the new approach also lies in significantly reducing the computational complexity compared to current methods.

Keywords: semantic similarity, taxonomy, hierarchical numbers, semantic networks, knowledge graphs, and computational complexity.

Семантическое сходство естественно-языковых конструкций (текстов, предложений, словосочетаний и слов) является одним из главных функциональных модулей интеллектуальных систем общения [1,2]. Они используются для решения следующих актуальных задач прикладного уровня [3]:

- поиска релевантных пользовательскому запросу фрагментов текста в больших информационных хранилищах или глобальных сетях;
- извлечения фактографических данных из предложений;
- формирования естественно-языковых ответов на запросы клиентов информационных систем;
- классификации текстов.

Семантическое сходство - это показатель, используемый для измерения степени сходства значений между двумя концептами (понятиями). Концептами могут быть два слова, предложения или абзацы. Интеллектуальная система пытается найти расстояние между двумя понятиями в семантическом пространстве, чем меньше расстояние, тем больше сходство. Существующие методы вычисляют сходство между двумя понятиями, которые лексикографически различаются. Методы, используемые для поиска семантического сходства между двумя словами, могут быть расширены для поиска сходства между предложениями, фразами или абзацами.

Подходы, используемые для расчета семантического сходства, можно классифицировать как модели, основанные на обучающем корпусе, и модели, основанные на знаниях [4]. Корпусные модели пытаются определить степень сходства между двумя терминами на основе информации, полученной при анализе большого набора документов (корпус). Модели, основанные на знаниях, пытаются определить, насколько схожи два термина, базирующиеся на какой-либо семантической сети или онтологии [5].

Одним из самых ранних предложенных методов вычисления семантического сходств был разработан Ландауэром [6]. Он основан на предположении, что слова, имеющие схожее значение, будут встречаться в похожих контекстах. Эта модель ана-

лизирует большой фрагмент текста, вычисляет частоту употребления терминов и создает матрицу, в которой столбцы представляют собой тестовые образцы, строки представляют уникальные термины, встречающиеся в тексте, а ячейки содержат частоту встречаемости для тестовой выборки. Полученная матрица обрабатывается методом, называемым разложением по сингулярным значениям (SVD). SVD - это метод уменьшения размерности, который уменьшает количество столбцов, в то время как информация о сходстве в строках сохраняется. Затем сходство вычисляется с помощью косинусного расстояния в семантическом пространстве.

Семантическое сходство с использованием web-поисковых систем: этот подход предложен в [7]. В этом случае метод пытается найти значение совпадения между двумя словами, используя результаты web-поиска. Метод использует такую информацию, как количество страниц и фрагментов, предоставляемых поисковыми системы для каждого запроса. Для определения сходства между двумя словами P и Q метод определяет количество страниц для запросов P и Q, выполняемых по отдельности, а также количество страниц для объединенного запроса P & Q. Затем выполняется поиск множества лексико-синтаксических шаблонов двух терминов запроса, а также определяется частота встречаемости этих шаблонов в найденных фрагментах.

Полученные шаблоны ранжируются в соответствии с их способностью выражать сходство между двумя словами. Сходство оценки, основанные на количестве страниц, интегрируются с лексико-синтаксическими шаблонами с использованием метода опорных векторов (SVM) для вычисления семантического сходства между заданными словами.

Вычисление сходства на основе корпусной статистики и лексической таксономии предложен в [8], в котором объединена лексическая таксономия и статистическая информация, полученная из корпуса. Метод основан на подходе, при котором информация, собранная из корпуса, используется в качестве решающего фактора. Этот метод по многим характеристикам превосходит предыдущие методы.

Метод явного семантического анализа основан на вычислении семантической взаимосвязи между двумя произвольными текстами путем расширения знаний о словах в текстовом представлении [9]. Значение слова явно представлено в виде взвешенного вектора понятий, заимствованного из Википедии, которая сама по себе является большим хранилищем статей. Два текста, которые необходимо сравнить, преобразуются во взвешенные векторы понятий с использованием машинного обучения. Затем сгенерированные таким образом векторы сравниваются с использованием какого-либо известного метода, например, косинусного подобия для вычисления семантической взаимосвязи между текстами.

Методы, основанные на знаниях, используют такой формальный инструментарий, как таксономии, онтологии, семантические сети. Все это — формы представления знаний, используемые, в частности, в информационном поиске. Они представляют сущности, объекты, их свойства и отношения между ними. При этом весьма часто используются иерархические представления [1]. Подход, основанный на знаниях, может быть дополнительно классифицирован как основанный на графах знаний. Весьма эффективно в этом случае вычисления семантической близости основаны на теории иерархических чисел [10].

Одной из новаторских работ, выполненных в этой области, была модель, предложенная Ву и Палмером [11]. Концептуальная область может быть представлена таксономией, в которой понятия расположены иерархически. Сходство между двумя концепциями, представленными в одной и той же области, может быть рассчитано как мера расстояния между двумя концепциями, представленными в иерархической структуре. Если два концепта C1 и C2 между которыми требуется вычислить сходство, и C3 является их ближайшим обобщающим понятием, то сходство между двумя концепциями выражается так:

$$\text{Similarity}(C1, C2) = (N3 * 2) / (N1 + N2),$$

где N_1 — количество ребер на пути между C_1 и корневой вершиной таксономии, N_2 — количество ребер на пути между C_2 и корневым понятием, а N_3 — количество ребер на пути между C_3 и корневым узлом. Этот способ подсчета ребер основан на предположении, что все ребра имеют равные веса.

В исследовании Резника [11] высказано предположение о том, что проблема с мерами, основанными на подсчете ребер, семантическое сходство заключается в том, что все ребра в таксономии не представляют равного семантического расстояния между терминами, которые они соединяют. Например, в таксономии WordNet[12] есть единственная связь между монетой и копеей и одна связь между сокровищем и золотом, но значение копейки более тесно связано с монетой, чем значение золота с сокровищем.

Резник заметил в своей статье, что нахождение семантического сходства между двумя словами — это, по сути дела, определение степени, в которой они определяют общую информацию. Это можно сделать, найдя наименьшее общее понятие. По мере продвижения вверх по иерархии в таксономии осуществляется переход от более конкретных понятий к более абстрактным. Далее предполагается, что чем меньше расстояние между двумя понятиями и их общим разделом текста, тем больше семантическое сходство между ними.

Таким образом, Резник предложил количественно оценить общий информационный контент для двух концептов, чтобы измерить сходство между ними. Чем больше общего информационного контента, тем концепты более похожи. Информационное содержание, разделяемое двумя концептами, является информационным содержанием концепта, который их объединяет. Взяв за основу теорию информации, модель вычисляет семантическое сходство между концептами C_1 и C_2 с использованием следующей формулы:

$$\text{Similarity}(C_1, C_2) = \text{Max}_{c \in S(C_1, C_2)} (-\log p(c)),$$

где $S(C_1, C_2)$ — это множество концептов, которые включают в как C_1 , так и C_2 , а $(-\log p(c))$ определяет количество информации для концепта C .

Рада и др. предложили метод, основанный на измеренных концептуальных расстояниях между терминами используя таксономию, определяющую эти термины [13]. Исходя из предположения, что чем меньшее расстояние существует между двумя узлами в таксономии, большее семантическое сходство будет между этими узлами. В подходе отыскивается кратчайший путь между двумя терминами в заданной таксономии. Сходство между двумя терминами затем вычисляется по следующей формуле:

$$\text{Similarity}(C_1, C_2) = 2 * \text{MAX} - L,$$

где MAX — максимальная длина пути между узлами C_1 и C_2 из родовидовой таксономии, а L — длина кратчайшего пути между этими узлами. Преимуществом этого метода является его низкая вычислительная сложность при расчете расстояний.

Автором настоящей статьи предложен новый метод оценки семантического сходства, основанный на десятичных иерархических числах. В этом случае при вычислении близости концептов a и b используется функция $\text{Max}(a_i, b_j)$ как максимальное значение иерархического индекса для этих концептов, рассчитанное на множестве нескольких таксономий. Такими таксономиями в минимальном случае являются родовидовая, причинно-следственная и меронимическая (таксономия «часть-целое»). Для вычисления используется следующее выражение:

$$\text{Similarity} = 2 * \frac{\sum_{i=1}^n \sum_{j=1}^m 2 * \text{Max}(a_i, b_j) / (|a_i| + |b_j|)}{|a| + |b|},$$

где операция «|a|» представляет собой вычисление сложности иерархического индекса понятия **a**, однозначно характеризующего глубину в таксономическом дереве, на которой расположен узел, соответствующий этому концепту.

Серия экспериментов в предметной области англоязычных политических статей из отечественных и западных средств массовой информации была проведена на основе анализа следующих электронных ресурсов [14]: msnbc, bloomberg, cnn, springer, nbcnew, theguardian, nytimes, france24, RT, Meduza, Kremlin, Globalaffairs, Themoscwtimes, RussiaBeyond, Rossiyasegodnya, Interfax, SputnikInternational.

Результаты испытаний показали высокую эффективность этого метода по характеристикам точности, сопоставимую с применением лучших из рассмотренных здесь методов, и с явным преимуществом в вычислительной сложности, которая в среднем на 25% ниже сложности рассмотренных методов.

Библиографический список

1. И.Ю. Каширин. Бинарные иерархические числа для вычисления семантической близости предложений естественного языка. // Вестник РГРТУ. 2023. № 86 / Vestnik of RSREU. 2023. С.110-121. DOI: 10.21667/1995-4565-2023-86-110-121.
2. Wael H Gomaa, Aly A Fahmy, "A survey of text similarity approaches", International Journal of Computer Applications, Foundation of Computer Science, vol 68, 2013.
3. И.Ю.Каширин. Формирование LLM-ориентированных ресурсов знаний на основе генерации дополненной информации. Вестник РГРТУ. 2025. № 94 / Vestnik of RSREU. 2025. No 94. С.85-97. DOI: 10.21667/1995-4565-2025-94-85-97.
4. Yung-Shen Lin, Jung-Yi Jiang, Shie-Jue Lee, "A similarity measure for text classification and clustering", IEEE transactions on knowledge and data engineering, vol 26, pp. 1575—1590, 2014.
5. Rada Mihalcea , Courtney Corley, Carlo Strapparava , "Corpus-based and knowledge-based measures of text semantic similarity", American Association for Artificial Intelligence, vol 6, pp. 775—780, 2006.
6. Landauer T.K. and Dumais S.T. (1997) A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104: 11–140.
7. Danushka Bollegala, Yutaka Matsuo, Mitsuru Ishizuka, "Measuring semantic similarity between words using web search engines", WWW, vol 7, pp. 757—766, 2007.
8. Jay J Jiang, David W Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", arXiv preprint cmp-lg/9709008, 1997.
9. Evgeniy Gabrilovich, Shaul Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis", IJCAI, vol 7, pp. 1606—1611, 2007.
10. И.Ю. Каширин. Эмбединги иерархических чисел для обогащения трансформерных языковых моделей внешними онтологическими знаниями. Вестник РГРТУ. 2025. № 93 / Vestnik of RSREU. 2025. No 93.С.72-82. DOI: 10.21667/1995-4565-2025-93-72-82.
11. Zhibiao Wu, Martha Palmer, "Verbs semantics and lexical selection", Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 133—138, 1994.
12. Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.
13. Roy Rada, Hamed Mili, Ellen Bicknell, Maria Blettner, "Development and application of a metric on semantic nets", IEEE transactions on systems, man, and cybernetics, Vol 19, 17—30, 1989.
14. И.Ю. Каширин. Токенизация политических текстов в BERT-моделях с использованием ICF+-онтологий. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ, Том 30, № 12, 2024, с.622-632, DOI: 10.17587/it.30.622-632.

УДК 004.01; 004.04; 004.5; 004.65; 006.01; 006.3/.8; 006.39; ГРНТИ 84.15.03; 28.15.15; 28.15.19

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ ФОРМИРОВАНИЯ ЭЛЕКТРОННЫХ ИНТЕРАКТИВНЫХ РАСЧЁТНЫХ СТАНДАРТОВ

Е.А. Игнатов

*ФГАОУ ВО «Московский политехнический университет»,
Российская Федерация, Москва, ignatov-evgenij@yandex.ru*

Аннотация. В работе рассмотрены расчётные методики, изложенные в различных стандартах. Выявлены основные математические функции и программные операторы, применяемые в этих методиках, в контексте построения информационной системы для автоматизации расчётов. Отмечено, что важным инструментом для автоматизации технической документации, в первую очередь стандартов, является концепция SMART-стандартов. Проведен анализ структуры классических стандартов и SMART-стандартов. Предложено расширение множества информационных блоков и информационных элементов SMART-стандартов, позволяющих задавать автоматизированные расчёты.

Ключевые слова: SMART-стандарты, электронные интерактивные расчётные стандарты, расчётные методики, автоматизация расчётов.

MATHEMATICAL AND SOFTWARE SUPPORT FOR THE INFORMATION SYSTEM FOR THE FORMATION OF ELECTRONIC INTERACTIVE CALCULATION STANDARDS

E.A. Ignatov

*Moscow Polytechnic University,
Russian Federation, Moscow, ignatov-evgenij@yandex.ru*

The summary. This paper examines the calculation methods outlined in various standards. The key mathematical functions and software operators used in these methods are identified in the context of building an information system for automated calculations. It is noted that the concept of SMART standards is an important tool for automating technical documentation, particularly standards. The structure of classical and SMART standards is analyzed. An expansion of the set of information blocks and information elements of SMART standards is proposed, enabling automated calculations.

Keywords: SMART standards, electronic interactive calculation standards, calculation methods, automation of calculations.

Необходимость автоматизации и цифровизации современных компаний – важнейшее условие повышения их конкурентоспособности. Одной из ключевых задач в данном направлении является перевод технической документации в цифровой вид. Особенно важным инструментом, для решения данной задачи являются SMART-стандарты [1, 2, 3, 4], позволяющие перевести стандарты и другую техническую документацию на новый уровень автоматизации. Данный инструментальный предоставляет возможность сделать её не только человеко-, но и машиночитаемой.

Многие стандарты содержат в своём составе методики расчёта, представляющие собой детерминированные модели [5]. Для автоматизации методик расчёта, содержащихся в стандартах и иной технической документации, требуется специализированное программное обеспечение, которое обладало бы инструментарием, позволяющим задавать необходимые математические функции и программные операторы, как для фиксации методики расчёта, так и для выполнения расчёта по данной методике. На сегодняшний день для автоматизации расчётных методик применяется ряд программных решений. В этот перечень входят такие известные системы, как: MathCad, Mathematica, Blockpad, CalcTree и другие. Однако, они не ориентированы на применение концепции SMART-стандартов, что снижает их эффективность и представляет собой проблему. Решением данной проблемы является объединение функционала SMART-стандартов с функционалом программ для автоматизации расчётов. Для этого необходимо разработать информационную систему (далее ИС), позволяющую объединить оба этих функционала в едином решении. Любая ИС для автоматиза-

ции расчётов должна базироваться на математическом и алгоритмическом обеспечении, позволяющем охватить ту предметную область, на которую она ориентирована. Соответственно, необходимо провести анализ стандартов, содержащих расчёты для определения требуемого функционала ИС, организованной по принципу SMART-стандартов и позволяющей автоматизировать расчёты.

Таким образом, целью данной работы является определение перечня математических и программных операторов, необходимых для автоматизации расчётных методик по принципу SMART-стандартов. Для достижения данной цели, требуется решение следующих задач:

- 1) Провести анализ стандартов, содержащих расчётные методики. Выявить структуру стандартов и SMART-стандартов;
- 2) Выявить перечень необходимого математического функционала для задания расчётных методик;
- 3) Выявить перечень необходимых программных операторов для задания расчётных методик.

Структура стандартов

Множество стандартов содержат расчётные методики. Для реализации ИС, осуществляющей автоматизацию расчётов, необходимо формализовать математический аппарат, позволяющий задавать расчётные методики.

В стандарте [6] описана структура элементов стандарта. Данная структура представляет собой множество: $St(t) = \{Ob, nOb\}$, где $St(t)$ – структура версии стандарта, утверждённая на определённую дату. $Ob(t) = \{Tl, Pr, Op, Naim, Onp, Bd\}$ – обязательные элементы стандарта, в число которых входят: $Tl(t)$ – титульный лист; $Pr(t)$ – предисловие; $Op(t)$ – область применения; $Naim(t)$ – наименование; $Onp(t)$ – основные нормативные положения; $Bd(t)$ – библиографические данные.

$nOb(t) = \{Sod, Vv, TiO, OiS, Ns, Pril, Bib\}$ – необязательные элементы стандарта, в число которых входят: $Sod(t)$ – содержание; $Vv(t)$ – введение; $TiO(t)$ – термины и определения; $OiS(t)$ – обозначения и сокращения; $Ns(t)$ – нормативные ссылки; $Pril(t)$ – приложения; $Bib(t)$ – библиография.

В стандарте [2] описана структура информационных блоков SMART-стандарта. Данная структура также представляет собой множество. $SmSt(t) = \{SmOb, nSmOb\}$, где $SmSt(t)$ – структура версии SMART-стандарта, утверждённая на определённую дату.

Здесь $SmOb(t) = \{At, Tl, Pr, Op, Onp, Bd, Ui, SvEp\}$ – обязательные блоки SMART-стандарта, в число которых входят: $At(t)$ – атрибуты; $Tl(t)$ – титульный лист; $Pr(t)$ – предисловие; $Op(t)$ – область применения; $Onp(t)$ – основные нормативные положения; $Bd(t)$ – библиографические данные; $Ui(t)$ – условия использования; $SvEp(t)$ – сведения об электронных подписях.

$nSmOb(t) = \{Sd, T, Vv, TiO, OiS, Og, Ns, Pril, Bib, Dmd\}$ – необязательные блоки SMART-стандарта, в число которых входят: $Sd(t)$ – сопутствующие данные; $T(t)$ – текст; $Vv(t)$ – введение; $TiO(t)$ – термины и определения; $OiS(t)$ – обозначения и сокращения; $Og(t)$ – оглавление; $Ns(t)$ – нормативные ссылки; $Pril(t)$ – приложения; $Bib(t)$ – библиография; $Dmd(t)$ – дополнительные метаданные.

Также в стандарте [2] описана структура информационных элементов SMART-стандарта. Данная структура представляет собой множество:

$SmIESt(t) = \{Atr, SE, Term, Sokr, Abz, Sp, Tabl, 2D, 3D, Vid, Aud, Form, Gip, SO, SEK, File, PC, Pok, PP, EP, FINP, NP\}$, где $SmIESt(t)$ – структура информационных элементов SMART-стандарта, утверждённая на определённую дату. $Atr(t)$ – атрибут; $SE(t)$ – структурный элемент; $Term(t)$ – термин; $Sokr(t)$ – сокращение; $Abz(t)$ – абзац; $Sp(t)$ – список, перечисление; $Tabl(t)$ – таблица; $2D(t)$ – графическое изображение (2D); $3D(t)$ – 3D-модель,

$Vid(t)$ – видео; $Aud(t)$ – аудио; $Form(t)$ – формула; $Gip(t)$ – гиперссылка; $SO(t)$ – ссылка на объект; $SEK(t)$ – ссылка на элемент классификатора; $File(t)$ – файл; $PC(t)$ – программный исполняемый код; $Pok(t)$ – показатель; $PP(t)$ – параметр показателя; $EP(t)$ – электронная подпись; $FINP(t)$ – формализованное изложение нормативного положения; $NP(t)$ – нормативное положение.

Необходимо отметить, что наилучшим образом структурированы стандарты, содержащие машиностроительные расчётные методики. В таких стандартах содержатся исходные данные, необходимые для расчёта, расчётные параметры с формулами и примечаниями, таблицы и графики для выбора значений переменных.

На рисунке 1 приведён пример представления исходных данных из стандарта [7].

2. РАСЧЕТ ГЕОМЕТРИЧЕСКИХ ПАРАМЕТРОВ

Таблица 1

Исходные данные для расчета			
Наименование параметров		Обозначения	Расчетные формулы и указания
Число зубьев	шестерни	z_1	—
	колеса	z_2	—
Модуль		m	—
Угол наклона		β	—
Нормальный исходный контур	Угол профиля	α	—
	Коэффициент высоты головки	h_a^*	—
	Коэффициент радиуса кривизны переходной кривой	ρ_f^*	—
	Коэффициент граничной высоты	h_f^*	—
	Коэффициент радиального зазора	c^*	—
	Линия модификации головки	—	—
	Коэффициент высоты модификации головки	h_g^*	—
	Коэффициент глубины модификации головки	Δ^*	—
Межосевое расстояние		a_w	Входит в состав исходных данных, если его значение задано

Рис. 1. Пример представления исходных данных для расчёта из машиностроительного стандарта [7]

На рисунке 2 приведён пример представления части расчёта из стандарта [7].

Расчет размеров для контроля взаимного положения одноименных профилей зубьев

Наименования параметров	Обозначения	Расчетные формулы и указания
1. Шаг зацепления	p_α	$p_\alpha = \pi m \cos \alpha$
2. Осевой шаг	p_x	$p_x = \frac{\pi m}{\sin \beta}$
3. Ход	p_z	$p_z = z p_x$

Рис. 2. Пример представления части расчёта из машиностроительного стандарта [7]

В стандартах часто применяется выбор значений переменных графическим способом.

На рисунке 3 приведён пример выбора значений переменных графическим способом из стандарта [7].

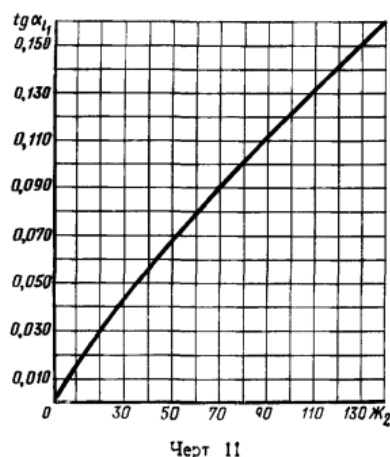
График для определения величины $\lg z_{f1}$ в зависимости от вспомогательной величины J_1 ($\alpha = 20^\circ$)

Рис. 3. Пример выбора значений переменных графическим способом из стандарта [7]

Одним из наиболее распространённых способов определения значений переменных, применяемых в стандартах, выступает выбор значений из таблиц по ключевому параметру или нескольким параметрам.

На рисунке 4 приведён пример выбора значений переменных посредством таблиц по ключевому параметру из стандарта [8].

Таблица 1 — Внутреннее кольцо нормального класса точности

Допуски в микрометрах																	
d, мм					Δ _{dmp}		V _{dsp}			V _{dmp}	K _{ia}	S _d	S _{ia} ¹⁾ , S _i ¹⁾	Δ _{Bs}			V _{Bs} , не более
							серии диаметров							подшипник			
							0, 8, 9	1, 7	2, 3, 4, 5, 6					любой	отдельный	комплектный ²⁾	
					верхн.	нижн.	не более						верхн.	нижн.			
До	0,6	включ.	0	−8	10	8	6	6	6	10	20	24	0	−40	—	12	
Св. 0,6	до	2,5	»	0	−8	10	8	6	6	10	20	24	0	−40	—	12	
»	2,5	»	10	»	0	−8	10	8	6	6	10	20	24	0	−120	−250	15
»	10	»	18	»	0	−8	10	8	6	6	10	20	24	0	−120	−250	20
»	18	»	30	»	0	−10	13	10	8	8	13	20	24	0	−120	−250	20

Рис. 4. Пример выбора значений переменных посредством таблиц по ключевому параметру из стандарта [8]

Исходя из вышеописанного следует, что в структуру информационных блоков SMART-стандартов [1, 2], целесообразно добавить блок расчёт. Информационный блок расчёт представляет собой законченную структуру, в котором происходит последовательный процесс расчёта, инициируемый заданием исходных данных с дальнейшим определением ряда параметров. Рассчитываемые параметры могут быть как определены по формулам, так и выбраны из баз данных по ключевому параметру, либо графическим способом. Таким образом в структуру необязательных информационных блоков SMART-стандарта необходимо добавить информационный блок расчёт:

$$nSmOb(t) = \{Sd, T, Vv, TiO, OiS, Og, Ns, Pril, Bib, Dmd, Calc\},$$

где $Calc(t)$ – расчёт.

В свою очередь, информационный блок расчёт также представляет собой множество, в которое входит 4 информационных элемента: $Calc(t) = \{Id, RP, DGr, DBd\}$, где $Id(t)$ – исходные данные; $RP(t)$ – расчётные параметры; $DGr(t)$ – данные, выбираемые графическим способом; $DBd(t)$ – данные, выбираемые из базы данных.

Общий перечень информационных элементов будет представлять собой расширенное множество: $SmIESt(t) = \{Atr, SE, Term, Sokr, Abz, Sp, Tabl, 2D, 3D, Vid, Aud, Form, Gip, SO, SEK, File, PC, Pok, PP, EP, FINP, NP, Id, RP, DGr, DBd\}$.

В соответствии с предлагаемой концепцией, SMART-стандарт, содержащий в своём составе информационный блок $Calc(t)$, представляет собой электронный интерактивный расчётный стандарт (далее ЭИРС).

Математическое обеспечение расчётных методик

С целью реализации ИС для автоматизации расчётов необходимо формализовать математический аппарат, позволяющий задавать расчётные методики.

В расчётных методиках, изложенных в стандартах, широко применяются арифметические функции, в том числе: тригонометрические, степенные, экспоненциальные, логарифмические и иные.

На рисунке 5 приведён пример использования тригонометрических функций для расчётной методики из стандарта [7].

Наименования параметров	Обозначения	Расчетные формулы и указания
Расчет постоянной хорды зуба и высоты до постоянной хорды		
1. Постоянная хорда зуба	шестерни	$\bar{s}_{c1} = \left(\frac{\pi}{2} \cos^2 \alpha + x_1 \sin 2\alpha \right) m$
	колеса	$\bar{s}_{c2} = \left(\frac{\pi}{2} \cos^2 \alpha + x_2 \sin 2\alpha \right) m$
		При $\alpha = 20^\circ$ (включая исходные контуры по ГОСТ 13755-81 и ГОСТ 9587-81) упрощенный расчет $\bar{s}_{c1}$, $\bar{s}_{c2}$ и $\bar{h}_{c1}$, $\bar{h}_{c2}$ приведен в табл.1 приложения 1 Должно выполняться условие: для шестерни $\rho_{s1} > \rho_{p1}$, для колеса $\rho_{s2} < \rho_{p2}$, где ρ_s - радиус кривизны разноименных профилей зуба зубчатого колеса в точках, определяющих постоянную хорду; $\rho_{s1} = 0,5 \left(d_{b1} \operatorname{tg} \alpha_t + \bar{s}_{c1} \frac{\cos \beta_b}{\cos \alpha} \right),$ $\rho_{s2} = 0,5 \left(d_{b2} \operatorname{tg} \alpha_t + \bar{s}_{c2} \frac{\cos \beta_b}{\cos \alpha} \right),$

Рис. 5. Пример расчётных методик с применением тригонометрических функций из стандарта [7]

На рисунке 6 приведён пример использования логарифмических и степенных функций для расчётной методики из стандарта [9].

Г.6 Расчет объединенных показателей дискомфорта

Вычисления UGR проводят, используя формулу

$$UGR = 8 \lg \sum_{i=1}^N \left[\frac{K_i I_{C_{\eta i}}^2}{A_i} \right] - 8 \lg E_{wid}. \quad (\text{Г.2})$$

При направлении линии зрения поперек продольной оси светильников

$$UGR = 8 \lg \left\{ \left(\frac{0,0041 \cdot 83,03^2}{0,169} \right) + \left(\frac{0,0054 \cdot 47,12^2}{0,115} \right) + \left(\frac{0,0047 \cdot 29,70^2}{0,088} \right) \right\} - 8 \lg(27,37);$$

$$UGR = 8 \lg \left\{ \left(\frac{0,0041 \cdot 83,03^2}{0,169} \right) + \left(\frac{0,0054 \cdot 47,12^2}{0,115} \right) + \left(\frac{0,0047 \cdot 29,70^2}{0,088} \right) \right\} - 8 \lg(27,37);$$

$$UGR = 8 \lg (167,25 + 104,26 + 48,21 + 167,25 + 104,26 + 48,21) - 8 \lg (27,37) =$$

$$= 22,45 - 11,50 = 10,95 \approx 11,0.$$

При направлении линии зрения вдоль продольной оси светильников:

$$UGR = 8 \lg \left\{ \left(\frac{0,0041 \cdot 106,55^2}{0,169} \right) + \left(\frac{0,0054 \cdot 65,57^2}{0,115} \right) + \left(\frac{0,0047 \cdot 46,26^2}{0,088} \right) \right\} - 8 \lg(27,37);$$

$$UGR = 8 \lg \left\{ \left(\frac{0,0041 \cdot 106,55^2}{0,169} \right) + \left(\frac{0,0054 \cdot 65,57^2}{0,115} \right) + \left(\frac{0,0047 \cdot 46,22^2}{0,088} \right) \right\} - 8 \lg(27,37);$$

$$UGR = 8 \lg (275,43 + 201,89 + 116,75 + 275,43 + 201,89 + 116,75) - 8 \lg (27,37) =$$

$$= 24,60 - 11,50 = 13,1.$$

Рис. 6. Пример расчётных методик с применением логарифмических и степенных функций из стандарта [9]

На рисунке 7 приведён пример использования экспоненциальных и степенных функций для расчётной методики из стандарта [10].

Значения показателя PMV рассчитывают по формулам (1)-(4):

$$PMV = [0,303 \exp(-0,036M) + 0,028 \cdot (M - W) - 3,05 \cdot 10^{-3} [5733 - 6,99 \cdot (M - W) - p_a] - 0,42 \cdot [(M - W) - 58,15] - 1,7 \cdot 10^{-5} \cdot M(5867 - p_a) - 0,0014 \cdot M(34 - t_a) - 3,96 \cdot 10^{-8} f_{cl} \cdot [(t_{cl} + 273)^4 - (t_r + 273)^4] - f_{cl} h_c (t_{cl} - t_a); \quad (1)$$

$$t_{cl} = 35,7 - 0,028(M - W) - I_{cl} - \left\{ 3,96 \cdot 10^{-8} f_{cl} [(t_{cl} + 273)^4 - (t_r + 273)^4] + f_{cl} h_c (t_{cl} - t_a) \right\}; \quad (2)$$

$$h_c = \begin{cases} 2,38 |t_{cl} - t_a|^{0,25}, & \text{если } 2,38 |t_{cl} - t_a|^{0,25} > 12,1 \sqrt{v_{ar}}, \\ 12,1 \sqrt{v_{ar}}, & \text{если } 2,38 |t_{cl} - t_a|^{0,25} < 12,1 \sqrt{v_{ar}}; \end{cases} \quad (3)$$

$$f_{cl} = \begin{cases} 1,00 + 1,290 I_{cl}, & \text{если } I_{cl} \leq 0,078 \text{ м}^2 \cdot \text{кВт}, \\ 1,05 + 1,645 I_{cl}, & \text{если } I_{cl} > 0,078 \text{ м}^2 \cdot \text{кВт}, \end{cases} \quad (4)$$

Рис. 7. Пример расчётных методик с применением экспоненциальных и степенных функций из стандарта [10]

На основе проведенных исследований было установлено, что в расчётных методиках, применяемых в стандартах и иной технической документации, наиболее часто используются следующие функции:

+
 -
 *
 /
 sin(x)
 cos(x)
 tan(x)
 cotang(x)
 arcsin(x)
 arccos(x)
 arctan(x)
 arccotang(x)
 log(x)
 ln(x)
 lg(x)
 e^x
 x^y
 sqr(x)
 random(x)
 π
 x!
 и другие

Применение вышеперечисленных функций позволяет решать подавляющую часть задач используемых в стандартах и иных технических документах.

Отрезки и интервалы

В стандартах часто используются отрезки и интервалы, которые предназначены для выполнения двух функций: 1) Выбор по отрезкам или интервалам значений, содержащихся в таблицах. В качестве примера можно привести выбор значений переменных из стандарта

[11]. В данном стандарте переменные назначаются по ключевым параметрам, заданным отрезками и интервалами, что представлено на рисунке 8.

Степень точности	Обозначение	Модуль m , мм	Ширина зубчатого венца (полушестерня) или длина контактной линии, мм					
			До 40	Св. 40 до 100	Св. 100 до 160	Св. 160 до 250	Св. 250 до 400	Св. 400 до 630
			мм					
10	F_k	От 1 до 3,5 Св. 3,5 » 6,3 » 6,3 » 10 » 10 » 16 » 16 » 25 » 25 » 40 » 40 » 55	90 100 125 — — — —	100 112 125 160 200 — —	100 112 140 160 200 280 —	112 125 140 180 224 280 355	125 140 160 200 224 315 400	— 180 200 224 250 315 400
	F_β	От 1 до 55	45	63	80	100	112	140
	f_α	От 1 до 55	45	63	80	100	112	140
	f_y	От 1 до 55	22	32	40	50	56	71
11	F_k	От 1 до 3,5 Св. 3,5 » 6,3 » 6,3 » 10 » 10 » 16 » 16 » 25 » 25 » 40 » 40 » 55	140 160 200 — — — —	160 180 200 250 315 — —	160 180 224 250 315 450 —	180 200 224 280 355 450 560	200 224 250 315 355 500 630	— 280 315 355 400 500 630
	F_β	От 1 до 55	71	100	125	160	180	224
	f_α	От 1 до 55	71	100	125	160	180	224
	f_y	От 1 до 55	36	50	63	80	90	112
12	F_k	От 1 до 3,5 Св. 3,5 » 6,3 » 6,3 » 10 » 10 » 16 » 16 » 25 » 25 » 40 » 40 » 55	224 350 315 — — — —	250 280 315 355 500 — —	250 280 355 400 500 630 —	280 315 355 450 560 710 900	315 355 400 500 560 710 1000	— 450 500 560 630 800 1000
	F_β	От 1 до 55	112	160	200	250	280	355
	f_α	От 1 до 55	112	160	200	250	280	355
	f_y	От 1 до 55	56	80	100	125	140	180

П р и м е ч а н и я:

1. Принятые обозначения:

$\pm F_{pm}$ — предельные отклонения осевых шагов по нормали;

F_k — допуск на суммарную погрешность контактной линии;

F_β — допуск на направление зуба;

f_α — допуск параллельности осей;

f_y — допуск на перекос осей.

2. F_k определяется в зависимости от длины контактной линии. Для зубчатых колес с номинальным коэффициентом осевого перекрытия ε_β до 0,8 допуск на суммарную погрешность линии F_k принимается:

$$\begin{aligned} &\text{при } \varepsilon_\beta \text{ от } 0 \text{ до } 0,2 & F_k &= F_\beta; \\ &\text{» } \varepsilon_\beta \text{ св. } 0,2 \text{ » } 0,4 & F_k &= 1,3 F_\beta; \\ &\text{» } \varepsilon_\beta \text{ » } 0,4 \text{ » } 0,6 & F_k &= 1,5 F_\beta; \\ &\text{» } \varepsilon_\beta \text{ » } 0,6 \text{ » } 0,8 & F_k &= 1,7 F_\beta. \end{aligned}$$

Рис. 8. Пример выбора значений переменных по ключевым параметрам, заданным отрезками и интервалами, из стандарта [11]

2) Представление значения переменных в виде отрезка или интервала. Например, при назначении прочностных параметров материалов из стандарта [12] $\sigma_b = 700-850$.

Обычно в стандартах с расчетными методиками используются следующие варианты представления отрезков и интервалов: 1) От 1 До 3; 2) Св. 2 До 2,8; 3) 1,5 – 2,5; 4) Св. 1,5; 5) До 6 и так далее. Однако, данное представление не всегда позволяет однозначно трактовать представленную информацию, поскольку неясно входят ли в диапазон представляемых величин их крайние значения (открытый интервал или закрытый). Для того, чтобы сделать запись более понятной, необходимо указывать, входит ли крайнее значение величины в представляемый диапазон или нет. Например: От 2 включительно До 3 включительно. Однако данное представление достаточно громоздкое. Для использования в ИС предлагается взять в качестве основы стандартные математические обозначения отрезков и интервалов: 1) $[a,b]$ – замкнутый интервал (отрезок); (a,b) – открытый интервал (интервал); 3) $[a,b)$ – полуоткрытый (или полузакрытый) интервал; $(a,b]$ – полуоткрытый (или полузакрытый) интервал.

Программные операторы и логические операции

Для представления логических операций предполагается использовать стандартные фразы: 1) НЕ; 2) И; 3) ИЛИ.

В расчетах часто приходится применять условия и циклы.

Пример задания условия из стандарта [13] представлен на рисунке 9.

3. долговечности	Коэффициент	Z_N	По графику (черт.4) или формулам: $Z_N = 6 \sqrt{\frac{N_{Hlim}}{N_K}} \quad \text{при } N_K \leq N_{Hlim},$ но не более 2,6 для однородной структуры материала и 1,8 для поверхностного упрочнения; $Z_N = 20 \sqrt{\frac{N_{Hlim}}{N_K}} \quad \text{при } N_K > N_{Hlim},$ но не менее 0,75. Примечание. При использовании метода эквивалентных циклов вместо N_K подставляют N_{HE}

Рис. 9. Пример задания условия из стандарта [13]

Из приведенного примера наглядно видно, условия расположены в той же ячейки таблицы, что и расчетные формулы. Этот же принцип предлагается использовать и в ИС для ЭИРС. Для циклов предлагается применить такой же подход. Однако в программе предпочтительнее использовать стандартную схему записи условия, применяемую в информатике: ЕСЛИ ТО ИНАЧЕ. В программе блок записи условия, будет выглядеть следующим образом:

1.1)

ЕСЛИ

(Условие)

ТО

[Оператор]

ИНАЧЕ

[Оператор]

Пример: ЕСЛИ $N_K \leq N_{Hlim}$ ТО $Z_N = 6 \sqrt{\frac{N_{Hlim}}{N_K}}$ ИНАЧЕ $Z_N = 20 \sqrt{\frac{N_{Hlim}}{N_K}}$

1.2)

ЕСЛИ

(Условие)

ТО

[Оператор 1],
[Оператор 2],

.....

[Оператор n];

ИНАЧЕ

[Оператор 1],
[Оператор 2],

.....

[Оператор n];

Циклы

В качестве классического примера цикла, можно привести проверку нагрузочной способности зубчатых колёс из стандарта [13], которую необходимо повторять, пока условие не будет выполнено. На практике при не выполнении условия часто приходится менять материал зубчатых колёс, их поверхностное упрочнение, степень точности и/или иные параметры. Данный пример приведён на рисунке 10.

1.4. Нагрузочная способность поверхностей зубьев. Нагрузочная способность поверхностей зубьев обеспечивается при выполнении условий любого критерия (табл. 1).

Таблица 1

Критерии	Условия нагрузочной способности
Напряжение	$\sigma_H \leq \sigma_{HP}$, (6)
	$\sigma_{Hmax} \leq \sigma_{HPmax}$ (7)
Безопасность	$S_H \geq S_{Hmin}$, (8)
	$S_{HSt} \geq S_{HStmin}$ (9)
Ресурс	$N_L \geq N_K$, (10)
	$\sigma_{Hmax} \leq \sigma_{HPmax}$ (11)
Вероятность безотказной работы	$P_H(N_L \geq N_K) \geq P_{Hmin}$, (12)
	$P_{HSt}(\sigma_{HSt} \geq \sigma_{Hmax}) \geq P_{HStmin}$ (13)

Рис. 10. Пример циклов из стандарта [13]

Циклы: 2) Цикл с параметром (for) Для $i = 1$ До 10 с шагом 1. В математике используется другое представление цикла с параметром: Для $i = 1(1)10$. Для параметра i от 1 до 10 с шагом 1.

2.1)

Для $i=1(k)j$
[Оператор]

2.2)

Для $i=1(k)j$
[Оператор 1],
[Оператор 2],
.....
[Оператор n];

3) Цикл с предусловием (repeat). Выполнять Пока.

3.1)

Выполнять Пока выполняется условие (Условие 1)

[Оператор 1],
[Оператор 2],
.....
[Оператор n];

Ключевым в данном случае является слово ПОКА.

Данное представление можно применять для неопытных пользователей, так, как оно характеризует весь смысл выполняемой операции.

3.2) Для краткости допускается использовать следующий вариант:

ПОКА (Условие 1)
[Оператор 1],
[Оператор 2],
.....
[Оператор n];

Данная конструкция более короткая и более удобна для ввода с клавиатуры.

4) Цикл с постусловием (repeat/until). Повторять Пока

4.1)
Выполнять
[Оператор 1],
[Оператор 2],
.....
[Оператор n];

Пока не выполнится условие (Условие 1).

Данное представление характеризует весь смысл выполняемой операции.

4.2)
Выполнять
[Оператор 1],
[Оператор 2],
.....
[Оператор n];

ПОКА (Условие 1).

Данная конструкция также более короткая и более удобна для ввода с клавиатуры.

В работе проведен анализ стандартов, содержащих расчётные методики. Выявлено, что для выполнения перечня задач по автоматизации расчётов, помимо математических функций, требуются программные операторы, а также выборка значений из таблиц по ключевой переменной. Предложен вариант записи условных и циклических алгоритмов.

Предложенное в работе расширение множества информационных блоков и информационных элементов SMART-стандартов позволит автоматизировать процесс выполнения как стандартизированных расчётов, так и формирования новых. Автоматизированные подобным образом расчётные методики, станут не только человекочитаемыми, но и приобретут функционал машиночитаемости. Такие расчётные методики возможно будет применять как в технических, так и в киберфизических автоматизированных системах.

Определенный в работе перечень математических функций и программных алгоритмов позволяет перейти к разработке программного обеспечения для задания расчётных методик по принципу ЭИРС и его последующему применению в практических задачах из широкого круга предметных областей.

Библиографический список

1. ПНСТ 1016-2025. Умные (Smart) стандарты. Архитектура и форматы данных // М.: Российский институт стандартизации. 2025.
2. ПНСТ 864-2023. Умные (Smart) стандарты. Общие положения // М.: Российский институт стандартизации. 2023.

3. Колмыков, Е. А. Как перейти к умным (машиночитаемым) стандартам / Е. А. Колмыков, Ю. В. Воронцова, А. Н. Воронцова // Известия Волгоградского государственного технического университета. – 2022. – № 1(260). – С. 17-20. – DOI 10.35211/1990-5297-2022-1-260-17-20. – EDN BTGBYO.
4. Матушкина, И. Ю. Умный стандарт как инструмент цифровой трансформации / И. Ю. Матушкина, Е. А. Михалева // Техническое регулирование в едином экономическом пространстве : Сборник статей IX Всероссийской научно-практической конференции с международным участием, Екатеринбург, 19 мая 2022 года / Под научной редакцией Б.Н. Гузанова. – Екатеринбург: Российский государственный профессионально-педагогический университет, 2022. – С. 44-48. – EDN VRSPZJ.
5. Игнатов, Е. А. Сочетание детерминированных и стохастических алгоритмов в киберфизических системах управления микроклиматом помещений / Е. А. Игнатов // Материалы пула научно-практических конференций : Материалы VII Международной научно-практической конференции, X Международной научно-практической конференции и VII Международной научно-практической конференции, Керчь, 23–27 января 2026 года. – Керчь: Керченский государственный морской технологический университет, 2026. – С. 184-191. – EDN DLXABJ.
6. ГОСТ Р 1.5—2001. Стандарты межгосударственные, правила и рекомендации по межгосударственной стандартизации. Общие требования к построению, изложению, оформлению, содержанию и обозначению // М.: Стандартинформ. 2010.
7. ГОСТ 19274—73. Передачи зубчатые цилиндрические эвольвентные внутреннего зацепления. Расчёт геометрии // М.: Государственный комитет стандартов совета министров СССР. 1973.
8. ГОСТ 520—2011. ПОДШИПНИКИ КАЧЕНИЯ Общие технические условия // М.: Стандартинформ. 2012.
9. ГОСТ 33392—2015. ЗДАНИЯ И СООРУЖЕНИЯ Метод определения показателя дискомфорта при искусственном освещении помещений // М.: Стандартинформ. 2016.
10. ГОСТ Р ИСО 7730-2009. Эргономика термальной среды. Аналитическое определение и интерпретация комфортности теплового режима с использованием расчета показателей PMV и PPD и критериев локального теплового комфорта // М.: Стандартинформ. 2011.
11. ГОСТ 1643—81. Основные нормы взаимозаменяемости. Передачи зубчатые цилиндрические. Допуски // М.: ИПК Издательство стандартов. 1981.
12. ГОСТ 1050—2013. Металлопродукция из нелегированных конструкционных качественных и специальных сталей. Общие технические условия // М.: Стандартинформ. 2014.
13. ГОСТ 21354—87. Передачи зубчатые цилиндрические эвольвентные внешнего зацепления. Расчёт на прочность // М.: Государственный комитет СССР по стандартам. 1987.

УДК 004.021; ГРНТИ 50.05

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ КВАНТОВО-ИНСПИРИРОВАННОЙ ОПТИМИЗАЦИИ В ЗАДАЧАХ С ВЕЩЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

В.В. Масленников

*МИРЭА – Российский технологический университет,
Российская Федерация, Москва, vldmsn@yahoo.com*

Аннотация. В работе представлено теоретическое обоснование концепции квантово-инспирированной оптимизации, ориентированной на решение задач с вещественными переменными. Предложена формальная вычислительная модель, в которой каждая непрерывная переменная кодируется посредством многоуровневой квантовой системы, а состояние системы описывается матрицей плотности в полном гильбертовом пространстве. На основе принципов квантовой суперпозиции, интерференции и запутанности показано, как вероятностные распределения над дискретными уровнями могут использоваться для стохастической генерации решений в непрерывных областях поиска. Рассмотрены механизмы управления эволюцией системы через унитарные и CPTP-отображения, обеспечивающие динамическую балансировку между глобальным исследованием и локальной эксплуатацией. Особое внимание уделено представлению состояний кудитов через разложение по базису операторов Гейзенберга-Вейля, что позволяет естественным образом интегрировать как когерентные, так и некогерентные эффекты, включая декогеренцию. Предложенная схема применима как к однокритериальным, так и к многокритериальным задачам оптимизации и обеспечивает строгую теоретическую основу для разработки новых квантово-инспирированных алгоритмов.

Ключевые слова: квантово-инспирированная оптимизация, вещественные переменные, кудит, матрица плотности, операторы Гейзенберга-Вейля, квантовая интерференция, управляемая эволюция.

THEORETICAL FOUNDATIONS OF QUANTUM-INSPIRED OPTIMIZATION IN PROBLEMS WITH REAL-VALUED VARIABLES

V.V. Maslennikov

MIREA - Russian Technological University,
Russia, Moscow, vldmsn@yahoo.com

The summary. This paper presents a theoretical foundation for the concept of quantum-inspired optimization tailored to problems with real-valued variables. A formal computational model is proposed, in which each continuous variable is encoded by means of a multilevel quantum system, and the state of the entire system is described by a density matrix in the full Hilbert space. Building upon the principles of quantum superposition, interference, and entanglement, it is shown how probability distributions over discrete levels can be employed for stochastic generation of solutions in continuous search domains. Mechanisms for controlling the system's evolution through both unitary and CPTP maps are discussed, enabling dynamic balancing between global exploration and local exploitation. Particular attention is given to representing qudit states via decomposition in the basis of Heisenberg-Weyl operators, which naturally accommodates both coherent and incoherent effects, including decoherence. The proposed framework is applicable to both single-objective and multiobjective optimization problems and provides a rigorous theoretical basis for the development of novel quantum-inspired algorithms.

Keywords: quantum-inspired optimization, real-valued variables, qudit, density matrix, Heisenberg-Weyl operators, quantum interference, controlled evolution.

Квантово-инспирированная оптимизация представляет собой вычислительную парадигму, основанную на фундаментальных принципах квантовой механики [1] и реализуемую с использованием классических вычислительных архитектур. В отличие от прямого использования квантовых процессоров [2], квантово-инспирированный подход к вычислениям воспроизводит ключевые квантовые явления посредством строгих математических моделей и алгоритмических конструкций, что позволяет извлекать преимущества квантового параллелизма без необходимости использования квантовых вычислительных устройств или систем. Применительно к задачам однокритериальной и многокритериальной оптимизации с вещественными переменными квантово-инспирированный подход к вычислениям обеспечивает эффективное кодирование потенциальных решений в виде вероятностных распределений, порождённых аналогами квантовых состояний. Такая парадигма позволяет одновременно исследовать множество регионов непрерывного пространства поиска и динамически регулировать соотношение между глобальным исследованием и локальной эксплуатацией. Теоретическую основу квантово-инспирированной оптимизации составляют формализмы квантовой теории, в первую очередь принципы квантовой суперпозиции, интерференции и запутанности, а также математические конструкции, описывающие двухуровневые и многоуровневые квантовые системы, включая представление состояний многоуровневых квантовых систем через матрицу плотности и разложение по базису операторов Гейзенберга-Вейля, что обеспечивает строгий математический аппарат, лежащий в основе данной парадигмы.

Одним из центральных положений квантовой механики является принцип суперпозиции [3], который обусловлен линейной структурой гильбертова пространства состояний [4]. Пусть $\mathcal{H}$ – комплексное сепарабельное гильбертово пространство, описывающее совокупность возможных состояний квантовой системы. Тогда любое чистое состояние такой системы может быть представлено нормированным вектором $|\psi\rangle \in \mathcal{H}$, удовлетворяющим условию $\langle\psi|\psi\rangle = 1$, где $\langle\psi| = (|\psi\rangle)^\dagger$ – бра-вектор, эрмитово сопряжённый к кет-вектору $|\psi\rangle$.

Пусть $\{|\phi_i\rangle\}_{i=1}^m$ – конечный набор из $m \in \mathbb{N}$ допустимых чистых состояний квантовой системы, принадлежащих гильбертову пространству $\mathcal{H}$. Согласно принципу суперпозиции, любая линейная комбинация этих состояний также описывает физически реализуемое состояние:

$$|\psi\rangle = \sum_{i=1}^m c_i |\phi_i\rangle, \quad (1)$$

где коэффициенты $c_i \in \mathbb{C}$ представляют собой комплексные амплитуды вероятностей соответствующих базисных состояний $|\phi_i\rangle$. Нормировка вектора состояния требует выполнения условия:

$$\langle\psi|\psi\rangle = \sum_{i=1}^m |c_i|^2 + \sum_{i=1}^m \sum_{j=1, j \neq i}^m c_i^* c_j \langle\phi_i|\phi_j\rangle = 1, \quad (2)$$

где c_i^* обозначает комплексно сопряжённое число к c_i .

Если базисные состояния ортонормированы, то есть удовлетворяют условию $\langle\phi_i|\phi_j\rangle = \delta_{ij}$ (δ_{ij} – символ Кронекера [5]), выражение для нормировки (2) принимает упрощённый вид:

$$\sum_{i=1}^m |c_i|^2 = 1. \quad (3)$$

В этом случае величина $|c_i|^2$ допускает вероятностную интерпретацию, то есть определяет вероятность обнаружения системы в состоянии $|\phi_i\rangle$ при выполнении измерения в данном базисе. Следует подчеркнуть, что до момента измерения система не пребывает ни в одном из базисных состояний, а находится в их когерентной суперпозиции.

Квантовая интерференция [6] является следствием волновой природы квантовых состояний и обусловлена комплексной структурой амплитуд вероятностей в суперпозиционных состояниях. В отличие от классической теории вероятностей, где вероятности различных исходов складываются линейно, в квантовой механике наблюдаемая вероятность определяется как квадрат модуля суммы соответствующих комплексных амплитуд. Эта особенность приводит к возникновению интерференционных членов, которые могут усиливать или ослаблять вероятность тех или иных исходов в зависимости от относительных фаз амплитуд.

Рассмотрим простейший случай двухкомпонентной суперпозиции:

$$|\psi\rangle = c_1 |\phi_1\rangle + c_2 |\phi_2\rangle, \quad (4)$$

где $|\phi_1\rangle$ и $|\phi_2\rangle$ – ортонормированные базисные состояния, а $c_1, c_2 \in \mathbb{C}$ – комплексные амплитуды, удовлетворяющие условию нормировки $|c_1|^2 + |c_2|^2 = 1$.

Вероятность обнаружить систему в состоянии $|\chi\rangle$ определяется по правилу Борна [7]:

$$P(\chi) = |\langle\chi|\psi\rangle|^2 = |c_1 \langle\chi|\phi_1\rangle + c_2 \langle\chi|\phi_2\rangle|^2. \quad (5)$$

Введём обозначения $\alpha_1 = \langle\chi|\phi_1\rangle$ и $\alpha_2 = \langle\chi|\phi_2\rangle$. Тогда уравнение (5) принимает вид:

$$P(\chi) = |c_1\alpha_1 + c_2\alpha_2|^2 = |c_1\alpha_1|^2 + |c_2\alpha_2|^2 + I, \quad (6)$$

где $I = 2 \operatorname{Re}(c_1\alpha_1 \cdot c_2^*\alpha_2^*)$ представляет собой интерференционный член [8], а $\operatorname{Re}(\cdot)$ обозначает действительную часть комплексного числа.

Представим комплексные амплитуды в полярной форме:

$$c_1 = |c_1|e^{i\theta_1}, \quad c_2 = |c_2|e^{i\theta_2}, \quad (7)$$

$$\alpha_1 = |\alpha_1|e^{i\varphi_1}, \quad \alpha_2 = |\alpha_2|e^{i\varphi_2}, \quad (8)$$

где $\theta_1, \theta_2, \varphi_1, \varphi_2$ – угловые параметры, определяющие положение соответствующих комплексных чисел на единичной окружности в комплексной плоскости.

Тогда интерференционный член принимает вид:

$$I = 2|c_1||c_2||\alpha_1||\alpha_2|\cos((\theta_1 - \theta_2) + (\varphi_1 - \varphi_2)). \quad (9)$$

Формула (9) показывает, что полная вероятность $P(\chi)$ может превышать или, наоборот, быть меньше суммы отдельных вкладов $|c_1\alpha_1|^2 + |c_2\alpha_2|^2$ в зависимости от разности фаз:

$$\Delta\Phi = \theta_1 - \theta_2. \quad (10)$$

В случае, когда $\Delta\Phi = 0$, интерференционный член принимает максимальное положительное значение, что соответствует конструктивной интерференции. Напротив, при $\Delta\Phi = \pi$ интерференционный вклад становится минимальным (отрицательным), что приводит к деструктивной интерференции.

С физической точки зрения интерференция обусловлена принципиальной неопределённостью траектории квантовой системы до момента измерения: система не следует одному конкретному пути, а развивается по всем допустимым траекториям одновременно. Вклады этих траекторий в результирующую вероятностную амплитуду суммируются с учётом их относительных фаз, что и порождает интерференционные эффекты. В квантово-инспирированных вычислительных моделях, ориентированных на решение задач оптимизации, управление этими эффектами осуществляется посредством унитарных преобразований, которые целенаправленно изменяют фазовые соотношения между компонентами суперпозиции. Такой подход позволяет усиливать амплитуды, соответствующие перспективным решениям, и подавлять, соответствующие менее эффективным решениям, тем самым направляя поиск в сторону глобального или близкого к нему оптимума [9].

Квантовая запутанность [10] является одним из наиболее характерных проявлений неклассического поведения квантовых систем и отражает наличие нелокальных корреляций между их подсистемами, которые не могут быть воспроизведены в рамках классической теории вероятностей. Рассмотрим составную квантовую систему, гильбертово пространство которой задаётся тензорным произведением пространств её частей:

$$\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B, \quad (11)$$

где $\mathcal{H}_A$ и $\mathcal{H}_B$ – гильбертовы пространства подсистем A и B соответственно.

Состояние $|\Psi\rangle \in \mathcal{H}$ называется запутанным, если его невозможно представить в виде сепарабельного тензорного произведения состояний подсистем:

$$|\Psi\rangle \neq |\psi_A\rangle \otimes |\psi_B\rangle, \quad (12)$$

где $|\psi_A\rangle \in \mathcal{H}_A$, $|\psi_B\rangle \in \mathcal{H}_B$.

Классическим примером запутанного состояния двух двухуровневых систем служит одно из состояний Белла [11]:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle). \quad (13)$$

В состоянии (13) результат измерения первого кубита мгновенно определяет состояние второго, независимо от расстояния между ними, что иллюстрирует сильную квантовую корреляцию, не имеющую аналогов в классической физике.

Для количественной характеристики степени запутанности применяются различные меры [12], такие как конкуренция, энтропия фон Неймана [13] или отрицательность. Однако в контексте квантово-инспирированной оптимизации принципиальное значение имеет не столько количественная оценка, сколько сам факт существования неразложимых суперпозиций в пространстве совместных состояний, позволяющих моделировать сложные взаимосвязи между переменными.

В общем случае для N -частичной системы ($N \in \mathbb{N}$) состояние $|\Psi\rangle$ является запутанным, если тензор амплитуд $c_{i_1 i_2 \dots i_N} \in \mathbb{C}$ не допускает факторизации на произведение тензоров меньшего порядка:

$$|\Psi\rangle = \sum_{i_1, i_2, \dots, i_N} c_{i_1 i_2 \dots i_N} |i_1\rangle \otimes |i_2\rangle \otimes \dots \otimes |i_N\rangle, \quad (14)$$

где $c_{i_1 i_2 \dots i_N}$ – комплексная амплитуда вероятности совместного нахождения составной квантовой системы в базисном состоянии, в котором первая подсистема находится в состоянии $|i_1\rangle$, вторая – в $|i_2\rangle$ и N -я – в $|i_N\rangle$, индексы $i_1, i_2, \dots, i_N$ – метки базисных состояний соответствующих подсистем.

В рамках квантово-инспирированной оптимизации, по аналогии с квантовыми вычислениями, элементарной единицей представления информации служит кубит – двухуровневая квантовая система, состояние которой описывается нормированным вектором в двумерном комплексном гильбертовом пространстве $\mathcal{H} \cong \mathbb{C}^2$. Формально состояние кубита задаётся суперпозицией ортонормированного вычислительного базиса $\{|0\rangle, |1\rangle\}$:

$$|\psi\rangle = \alpha |0\rangle + \beta |1\rangle, \quad (15)$$

где $\alpha, \beta \in \mathbb{C}$ – комплексные амплитуды, удовлетворяющие условию нормировки $|\alpha|^2 + |\beta|^2 = 1$, $|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $|1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. При измерении в стандартном базисе кубит коллапсирует в одно из базисных состояний: $|0\rangle$ с вероятностью $|\alpha|^2$ или $|1\rangle$ с вероятностью $|\beta|^2$.

Кудит представляет собой обобщение кубита на d -уровневую квантовую систему ($d \geq 3$) и допускает нахождение в суперпозиции более чем двух базисных состояний. В отличие от двухуровневого кубита, состояние кудита описывается нормированным вектором в d -мерном комплексном гильбертовом пространстве $\mathcal{H} \cong \mathbb{C}^d$.

Формально чистое состояние кудита задаётся линейной комбинацией ортонормированного вычислительного базиса $\{|0\rangle, |1\rangle, \dots, |d-1\rangle\}$:

$$|\psi\rangle = \sum_{k=0}^{d-1} \alpha_k |k\rangle = (\alpha_0, \alpha_1, \dots, \alpha_{d-1})^T \in \mathbb{C}^d, \quad (16)$$

где $\alpha_k \in \mathbb{C}$ – комплексные амплитуды вероностей, удовлетворяющие условию нормировки $\sum_{k=0}^{d-1} |\alpha_k|^2 = 1$. При измерении в стандартном базисе система коллапсирует в одно из базисных состояний $|k\rangle$ с вероятностью $|\alpha_k|^2$.

В общем случае состояние квантовой системы может быть не только чистым, но и смешанным, что возникает либо из-за классической неопределённости относительно реализуемого состояния, либо вследствие взаимодействия с окружением, приводящего к декогеренции [14]. Для описания таких состояний используется матрица плотности ρ – статистический оператор, обобщающий понятие чистого состояния.

Пусть квантовая система описывается ансамблем $\{p_i, |\psi_i\rangle\}_{i=1}^m$, где $|\psi_i\rangle \in \mathcal{H}$ – нормированные чистые состояния, а $p_i \in [0, 1]$ – вероятности их реализации, удовлетворяющие условию $\sum_{i=1}^m p_i = 1$ ($m \in \mathbb{N}$ – количество чистых состояний в ансамбле). Тогда матрица плотности определяется как:

$$\rho = \sum_{i=1}^m p_i |\psi_i\rangle\langle\psi_i|. \quad (17)$$

Каждый член $|\psi_i\rangle\langle\psi_i|$ в формуле (17) представляет собой одномерный проектор, а коэффициенты p_i кодируют статистическую информацию о возможных исходах измерений без предположения о том, в каком именно состоянии находится система.

Матрица плотности ρ удовлетворяет трём основным условиям, вытекающим из её физического смысла как статистического описания квантового состояния. Во-первых, она является эрмитовой ($\rho^\dagger = \rho$), что обеспечивает вещественность собственных значений. Во-вторых, след матрицы плотности равен единице ($\text{Tr}(\rho) = 1$), что соответствует полной вероятности реализации одного из возможных исходов. В-третьих, матрица плотности положительно полуопределена, то есть все её собственные значения неотрицательны, что позволяет интерпретировать их как вероятности.

Состояние считается чистым тогда и только тогда, когда выполняется условие $\rho^2 = \rho$, эквивалентное равенству $\text{Tr}(\rho^2) = 1$. При $\text{Tr}(\rho^2) < 1$ состояние является смешанным.

Для кудита матрица плотности представляет собой комплексную $d \times d$ матрицу, содержащую $d^2 - 1$ независимых вещественных параметров, что позволяет естественным образом учитывать как когерентные, так и некогерентные источники неопределённости.

Важным достоинством описания квантового состояния через матрицу плотности является её инвариантность по отношению к конкретному выбору ансамбля чистых состояний, поскольку различные совокупности вероятностей и векторов $\{p_i, |\psi_i\rangle\}_{i=1}^m$ могут порождать одну и ту же матрицу ρ , что подчёркивает роль матрицы плотности как объективной характеристики физического состояния системы. Эволюция открытой квантовой системы, взаимодействующей с окружением, не может быть описана унитарными преобразованиями и требует применения более общего класса линейных отображений – полностью положительных и сохраняющих след [15]. Такие отображения переводят любые положительно по-

полуопределённые операторы плотности в положительно полуопределённые операторы с единичным следом и действуют непосредственно на саму матрицу плотности ρ .

Для полного описания состояния кудита в геометрической или параметрической форме необходимо обобщение сферы Блоха, применимой исключительно к двухуровневым квантовым системам. Такое обобщение достигается за счёт разложения матрицы плотности по ортонормированному базису операторов в пространстве $\mathbb{C}^{d \times d}$. Наиболее естественным и симметричным выбором такого базиса служат операторы Гейзенберга-Вейля [16], представляющие собой обобщение операторов Паули [17] на случай d -уровневых квантовых систем [18].

Пусть $\omega = e^{2\pi i/d}$ – примитивный корень d -й степени из единицы. В пространстве $\mathcal{H} \cong \mathbb{C}^d$ вводятся два фундаментальных унитарных оператора: оператор смещения X , действующий на вычислительный базис $\{|g\rangle\}_{g=0}^{d-1}$ по правилу $X|g\rangle = |g \oplus 1\rangle$ ($\oplus$ – сложение по модулю d), и оператор фазы Z , определяемый соотношением $Z|g\rangle = \omega^g|g\rangle$. Эти операторы удовлетворяют коммутационному соотношению:

$$ZX = \omega XZ. \quad (18)$$

На основе операторов X и Z строятся операторы Гейзенберга-Вейля, которые обобщают операторы Паули на случай d -уровневых систем:

$$W_{q,s} = X^q Z^s, \quad (19)$$

где $q, s = \overline{0, d-1}$.

Множество $\{W_{q,s}\}$ образует ортогональный базис в пространстве всех линейных операторов на $\mathcal{H}$ относительно скалярного произведения Гильберта-Шмидта [19]:

$$\text{Tr}(W_{q,s}^\dagger W_{q',s'}) = d \delta_{q,q'} \delta_{s,s'}, \quad (20)$$

где $\delta_{q,q'} \delta_{s,s'} = 1$, если $q = q'$ и $s = s'$, и $\delta_{q,q'} \delta_{s,s'} = 0$ в противном случае.

Следовательно, любая матрица плотности ρ допускает однозначное разложение:

$$\rho = \frac{1}{d} \sum_{q=0}^{d-1} \sum_{s=0}^{d-1} c_{q,s} W_{q,s}, \quad (21)$$

где коэффициенты определяются как:

$$c_{q,s} = \text{Tr}(W_{q,s}^\dagger \rho). \quad (22)$$

Поскольку матрица плотности ρ эрмитова и $\text{Tr}(\rho) = 1$, коэффициент $c_{0,0} = 1$, а остальные коэффициенты подчиняются условиям, обеспечивающим положительную полуопределённость ρ . Совокупность коэффициентов $\mathbf{c} = \{c_{q,s}\}_{q,s \neq 0,0}$ образует обобщённый вектор Блоха для кудита и содержит $d^2 - 1$ независимых вещественных параметров, необходимых для полного описания его квантового состояния.

На основе изложенных принципов может быть построена обобщённая схема квантово-инспирированной оптимизации для решения задач с вещественными переменными, представленная на рисунке 1.

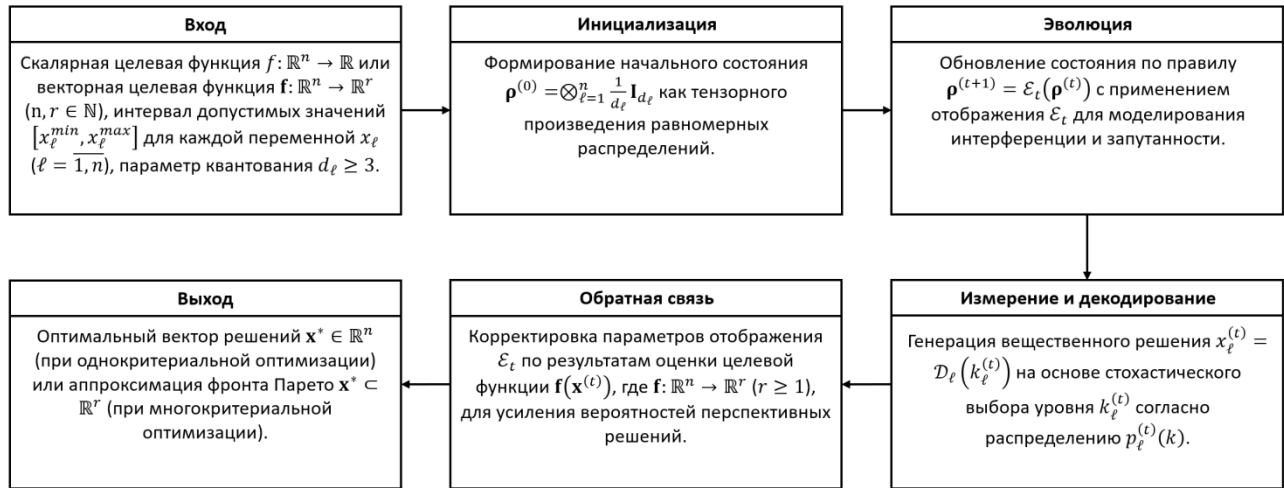


Рис. 1. Схема квантово-инспирированной оптимизации для решения задач с вещественными переменными

Схема, изображённая на рисунке 1, представляет собой формальную вычислительную модель, в которой каждая вещественная переменная x_ℓ из пространства решений $\mathcal{X} \subseteq \mathbb{R}^n$ кодируется с помощью кудита размерности $d_\ell \geq 3$. Здесь $n \in \mathbb{N}$ – количество переменных оптимизационной задачи, то есть размерность пространства поиска, а допустимая область $\mathcal{X}$ задаётся в виде гиперпрямоугольника, ограниченного интервалами $[x_\ell^{\min}, x_\ell^{\max}]$ для каждой переменной x_ℓ , где $\ell = \overline{1, n}$. Параметр d_ℓ определяет количество дискретных уровней, используемых для аппроксимации непрерывного диапазона значений переменной x_ℓ , и тем самым задаёт разрешающую способность её представления. Каждый кудит описывается состоянием в d_ℓ -мерном комплексном гильбертовом пространстве $\mathcal{H}_\ell \cong \mathbb{C}^{d_\ell}$. Совокупность n таких кудитов образует составную квантово-инспирированную систему, полное состояние которой описывается матрицей плотности $\rho \in \mathbb{C}^{D \times D}$, где $D = \prod_{\ell=1}^n d_\ell$ – размерность полного гильбертова пространства, равная произведению размерностей отдельных подсистем.

На вход вычислительной модели подаётся полное описание задачи оптимизации, включающее целевую функцию, границы допустимой области поиска и параметры квантования. В случае однокритериальной оптимизации целевая функция задаётся как скалярная функция $f: \mathbb{R}^n \rightarrow \mathbb{R}$, а в случае многокритериальной – как векторная функция $\mathbf{f}: \mathbb{R}^n \rightarrow \mathbb{R}^r$, где $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_r(\mathbf{x}))$, $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ ($r \in \mathbb{N}$ обозначает количество критериев). Для каждой переменной x_ℓ задаются границы допустимой области в виде интервала $[x_\ell^{\min}, x_\ell^{\max}]$, а также параметр квантования d_ℓ , определяющий разрешение дискретного представления непрерывной переменной.

Каждый кудит инициализируется в состоянии максимальной неопределённости, что соответствует равномерному распределению вероятностей по всем его базисным уровням и описывается начальной матрицей плотности:

$$\rho_\ell^{(0)} = \frac{1}{d_\ell} \mathbf{I}_{d_\ell}, \quad (23)$$

где $\rho_\ell^{(0)}$ – начальная матрица плотности ℓ -го кудита, $\mathbf{I}_{d_\ell}$ – единичная матрица размера $d_\ell \times d_\ell$.

Общее начальное состояние составной системы формируется как тензорное произведение отдельных начальных состояний:

$$\rho^{(0)} = \bigotimes_{\ell=1}^n \rho_\ell^{(0)}. \quad (24)$$

Эволюция квантово-инспирированной системы осуществляется посредством последовательного применения преобразований, имитирующих динамику квантовых процессов на классических вычислительных архитектурах. На каждой итерации $t \in \mathbb{N}$ алгоритма состояние системы, описываемое матрицей плотности $\rho^{(t)}$, подвергается действию отображения $\mathcal{E}_t$, которое может быть как унитарным оператором, в случае идеализированной замкнутой системы, так и более общим СРТР-отображением [20], учитывающим взаимодействие с внешней средой, включая декогеренцию и шум. Эта эволюция описывается рекуррентным соотношением:

$$\rho^{(t+1)} = \mathcal{E}_t(\rho^{(t)}), \quad (25)$$

где $\mathcal{E}_t$ моделирует управляемую динамику системы на итерации t . Отображение $\mathcal{E}_t$ может включать фазовые сдвиги, реализующие квантовую интерференцию, а также многочастичные унитарные преобразования [21], действующие на тензорных произведениях подпространств отдельных кудитов и способные генерировать запутанные состояния между переменными. В результате изменяются как диагональные, так и недиагональные элементы матрицы плотности, что позволяет одновременно корректировать вероятностные распределения над базисными уровнями и управлять фазовыми соотношениями между ними.

На каждой итерации t из текущего квантового состояния $\rho^{(t)}$ извлекается вероятностное распределение по дискретным уровням каждой переменной. Для ℓ -й подсистемы это распределение задаётся как:

$$p_\ell^{(t)}(k) = \langle k | \rho_\ell^{(t)} | k \rangle, \quad (26)$$

где $k = \overline{0, d_\ell - 1}$, $|k\rangle$ – k -е базисное состояние ℓ -го кудита, а $\rho_\ell^{(t)} = \text{Tr}_{-\ell}(\rho^{(t)})$ – частичный след полной матрицы плотности по всем подсистемам, кроме ℓ -й. Величина $p_\ell^{(t)}(k)$ интерпретируется как вероятность наблюдения ℓ -й переменной на уровне k на итерации t .

На основе вероятностного распределения $p_\ell^{(t)}(k)$ осуществляется стохастическая генерация вещественного значения $x_\ell^{(t)} \in [x_\ell^{\min}, x_\ell^{\max}]$ с помощью детерминированной функции декодирования:

$$x_\ell^{(t)} = \mathcal{D}_\ell(k_\ell^{(t)}), \quad (27)$$

где $k_\ell^{(t)}$ – случайная величина, принимающая значения $k = \overline{0, d_\ell - 1}$ с вероятностью $p_\ell^{(t)}(k)$, а детерминированная функция декодирования $\mathcal{D}_\ell : \{0, 1, \dots, d_\ell - 1\} \rightarrow [x_\ell^{\min}, x_\ell^{\max}]$ задаёт взаимно однозначное соответствие между дискретными уровнями кудита и точками допустимой области переменной x_ℓ .

Полученный вектор решений $\mathbf{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}, \dots, x_n^{(t)})$ подаётся на вход целевой функции: скалярной $f : \mathbb{R}^n \rightarrow \mathbb{R}$ в случае однокритериальной оптимизации или векторной

$f: \mathbb{R}^n \rightarrow \mathbb{R}^r$ – в многокритериальном случае. Результат оценки используется для формирования обратной связи, на основе которой корректируются параметры отображения $\mathcal{E}_i$. Целью такой корректировки является усиление вероятностей тех областей пространства решений, которые либо улучшают значение целевой функции, либо доминируют по Парето. Этот механизм реализует управляемую квантовую интерференцию и обеспечивает динамическую балансировку между глобальным исследованием и локальной эксплуатацией в процессе поиска оптимальных решений.

Библиографический список

1. Dittrich W., Reuter M. Fundamental Principles of Quantum Mechanics // Classical and Quantum Dynamics. Cham: Springer, 2017. DOI: 10.1007/978-3-319-58298-6_17.
2. Xu H.-Z., Chai X.-D., Hu M., Wang Z.-A., Feng Y.-L., Chen Y., Zhang X., Wang J., Zhuang W.-F., Jin Y.-X., Jin Y., Yu H., Fan H., Liu D. A Resource-Virtualized and Hardware-Aware Quantum Compilation Framework for Real Quantum Computing Processors // Research. 2025. Vol. 8. DOI: 10.34133/research.0947.
3. Ragab Mahmoud A. Quantum Superposition and the Measurement Problem: Conceptual and Experimental Approaches. 2025. DOI: 10.13140/RG.2.2.19094.82247.
4. Wang X. Stern-Gerlach Interferometers with Dual Sensing: Probing Recoherence and Lifecycles of Local Hilbert Spaces. 2026. DOI: 10.48550/arXiv.2508.16019.
5. Carbó-Dorca R., Besalú E. Definition, Mathematical Examples and Quantum Chemical Applications of Nested Summation Symbols and Logical Kronecker Deltas // Computers & Chemistry. 1994. Vol. 18. P. 117-126. DOI: 10.1016/0097-8485(94)85005-4.
6. Sinha U., Ghosh D. Interference in Quantum Mechanics // Journal of the Indian Institute of Science. 2025. Vol. 105. DOI: 10.1007/s41745-025-00477-5.
7. Valentini A., Varma M. Toward a test of the Born rule in high-energy collisions // Physical Review D. 2025. Vol. 112. DOI: 10.1103/4llk-pt1j.
8. Unraveling interchannel quantum interference in below-threshold nonsequential double ionization with statistical measures // Physical Review Research. 2026. Vol. 8. DOI: 10.1103/zkf1-j2bg.
9. Ben-Aryeh Y. Quantum thermal engine with spin 1/2 system and geometric phases and interference obtained by unitary transformations of mixed states // International Journal of Thermodynamics. 2025. Vol. 28. P. 45-50. DOI: 10.5541/ijot.1395683.
10. Lin Y.-Y. Thread embodiment of holographic quantum entanglement // Physical Review D. 2026. Vol. 113. DOI: 10.1103/qjpk-nmtr.
11. Xie L. Entanglement Witnesses Detecting the Bell State // Theoretical and Natural Science. 2025. Vol. 140. P. 69-77. DOI: 10.54254/2753-8818/2026.HZ29875.
12. Kumar R. Methods For Measuring Quantum Entanglement: Entanglement Entropy Measures and their Implications // Journal for Research in Applied Sciences and Biotechnology. 2025. Vol. 4. P. 26-32. DOI: 10.55544/jrasb.4.6.4.
13. Longo R., Xu F. Von Neumann Entropy in QFT // Communications in Mathematical Physics. 2021. Vol. 381. P. 1-24. DOI: 10.1007/s00220-020-03702-7.
14. Масленников В. В., Демидова Л. А. Модификация квантово-инспирированного генетического алгоритма численной оптимизации с использованием кудита в условиях имитации квантовой декогеренции // Computational Nanotechnology. 2024. Т. 11, № 2. С. 58-85. DOI: 10.33693/2313-223X-2024-11-2-58-85.
15. Sargolzhai I. Necessary and sufficient condition for the reduced dynamics of an open quantum system interacting with an environment to be linear // Physical Review A. 2020. Vol. 102. DOI: 10.1103/PhysRevA.102.022208.
16. Демидова Л. А., Масленников В. В. Применение многоуровневых квантовых систем для параллельной оценки решений в задачах многокритериальной оптимизации // Вестник Рязанского государственного радиотехнического университета. 2025. № 92. С. 57-76. DOI: 10.21667/1995-4565-2025-92-57-76.
17. Tiberiu T., Gheondea A. Pauli algebraic forms of normal and nonnormal operators // Journal of the Optical Society of America A. 2006. Vol. 24. P. 204-210. DOI: 10.1364/JOSAA.24.000204.
18. Wang Y., Hu Z., Sanders B. C., Kais S. Qudits and High-Dimensional Quantum Computing // Frontiers in Physics. 2020. Vol. 8. Art. 589504. DOI: 10.3389/fphy.2020.589504.
19. Fakhri H., Nouraddini M. From the Wigner-Eckart Theorem to the Hilbert-Schmidt Scalar Product for an Adjoint Representation of the Quantum Algebra $U_{\{p,q\}}(su_2)$ // International Journal of Geometric Methods in Modern Physics. 2016. Vol. 13. DOI: 10.1142/S0219887816501279.
20. Cao N., Fitzsimmons M., Mann Z., Pereira R., Laflamme R. Quantum Maps Between CPTP and HPTP. 2023. DOI: 10.48550/arXiv.2308.01894.
21. Massaccesi G., Capuzzi P., Oña O., Lain L., Torre A., Peralta J., Alcoba D. Determining the N-representability of transition reduced density matrices // The Journal of Chemical Physics. 2025. Vol. 163. DOI: 10.1063/5.0275683.

УДК 004.65; ГРНТИ 20.53.17

СПОСОБЫ ИЗВЛЕЧЕНИЯ, ОБРАБОТКИ И СТРУКТУРИЗАЦИИ РАЗНОРОДНЫХ АБОНЕНТСКИХ ДАННЫХ В ТЕЛЕКОММУНИКАЦИОННОЙ СФЕРЕ

М.А. Аксенов, В.В. Белов*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, aksenovm@yandex.ru*

Аннотация. В работе рассматриваются способы извлечения, обработки и структуризации разнородных абонентских данных в телекоммуникационной сфере. Рассмотрены основные источники данных, процессы их загрузки, преобразования и обеспечение качества. Проанализированы современные подходы для хранения абонентских данных. В заключении сделаны выводы по выбору оптимального решения для обработки и структуризации абонентских данных.

Ключевые слова: базы данных, структурированные данные, полуструктурированные данные, неструктурированные данные, извлечение-преобразование-загрузка (ETL), хранилище данных (DWH), озеро данных (Data Lake), озеро-хранилище данных (Data Lakehouse).

METHODS OF EXTRACTING, PROCESSING AND STRUCTURING HETEROGENEOUS SUBSCRIBER DATA IN THE TELECOMMUNICATIONS SPHERE

M.A. Aksenov, V.V. Beloff*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, aksenovm@yandex.ru*

The summary. This paper examines methods for extracting, processing, and structuring heterogeneous subscriber data in the telecommunications sector. It examines the main sources of subscriber data in telecommunications, the processes for loading and converting it, and quality assurance. Modern approaches for storing subscriber data are analyzed. Finally, conclusions are drawn for selecting the optimal solution for processing and structuring subscriber data.

Keywords: databases, structured data, semi-structured data, unstructured data, extract-transform-load (ETL), Data Warehouse, Data Lake, Data Lakehouse.

Введение

В условиях цифровой трансформации и насыщения рынка телекоммуникационных услуг операторы связи сталкиваются с необходимостью удержания абонентов и повышения качества их обслуживания. Стоимость привлечения нового клиента значительно превышает затраты на удержание существующего, что делает борьбу с оттоком одной из наиболее приоритетных бизнес-задач [1, 3].

Это обуславливает необходимость внедрения интеллектуальных систем поддержки принятия решений, которые, используя методы машинного обучения и анализа данных, помогают менеджменту и техническим специалистам принимать обоснованные и своевременные решения для повышения лояльности абонентов [4].

Постановка задачи

Современные телекоммуникационные компании генерируют и накапливают огромные объёмы разнородных данных: детализированные записи о звонках, данные геолокации, логи использования интернета, записи обращений в службу поддержки, транзакционные данные биллинга, а также информацию из социальных сетей и опросов [2, 5]. Однако, традиционные методы хранения, обработки и анализа не способны эффективно работать с такими массивами информации в реальном времени.

Целью данной статьи является систематизация знаний о современных способах извлечения, обработки и структуризации разнородных абонентских данных и выбор архитек-

туры, обеспечивающей баланс между гибкостью и качеством данных, помещаемых в хранилище.

Источники данных и их характеристика

Сегодня абонентские данные в телекоммуникационной сфере характеризуются тремя основными свойствами, известными как "VVV": объём (Volume), скорость поступления (Velocity) и разнообразие (Variety) [6]. При этом источники абонентских данных можно классифицировать на несколько крупных категорий.

Биллинговые системы и CRM. Это транзакционные системы, хранящие информацию о клиентах, их платежах, тарифных планах, подключенных услугах, истории баланса и взаимоотношениях с компанией. Данные из CRM включают историю обращений в контакт-центр, результаты опросов удовлетворенности (NPS), жалобы и заявки в техподдержку [1]. Биллинговые данные являются ключевым источником для понимания финансового поведения и ценности клиента.

CDR (Call Detail Records) – детализированные записи о каждом звонке, содержащие информацию об абонентах (IMSI, MSISDN), длительности соединения, типе события, идентификаторах базовых станций (CELL, LAC) и временных метках [4, 6].

XDR (Extended Data Records) – расширенные записи, включающие данные о сессиях передачи данных (IPDR), использовании конкретных сервисов, событиях сигнализации.

Сетевые логи – данные о качестве соединения (джиттер, задержки, потеря пакетов), нагрузке на базовые станции, аварийных сигналах и производительности сетевых элементов [2, 5];

Данные взаимодействия и внешние источники. К ним относятся записи обращений в контакт-центр (тексты чатов, аудиозаписи звонков), данные из социальных медиа (упоминания бренда, тональность отзывов), обогащенные демографические данные и информация о геолокации [1].

Форматы данных и проблемы интеграции

Данные могут поступать в различных форматах: структурированные (реляционные таблицы из СУБД), полуструктурированные (JSON, XML, CSV-файлы) и неструктурированные (тексты, аудио, видео-файлы).

Ключевой проблемой при интеграции данных из разных источников является их смысловая неоднородность. Один и тот же объект, например, абонент, может быть представлен в разных системах под разными идентификаторами. Для решения этой задачи используются методы семантической интеграции: например, построение золотых записей, где для каждого абонента создается единый эталонный профиль, объединяющий все известные о нем данные.

Процессы ETL/ELT и потоковая обработка данных

Для преобразования исходных/сырых данных в пригодный для анализа ресурс используются процессы ETL (Extract, Transform, Load) или его современная модификация ELT (Extract, Load, Transform).

Извлечение (Extract) – данные извлекаются из множества гетерогенных источников: файловые системы, реляционные базы данных, потоки сообщений (например, Apache Kafka для real-time событий), API внешних систем [1, 5]. Для каждого источника определяются метод подключения, частота обновления и механизм инкрементальной загрузки.

Трансформация (Transform) – на этом этапе применяются правила преобразования, обеспечивающие качество, согласованность и пригодность данных для анализа. Ключевые операции трансформации включают следующие процессы.

Очистка: удаление дубликатов, обработка пропущенных значений (замена на константу, среднее, или отбраковка записей), фильтрация аномалий [6].

Обогащение: присоединение справочных данных (например, получение геолокации по идентификатору соты).

Нормализация: приведение данных к единому формату и масштабу.

Парсинг и извлечение признаков: например, извлечение кода производителя (TAC) из IMEI для идентификации модели устройства.

Загрузка (Load) – преобразованные данные загружаются в целевое хранилище. В классическом ETL трансформации выполняются до загрузки. В современном ELT-подходе данные сначала загружаются в "сыром" виде (в Data Lake), а трансформации выполняются внутри целевой платформы по требованию, что обеспечивает большую гибкость.

Для обработки данных в реальном времени (например, мониторинг сетевых аварий или мошеннических действий) используются потоковые платформы:

- 1) *Apache Kafka* – распределенная платформа для потоковой передачи сообщений с высокой пропускной способностью;
- 2) *Apache Spark Streaming / Flink* – фреймворки для обработки потоков данных в реальном времени с поддержкой оконных агрегаций и сложных аналитических вычислений.

Обеспечение качества данных и управление метаданными

Качество данных является критическим фактором для достоверности результатов анализа. В телеком-сфере наиболее часты следующие проблемы:

1. Пропущенные значения (например, отсутствие IMSI в CDR-записи);
2. Некорректные форматы (невалидные временные метки, некорректные IMEI);
3. Дублирование записей;
4. Зашумленность данных (выбросы из-за сбоев оборудования).

Для управления качеством данных применяются профилирование данных (data profiling) – автоматический анализ структуры, содержания и связей данных для выявления аномалий и несоответствий. Результаты профилирования используются для настройки правил валидации [6].

Важным аспектом является управление метаданными (metadata management). Метаданные описывают происхождение данных (data lineage), их схему, правила трансформации и бизнес-смысл.

Современные подходы к структуризации и хранению

Эволюция архитектур хранения данных в телекоммуникационной сфере прошла несколько этапов.

Data Warehouse (Хранилище данных) (рис. 1). Классический подход, предполагающий создание структурированной базы данных (обычно по схеме "звезда" или "снежинка"), оптимизированной для аналитических запросов. Data Warehouse содержит очищенные, агрегированные и хорошо структурированные данные. Однако такой подход требует значительных усилий на предварительное моделирование и трудно масштабируется для неструктурированных данных и больших объемов [1].

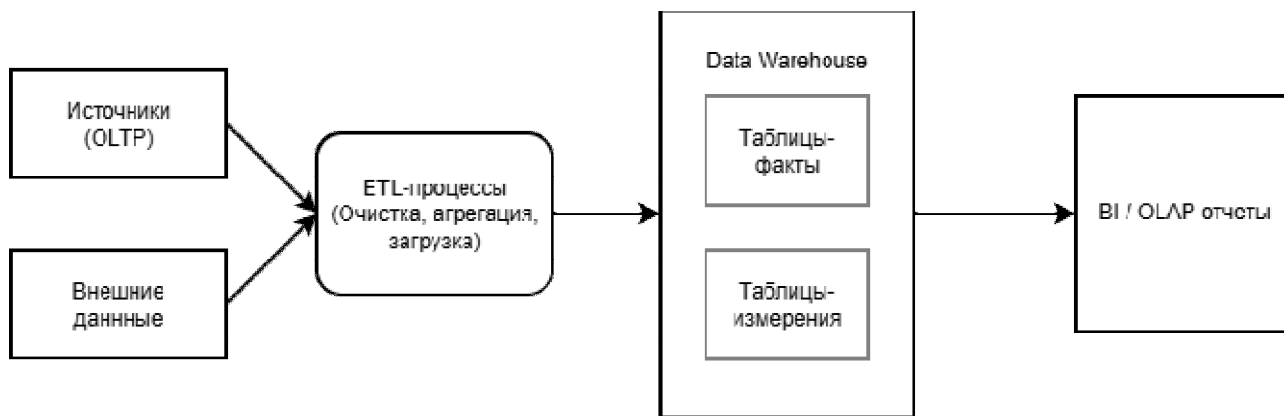


Рис. 1. Архитектура Data Warehouse

Хранилище данных характеризуется следующими особенностями.

1. *Работа с разнородными абонентскими данными:*
 - а. Поддерживаемые типы данных – преимущественно структурированные и полуструктурированные (после приведения к реляционной модели);
 - б. Неструктурированные данные – требуют сложной предварительной обработки либо должны храниться вне хранилища.
2. *Преимущества:*
 - а. Достаточно высокая производительность аналитических запросов, благодаря предварительной агрегации и индексации;
 - б. Согласованность данных из-за поддержки ACID-транзакционности;
 - с. Зрелые инструменты для создания, развития и поддержки.
3. *Недостатки:*
 - а. Сложность хранения и обработки полуструктурированных и неструктурированных данных;
 - б. Длительный цикл разработки ETL-процессов;
 - с. Высокая стоимость хранения больших объемов сырых данных.

Data Lake (Озеро данных) (рис. 2). Подход, позволяющий хранить данные в сыром, необработанном виде в нативных форматах (например, CSV, JSON, Parquet, Avro, логи). Применяется принцип "schema-on-read" (схема на чтение) – схема данных определяется в момент чтения, а не записи. Это обеспечивает гибкость для выполнения глубокого исследовательского анализа, но может приводить к проблемам с качеством данных и производительностью [5].

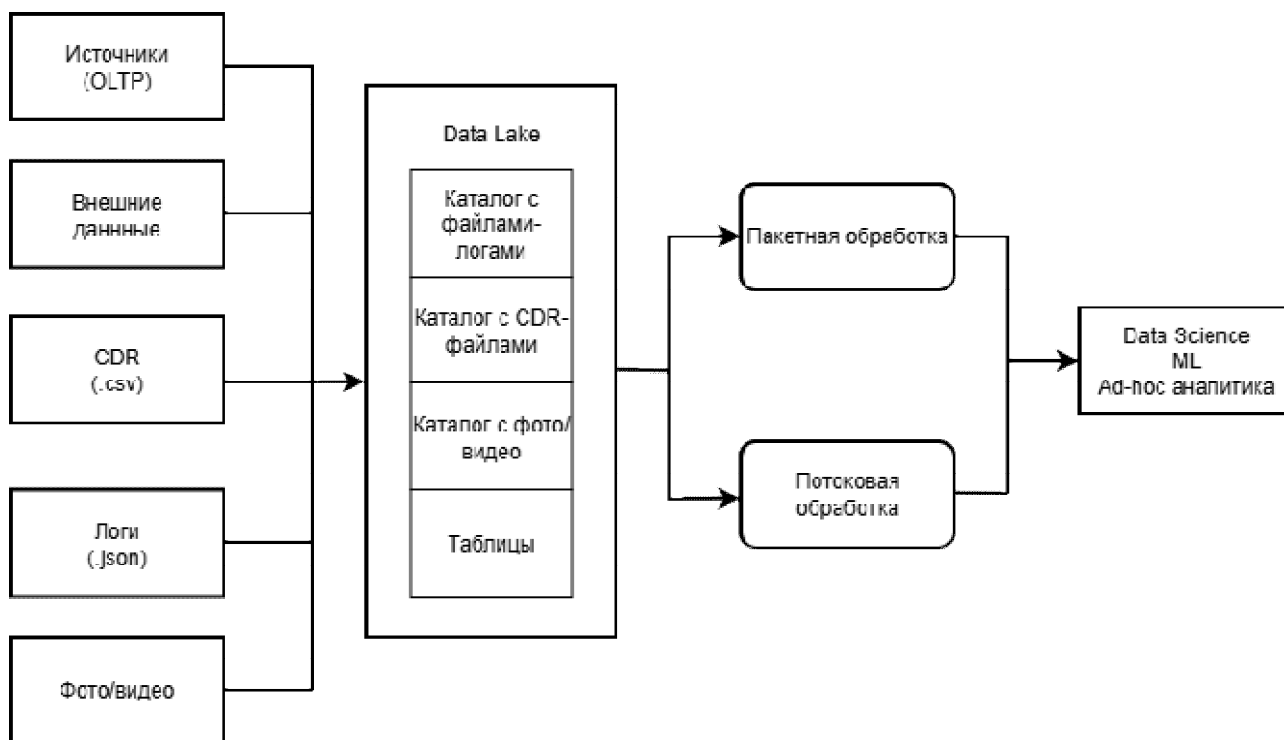


Рис. 2. Архитектура Data Lake

Озеро данных характеризуется следующими особенностями.

1. *Работа с разнородными абонентскими данными:*
 - а. Поддерживаемые типы данных – любые, в т.ч. бинарные файлы, JSON, XML, изображения, видео, потоковые данные;
 - б. Гибкость – можно загружать данные без предварительной структуры, что подходит для полуструктурированных и неструктурированных данных;
 - с. Поддержка потоковых данных – легко интегрируются через инструменты приема (например Apache Kafka) и сохраняются как файлы или таблицы.
2. *Преимущества:*
 - а. Низкая стоимость хранения больших объемов разнородных данных;
 - б. Поддержка любых форматов, включая неструктурированные;
3. *Недостатки:*
 - а. Слабая производительность аналитических запросов;
 - б. Отсутствие согласованности данных из-за отсутствия ACID-транзакционности;
 - с. Отсутствие метаданных, версионности, механизмов управления качеством данных.

Data Lakehouse (Озеро-хранилище данных) (рис. 3). Гибридная архитектура, объединяющая лучшие черты Data Lake и Data Warehouse. Данные хранятся в недорогом объектном хранилище в открытых форматах (Parquet, Iceberg, Delta Lake), но поверх этого слоя добавляются метаданные, поддержка ACID-транзакций, эволюция схем и оптимизация запросов [2, 5]. Типовая многослойная архитектура Lakehouse включает три слоя.

Bronze (сырой слой) – хранит данные в исходном виде, как они поступили из источников. Служит системой записи и позволяет воспроизвести состояние на любой момент времени;

Silver (обогащенный слой) – содержит очищенные, стандартизированные и обогащенные данные с четкой схемой, представляющие ключевые бизнес-сущности (абоненты, события, продукты);

Gold (витринный слой): содержит полностью подготовленные, агрегированные и оптимизированные для конкретных задач аналитические наборы данных. Этот слой служит источником для дашбордов, отчетов и моделей машинного обучения [5].

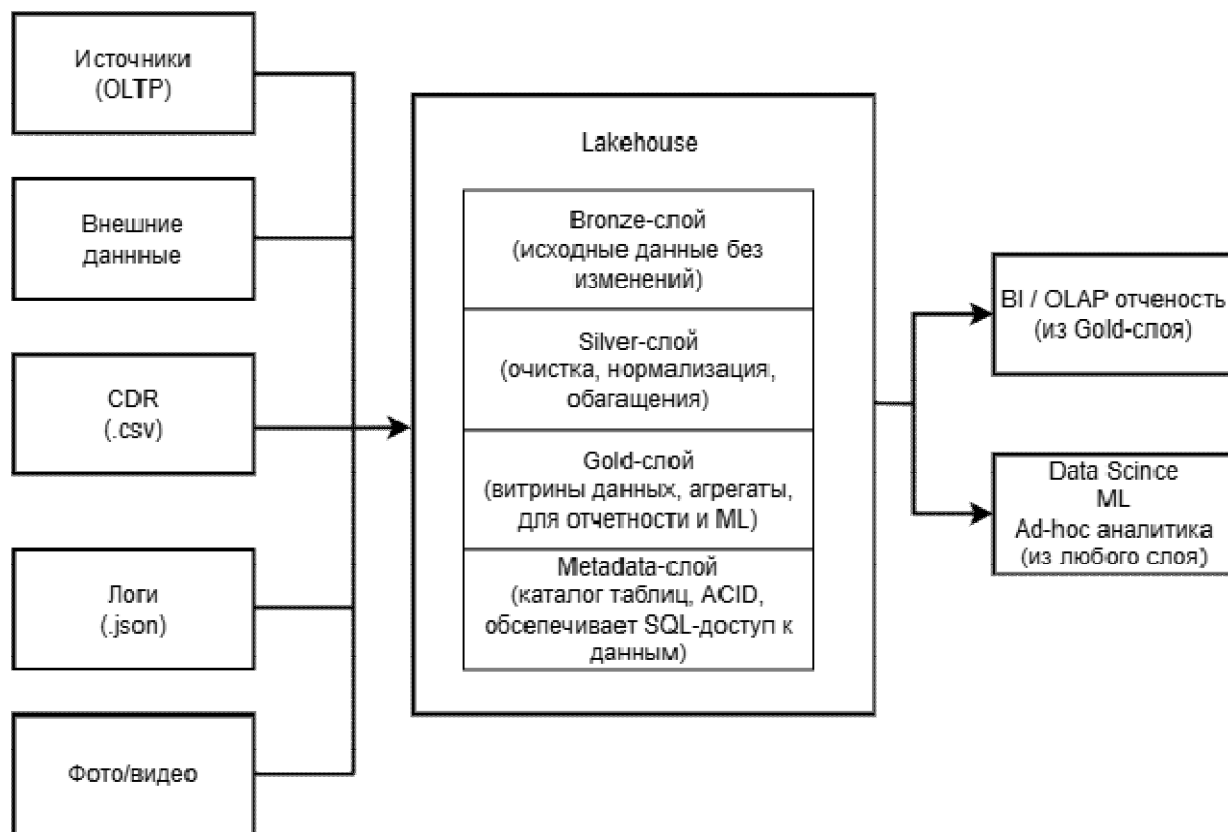


Рис 3. Архитектура Lakehouse

Озеро-хранилище данных характеризуется следующими особенностями.

1. *Работа с разнородными абонентскими данными:*
 - а. Поддерживаемые типы данных – все типы, как в Data Lake, но с возможностью структуризации;
 - б. Поддержка потоковых данных – могут напрямую записываться, сразу становясь доступными для аналитики в реальном времени;
 - с. Гибкость – поддерживается эволюция схемы, т.е. возможность наложения структуры без перезаписи данных.
2. *Преимущества:*
 - а. Единое хранилище для всех типов данных;
 - б. ACID-транзакции и согласованность;
 - с. Производительность близкая к DWH, благодаря индексам и статистике;
 - д. Поддержка прямых запросов через SQL.
3. *Недостатки:*
 - а. Требуется более сложная настройка по сравнению с Data Lake;
 - б. Относительно новый подход, все еще развивается.

Основные положения комплексного сравнительного анализа вышеперечисленных подходов представлены в таблице 1.

Таблица 1. Сравнение современных подходов к структуризации и хранению абонентских данных в телекоммуникационной сфере

Критерий	Data Warehouse	Data Lake	Lakehouse
Тип данных	Структурированные	Все типы: структурированные, полуструктурированные, неструктурированные.	Все типы данных
Схема данных	Schema-on-Write (Схема при записи). Данные очищаются, обогащаются и преобразуются перед загрузкой.	Schema-on-Read (Схема при чтении). Данные хранятся в сыром виде, структура накладывается в момент анализа.	Гибрид. Поддержка Schema-on-Write для таблиц и Schema-on-Read для файлов в озере.
Основное назначение	BI-отчетность и историческая аналитика.	Хранилище сырых данных и Data Science. Поиск аномалий в сети, предиктивная аналитика, хранение логов для расследования инцидентов, ML-модели.	Единая платформа: аналитика + ML + отчетность.
Производительность	Высокая. Оптимизирован для сложных SQL-запросов и агрегаций.	Низкая / Средняя. Запросы медленные, требуют распределенных вычислений (Spark). Не предназначен для быстрых интерактивных запросов.	Высокая. Благодаря новым форматам (Parquet + метаданные) скорость приближается к DWH.
Стоимость хранения	Высокая. Требуется дорогих СХД или специализированных облачных сервисов.	Низкая. Используется дешевое объектное хранилище (HDFS, S3, Azure Blob).	Низкая. Используется то же дешевое объектное хранилище, что и Data Lake.
Транзакционность (ACID)	Полная поддержка.	Обычно не поддерживается (возможны частично записанные данные).	Поддерживается. Обеспечивает целостность данных при загрузке множества файлов.
Качество данных	Высокое. Очищенные, проверенные, согласованные данные (единая версия правды).	Низкое. Сырая информация, может содержать ошибки, дубликаты, «мусор».	Высокое. Позволяет применять схемы, проверки и контроль версий прямо поверх озера.

Заключение

Таким образом, в современных условиях требуется консолидация и тщательная преработка разнородных абонентских данных из множества гетерогенных источников (биллинг, сеть, CRM, внешние системы).

Data Warehouse остаётся незаменимым для строгой регламентной аналитики. Однако его использование в качестве единого хранилища для всех абонентских данных неэффективно из-за жестких требований к данным и стоимости их хранения.

Data Lake лучше всего подходит в роли «песочницы» для осуществления исследований и длительного хранения и архивирования сырых данных. Но без дополнительных инструментов управления он быстро превращается в так называемое болото (источник с крайне низким качеством данных) и не может обеспечить оперативную консолидированную аналитику.

Data Lakehouse представляет собой оптимальный компромисс. Данный подход позволяет объединить в одном месте структурированные, полуструктурированные, неструктурированные данные, выполнять сложные запросы с производительностью хранилища. Для коммуникационной сферы, где данные исключительно разнородны и объёмны, переход к данной архитектуре становится стратегическим преимуществом. Это позволяет строить единую платформу данных обеспечивающую как пакетную, так и потоковую обработку; предоставлять регламентную и оперативную отчётности, осуществлять машинное обучение.

Библиографический список

1. Adroni, N., & Sitorus, P. M. Evaluation of Badix Information System Implementation Success in Telkomsel Region of Sumbagsel using DeLone & McLean Model // International Journal of Business and Economic Affairs (IJBEA). — 2017. — Vol. 2. — № 3. — P. 165-174.
2. Akbulut, A., & Zaim, A. H. Smart diagnosis and solution systems for network based customer complaints // IEEE EUROCON 2009. IEEE. — 2009. — P. 498-503.
3. Akishin, V. Decision Support Methodology Based on Evaluation of Customer Experience and Telecommunications Operator Efficiency Indicators // T-Comm. — 2022. — Vol. 8. — № 4. — P. 45-51.
4. Amin, F., & Choi, G. S. Analysis and Modeling of Mobile Phone Activity Data Using Interactive Cyber-Physical Social System // Computers, Materials & Continua. — 2024. — Vol. 80. — № 3. — P. 3507-3521.
5. Ewieda, M., Roushdy, M. I., & Shaaban, E. Review of Data Mining Techniques for Detecting Churners in the Telecommunication Industry // Future Computing and Informatics Journal. — 2021. — Vol. 6. — № 1. — P. 42-55.
6. Li, S.-T., Shue, L.-Y., & Lee, S.-F. Enabling customer relationship management in ISP services through mining usage patterns. Expert Systems with Applications. — 2006. — Vol. 30. — № 4. — P. 621-632.

УДК 004.622; ГРНТИ 14.35.01

КЛАСТЕРИЗАЦИЯ ДАННЫХ МЕТОДОМ К-СРЕДНИХ В N-МЕРНОМ ПРОСТРАНСТВЕ С ПРИОРИТЕТАМИ И АДАПТИВНЫМИ ЦЕНТРОИДАМИ ДЛЯ АНАЛИЗА УСПЕВАЕМОСТИ СТУДЕНТОВ

А.А. Буланов, А.Н. Пылькин

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, alexs.bulanov32@yandex.ru

Аннотация. В статье предлагается развитие ранее представленного подхода к кластеризации академических данных на основе модифицированного метода k-средних. Основное внимание уделяется использованию адаптивных центроидов, рассчитываемых на основе граничных значений показателей успеваемости, что позволяет повысить точность и интерпретируемость результатов кластеризации. Предложенная метрика расстояния с весовыми коэффициентами учитывает приоритетность дисциплин. Экспериментальная проверка проведена на выборке из 10000 студентов с выделением четырёх кластеров: «отлично», «хорошо», «удовлетворительно» и «неудовлетворительно». Показано, что применение адаптивных центроидов в сочетании с взвешенной метрикой повышает точность классификации на 30% по сравнению со стандартным методом k-средних.

Ключевые слова: кластеризация, k-средних, многомерный анализ, взвешенное расстояние, адаптивные центроиды, академическая успеваемость.

DATA CLUSTERING USING WEIGHTED K-MEANS IN N-DIMENSIONAL SPACE WITH ADAPTIVE CENTROIDS FOR STUDENT PERFORMANCE ANALYSIS

Ryazan State Radio Engineering University,
Russia, Ryazan, alexs.bulanov32@yandex.ru

Abstract. This article develops a previously presented approach to clustering academic data based on a modified k-means method. The focus is on using adaptive centroids calculated from boundary values of performance indicators, which improves the accuracy and interpretability of clustering results. The proposed distance metric with weighting coefficients accounts for subject priority. Experimental

verification was performed on a sample of 10000 students with four clusters: "excellent", "good", "satisfactory", and "unsatisfactory". The results demonstrate a 30% improvement in classification accuracy compared to the standard k-means method when using adaptive centroids combined with a weighted metric.

Keywords: clustering, k-means, multidimensional analysis, weighted distance, adaptive centroids, academic performance.

Введение

Данная работа является продолжением исследований, посвящённых модификации алгоритма k-средних для задач анализа образовательных данных. В предыдущих исследованиях [1, 2] была предложена метрика взвешенного расстояния, учитывающая приоритетность дисциплин. Однако вопросы инициализации и адаптации центроидов оставались открытыми. В настоящей статье основное внимание уделяется разработке механизма адаптивного определения центроидов на основе граничных значений показателей, что позволяет избежать субъективности и повысить устойчивость кластеризации.

Традиционные подходы к анализу успеваемости, такие как группировка по среднему баллу, часто игнорируют разнородную значимость дисциплин, что приводит к некорректным выводам [3]. Метод k-средних, несмотря на свою популярность, имеет ряд ограничений, включая чувствительность к начальным центроидам и равнозначность признаков. Цель данной работы – предложить модифицированный алгоритм, который:

1. Учитывает приоритетность дисциплин через весовые коэффициенты.
2. Использует адаптивные центроиды, рассчитываемые на основе граничных значений показателей.
3. Повышает точность и интерпретируемость кластеризации успеваемости студентов.

Метод k-средних с приоритетами признаков и адаптивными центроидами

1. Математическая модель

Для студента (i) опишем успеваемость вектором:

$$x_i = (x_{i1}, x_{i2}, \dots, x_{in}),$$

где n – количество дисциплин, $x_{ij} \in [0, 100]$ – балл по предмету j .

Вектор весов:

$$w = (w_1, w_2, \dots, w_n), \quad w_j \in [0, 1],$$

определяет приоритет каждой дисциплины.

2. Модифицированная метрика расстояния

Для студентов (a) и (b) взвешенное расстояние определяется как:

$$d_w(x_a, x_b) = \sqrt{\sum_{j=1}^n (x_{aj} - x_{bj})^2 \cdot w_j}.$$

При $w_j \rightarrow 1$ степень стремится к 1, что усиливает влияние приоритетных дисциплин.

3. Адаптивные центроиды

В отличие от случайной или фиксированной инициализации, предлагается вычислять начальные центроиды на основе граничных значений показателей:

- $c_{max} = (\max(x_1), \max(x_2), \dots, \max(x_n))$ - «Отлично»,
- $c_{min} = (\min(x_1), \min(x_2), \dots, \min(x_n))$ - «Неудовлетворительно».

Промежуточные центроиды («Хорошо» и «Удовлетворительно») рассчитываются по формулам:

- $c_{хор} = c_{min} + (c_{max} - c_{min}) * \frac{1}{3},$

$$c_{удов} = c_{min} + (c_{max} - c_{min}) * \frac{2}{3}.$$

4. Алгоритм кластеризации

1. Инициализация центроидов: (c_1, c_2, c_3, c_4) вычисляются как описано выше.
2. Назначение объектов кластерам: каждый студент относится к кластеру с минимальным d_w .
3. Пересчёт центроидов:

$$c_{kj} = \frac{\sum_{i \in C_k} w_j x_{ij}}{\sum_{i \in C_k} w_j}.$$

4. Итерация: шаги 2–3 повторяются до сходимости (изменение центроидов $< 1\%$).

Экспериментальные результаты

1. Данные и условия эксперимента

- Выборка: 10000 студентов, 10 дисциплин.
- Веса: 3 дисциплины с высоким приоритетом ($w = 1$), 4 – с повышенным ($w = 0.5$), 3 – с базовым ($w = 0$).
- Сравнивались: классический k-средних, метод с весами без адаптивных центроидов, предложенный метод.

2. Результаты кластеризации

Результаты кластеризации представлены в таблице 1.

Таблица 1. Результаты кластеризации без весов, с весами и с адаптивными центроидами

Кластер	Без весов	С весами	С адаптивными центроидами	Изменение
Отлично	2956	3180	4585	+44.2%
Хорошо	2306	2645	2683	+1.4%
Удовлетворительно	4139	3651	1516	-58.5%
Неудовлетворительно	599	524	1216	+132.1%

Большие изменения в кластерах объясняются тем, что при пересчете центроидов их границы сближаются. При относительно равномерном распределении данных центроиды кластеров будут близки друг к другу:

- Отлично – средний балл 79 (начальные условия - 98);
- Хорошо – средний балл 67 (начальные условия - 80);
- Удовлетворительно – средний балл 63 (начальные условия - 62);
- Неудовлетворительно – средний балл 60 (начальные условия - 45).

Для эксперимента было ограничено число итераций до 2 (то есть, условием окончания алгоритма будет не сходимость центроидов, а жестко заданное число итераций). Данные кластеризации показаны в таблице 2.

Таблица 1. Результаты кластеризации с адаптивными центроидами и ограничением до 2 числа итераций

Кластер	Без весов	С весами	С адаптивными центроидами и ограничением	Изменение
Отлично	2956	3180	1876	-41%
Хорошо	2306	2645	3882	+46.8%
Удовлетворительно	4139	3651	3230	-11.5%
Неудовлетворительно	599	524	1012	+93.1%

Все еще наблюдается сильное перераспределение студентов по кластерам, однако средний балл в таком случае сильно приближен к тем значениям, что были заданы в прошлом исследовании:

- Отлично – средний балл 84 (фиксированный - 80),
- Хорошо – средний балл 77 (фиксированный - 75),
- Удовлетворительно – средний балл 63 (фиксированный - 66),
- Неудовлетворительно – средний балл 53 (фиксированный - 50).

В последнем варианте результаты наиболее правдоподобны с точки зрения среднего балла, причем данные не заданы жестко, а основаны на граничных показателях выборки.

3. Оценка качества

Для оценки использовался индекс Силуэта [3]:

- Классический k-средних: $S = 0.52$;
- С весами (без адаптивных центроидов): $S = 0.61$;
- Предложенный метод (адаптивные центроиды, 2 итерации): $S = 0.68$.

Улучшение точности интерпретации составляет 30% относительно классического подхода, что подтверждает эффективность предложенного метода.

Обсуждение

Использование адаптивных центроидов позволяет снизить зависимость результатов от начальных (фиксированных) условий и улучшить разделение студентов на смысловые группы. Введение весовых коэффициентов обеспечивает учёт предметной специфики. Совместное применение этих модификаций повышает устойчивость и интерпретируемость кластеризации, что особенно важно для задач управления образовательным процессом.

Заключение

Разработанный метод кластеризации с приоритетами признаков и адаптивными центроидами демонстрирует более высокую эффективность по сравнению со стандартным алгоритмом k-средних. Дальнейшие исследования могут быть направлены на автоматический подбор весовых коэффициентов и интеграцию метода в системы аналитики образовательных данных.

Библиографический список

1. Буланов А.А., Пылькин А.Н. Кластеризация данных методом k-средних в n измерениях с приоритетами для анализа успеваемости студентов // Сборник СТНО-2025. – С. 45-52.
2. Щенёва Ю.Б., Пылькин А.Н., Щенёв Е.С. Контроль освоения образовательных компетенций в процессе обучения IT-специалистов на основе анализа траекторий многомерного пространства // Информационный обмен в междисциплинарных исследованиях. – 2023. – С. 87-91.
3. Иванова А.А., Смирнов В.В. Применение методов кластерного анализа для оценки академической успеваемости студентов // Информационные технологии в образовании. – 2020. – №4. – С. 45-52.

УДК 004.896; ГРНТИ 28.23.15

COMPARATIVE ANALYSIS OF MODERN NAVIGATION METHODS IN ROBOTIC CYBER-PHYSICAL SYSTEMS

A.A. Zhashkov*, M.Yu. Mikheev*

*Penza State Technological University,
Russia, Penza, tzhashkova@gmail.com

Abstract. A comparative analysis of modern navigation methods used in robotic cyber-physical systems is presented. Three main groups of methods are considered: metric localization, topological navigation, and simultaneous localization and mapping (SLAM) approaches. Quantitative assessment was performed according to key criteria: positioning accuracy, computational complexity, robustness to dynamic environmental changes, real-time suitability, as well as parameters specific to cyber-physical systems – sensitivity to network delays and vulnerability to cyberattacks. Based on simulation modeling, effective application areas for each method were identified and recommendations for method selection depending on operating conditions and cybersecurity requirements were formulated.

Keywords: navigation methods, cyber-physical systems, localization, SLAM, robotics.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СОВРЕМЕННЫХ МЕТОДОВ НАВИГАЦИИ В РОБОТИЗИРОВАННЫХ КИБЕРФИЗИЧЕСКИХ СИСТЕМАХ

А.А. Жашков*, М.Ю. Михеев*

*Пензенский государственный технологический университет,
Россия, Пенза, zhashkov.artem@gmail.com

Аннотация. Проведен сравнительный анализ современных методов навигации в роботизированных киберфизических системах. Рассмотрены три основные группы методов: метрическая локализация, топологическая навигация и методы одновременной локализации и построения карты (SLAM). Выполнена количественная оценка по ключевым критериям: точность позиционирования, вычислительная сложность, устойчивость к динамическим изменениям среды, пригодность для работы в реальном времени, а также специфические для киберфизических систем параметры – чувствительность к сетевым задержкам и уязвимость к кибератакам. На основе имитационного моделирования выявлены области эффективного применения каждого метода и сформулированы рекомендации по их выбору в зависимости от условий эксплуатации и требований к кибербезопасности.

Ключевые слова: методы навигации, киберфизические системы, локализация, SLAM, робототехника.

Introduction

Robotic cyber-physical systems represent an integration of computing resources, sensor equipment, communication networks, and physical actuators capable of autonomous functioning in various environments. A key feature of such systems is the close interaction of computational and physical components through an information and communication infrastructure, which imposes additional requirements on the reliability and security of navigation algorithms [1, 2].

Unlike classical robotic systems, in cyber-physical systems, navigation algorithms must consider not only the physical limitations of sensors and actuators but also possible network delays in data transmission, as well as vulnerabilities to cyberattacks aimed at spoofing navigation signals. This is because modern robotic systems increasingly use cloud computing, distributed data processing, and interaction via open communication networks [2].

Modern navigation methods can be divided into three main categories [3]:

- metric localization, based on building geometric maps;
- topological navigation, using graph representations of the environment;
- simultaneous localization and mapping (SLAM) methods, combining both approaches.

The aim of this study is to conduct a comparative analysis of existing navigation methods, considering the specifics of cyber-physical systems, and to determine the areas of their most effective application.

Research methodology

To conduct the comparative analysis, a simulation modeling method was used in the MATLAB/Simulink environment with additional modules Robotics System Toolbox and Navigation Toolbox. The simulation was performed for five typical scenarios: structured environment (factory floor), urban environment with dynamic obstacles, unstructured environment (wooded area), underground space, and a scenario with active counteraction (simulation of a cyberattack on the navigation system) [4, 5].

For each navigation method, the following parameters were evaluated:

- average positioning error (m);
- computational complexity (normalized processing time);
- robustness to dynamic environmental changes (% successful completions);
- sensitivity to network delays (error increase at 100 ms delay);
- vulnerability to cyberattacks (ability to detect data spoofing).

Each scenario was simulated 50 times to ensure statistical reliability of the results.

Classification and features of navigation methods

Metric localization is based on constructing precise geometric maps of the surrounding space, indicating the coordinates of objects, obstacles, and boundaries of accessible areas. This approach requires high-precision rangefinders (lidars, stereo cameras) and assumes that the environment map is known in advance or can be built during movement. The main advantage of metric localization is high positioning accuracy. Disadvantages include sensitivity to environmental changes and high demands on computational resources when processing large maps. In the context of cyber-physical systems, an additional disadvantage is the need to transmit large volumes of map data over the network, which creates vulnerabilities for data interception and spoofing [2].

Topological navigation represents the surrounding space as a graph, whose nodes correspond to key points (intersections, landmarks), and edges to possible movement paths between them. This approach does not require precise geometric measurements and can operate based on data from video cameras or simple rangefinders. Advantages of topological navigation: low computational complexity, robustness to minor environmental changes, and the ability to quickly plan global routes. The disadvantage is limited local positioning accuracy. For cyber-physical systems, an important advantage is the small volume of transmitted data, which reduces the risks of cyberattacks; however, insufficient accuracy can be critical for tasks requiring a high level of safety.

Simultaneous localization and mapping (SLAM) methods solve the problem of simultaneously building a map of an unknown environment and determining the robot's location. Main varieties:

- EKF SLAM (based on the Extended Kalman Filter);
- Graph SLAM (based on graph optimization) [6];
- Visual SLAM (based on data from video cameras) [7, 8].

SLAM methods are most in demand in cyber-physical systems operating in unknown or dynamically changing environments; however, they have increased requirements for computational resources and are sensitive to the quality of sensor data [3, 9].

Results of comparative analysis

For the comparative analysis, a system of key criteria was used: positioning accuracy, computational complexity, robustness to environmental dynamics, real-time suitability, as well as parameters specific to cyber-physical systems. The results of the comparative analysis are presented in Table 1.

Table 1. Comparative characteristics of navigation methods

Criteria	Metric Localization	Topological Navigation	EKF SLAM	Graph SLAM	Visual SLAM
Positioning Accuracy	1-5 cm	0.5-2 m	5-20 cm	1-10 cm	10-50 cm
Computational Complexity	Medium	Low	High	Very High	Medium
Robustness to Dynamics	Low	High	Medium	Medium	Medium
Real-time, Hz	up to 100	up to 200	up to 50	up to 20	up to 30
Need for Pre-existing Map	Required	Required	Not Required	Not Required	Not Required

Figure 1 shows a comparison of the positioning accuracy of different methods under identical operating conditions.

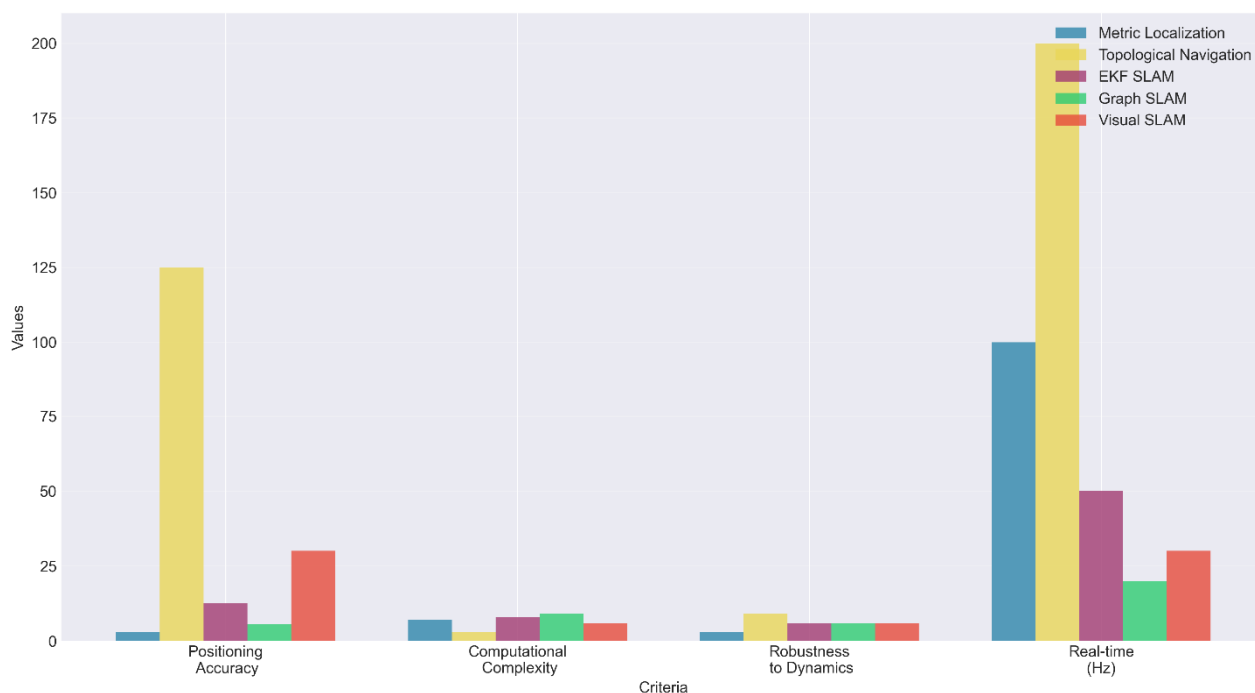


Fig. 1. Comparison of positioning accuracy

The best accuracy is provided by Graph SLAM (average error 5-8 cm); however, the method requires significant computational resources and has limited suitability for hard real-time operation (maximum update rate up to 20 Hz). Metric localization, given an accurate map, demonstrates comparable results (8-12 cm), but critically depends on environmental stability and is vulnerable to cyberattacks aimed at spoofing map data. Topological methods are inferior in accuracy (error 50-200 cm), which is due to the very nature of the graph representation of space; however, they excel in processing speed (up to 200 Hz) and robustness to dynamic changes.

An important characteristic of SLAM methods is the convergence rate -- the time it takes for the algorithm to achieve acceptable accuracy in pose estimation and mapping. Figure 2 shows the convergence dynamics of SLAM algorithms.

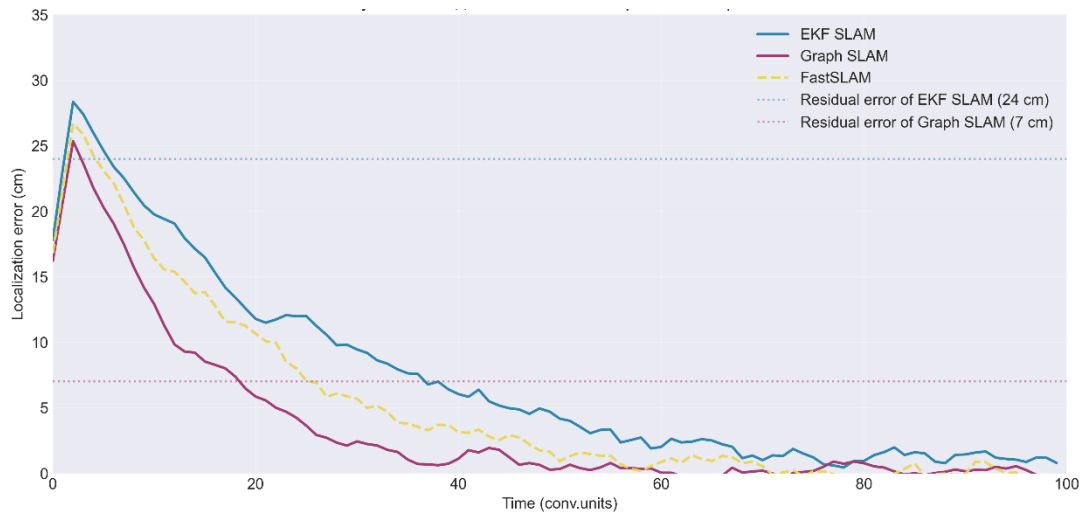


Fig. 2. Convergence of SLAM algorithms over time

EKF SLAM shows a smooth but relatively slow error reduction, associated with the sequential nature of filtering. Graph SLAM demonstrates faster convergence due to batch optimization but requires accumulating a sufficient number of observations. FastSLAM (not included in the main comparison but shown in the graph for completeness) occupies an intermediate position, combining the advantages of both approaches.

Specifics of navigation in cyber-physical systems

The conducted modeling revealed a number of features characteristic specifically of cyber-physical systems. Sensitivity to network delays is most critical for methods requiring the transmission of large data volumes (metric localization, Visual SLAM). With delays exceeding 100 ms, the positioning error for these methods increases by 35-40%, whereas for topological navigation and Graph SLAM, the error increase does not exceed 8-10%.

Vulnerability to cyberattacks was assessed by simulating a navigation data spoofing attack scenario. Visual SLAM and metric localization proved to be the most vulnerable, as they rely heavily on external data. Graph SLAM demonstrated the best resilience due to the use of graph optimization, which helps identify anomalies in the data [6]. Topological navigation showed average resilience due to the possibility of quickly switching to alternative landmarks.

Applicability of methods in various conditions

Navigation technologies find application in various fields, each imposing specific requirements for accuracy, speed, autonomy, and cybersecurity. Application areas with recommended methods are presented in Table 2.

Table 2. Application areas with methods and requirements

Field	Navigation Methods	Accuracy	Speed	Special Requirements
Autonomous Vehicles	Graph SLAM, Deep Learning, RTK-GPS	5-10 cm	Up to 200 km/h	Safety, fault tolerance
Urban Delivery	Topological, Computer Vision	0.5-1 m	Up to 20 km/h	Pedestrian recognition
Indoor Navigation	Visual Odometry, Markers	10-20 cm	Up to 5 km/h	Operation in confined spaces
Agriculture	GPS + RTK, Multispectral Cameras	2-5 cm	Up to 15 km/h	Dust resistance
Space Navigation	Star Trackers, Inertial	Meters–Kilometers	Up to 100,000 km/h	Radiation, communication delays

The conducted analysis allows drawing the following conclusions regarding the applicability of various navigation methods in cyber-physical systems:

Metric localization is preferable for structured environments with a known map, requiring high accuracy over secure transmission channels. Vulnerability of 85% limits its use in open networks without encryption [2].

Topological navigation is optimal for resource-constrained systems and dynamic environments where robustness is more important than accuracy. Low vulnerability (40%) and insensitivity to delays make it ideal for urban delivery with unstable communication [1].

EKF SLAM is a classic solution for small spaces with sufficient computational resources. Medium vulnerability (55%) and the ability to operate without a pre-existing map recommend it for underground and research missions.

Graph SLAM provides the best accuracy and scalability with minimal vulnerability (25%) due to anomaly detection during graph optimization. Despite high resource demands, it is most promising for critical applications.

Visual SLAM is a compromise solution with inexpensive sensors but high vulnerability (80%) and sensitivity to lighting conditions [7, 8]. It is recommended only for indoor environments with controlled conditions and low cybersecurity requirements.

Conclusion

This paper presents a comparative analysis of modern navigation methods used in robotic cyber-physical systems.

It was revealed that no single method is universal -- the choice depends on specific operating conditions and cybersecurity requirements. The most accurate methods (Graph SLAM, metric localization) require significant computational resources and have limitations for real-time operation, while fast methods (topological navigation) are inferior in positioning accuracy.

Graph SLAM provides the best accuracy (5-8 cm) and resilience to cyberattacks (25% vulnerability), but requires maximum computational resources and has a limited update rate (up to 20 Hz). The method is recommended for autonomous vehicles and space missions.

Topological methods are optimal for resource-constrained systems and for operation in dynamic environments with high robustness requirements. Sensitivity to network delays is minimal (error increase less than 10% at 200 ms delay).

Visual SLAM represents the most accessible solution using inexpensive sensors, but is critical to lighting conditions and vulnerable to cyberattacks (80%). It is recommended only for applications with controlled conditions.

Sensor fusion and multi-level data validation are critically important for ensuring navigation reliability in cyber-physical systems, especially when operating in open networks.

A promising direction for further research is the development of hybrid navigation systems that combine the advantages of different methods and adaptively switch between them depending on operating conditions and the level of cyber threats. Also relevant is the creation of specialized protocols for protecting navigation data in open networks, resilient to spoofing attacks and ensuring information integrity under unstable communication channel conditions [2].

References

1. Thrun S., Burgard W., Fox D. Probabilistic robotics. – MIT Press, 2005. – 647 p.
2. Petrov A.V., Ivanov P.S. Cybersecurity challenges in robotic navigation systems // Journal of Robotics and Autonomous Systems. 2024. – Vol. 168. – pp. 104-118.
3. Cadena C., Carlone L., Carrillo H. et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age // IEEE Transactions on Robotics. 2016. – Vol. 32, № 6. – pp. 1309-1332.
4. Roganov V.R. New directions in the development of simulators synthesizing external environment images in training complexes // Труды международного симпозиума "Надежность и качество". 2024. – Vol. 1. – pp. 77-79. (In Russ.).

5. Roganov V.R., Mischev M.J., Zhashkova T.V., Roganova E.V., Kuvshinova O.A., Gushchin S.M. FLIGHT SIMULATOR INFORMATION SUPPORT // Defence S and T Technical Bulletin. 2018. – Vol. 11. – № 1. – pp. 90-98.
6. Grisetti G., Kümmerle R., Stachniss C., Burgard W. A tutorial on graph-based SLAM // IEEE Intelligent Transportation Systems Magazine. 2010. – Vol. 2, № 4. – pp. 31-43.
7. Mur-Artal R., Tardós J.D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras // IEEE Transactions on Robotics. 2017. – Vol. 33, № 5. – pp. 1255-1262.
8. Engel J., Schöps T., Cremers D. LSD-SLAM: Large-scale direct monocular SLAM // European Conference on Computer Vision. – Springer. 2014. – pp. 834-849.
9. Kim A., Eustice R.M. Real-time visual SLAM for autonomous underwater hull inspection using visual saliency // IEEE Transactions on Robotics. 2013. – Vol. 29, № 3. – pp. 719-733.

УДК 004.42

СИСТЕМНЫЙ ПОДХОД К ВИЗУАЛИЗАЦИИ БОЛЬШИХ ГРАФОВ В PYTHON С ИСПОЛЬЗОВАНИЕМ OPENGL: ПРОЕКТИРОВАНИЕ ПАЙПЛАЙНА ЦП–ГП И АНАЛИЗ МАСШТАБИРУЕМОСТИ

М.Е. Ушков*, В.Л. Бурковский**

**Воронежский государственный технический университет, Россия, Воронеж, ushkov.maksim@mail.ru*

***Воронежский государственный технический университет, Россия, Воронеж, bvl@vorstu.ru*

Аннотация. В науке экспериментальные методы быстро развиваются, все чаще встречаются большие наборы данных. Получение интерпретируемых результатов из необработанных данных иногда может осуществляться автоматически; однако существует множество ситуаций, когда на всех этапах обработки возникает необходимость интерактивной визуализации данных. Это позволяет ученому получить представление о происходящем, обнаружить неожиданные закономерности и найти ориентиры для последующих этапов анализа. Существующие инструменты визуализации в основном ориентированы на статические изображения высокого качества для публикаций и не поддерживают интерактивную визуализацию больших наборов данных. В ходе разработки программного обеспечения на Python для визуализации данных больших объемов были определены методы, позволяющие использовать вычислительную мощность современных видеокарт для высокопроизводительной интерактивной визуализации данных. В работе представлены примеры применения этих методов для визуализации графов больших размеров. Несмотря на интерпретируемый и динамический характер Python, удалось достичь очень высокой производительности, используя современные и быстрые библиотеки – NumPy и PyOpenGL. Предложенные методы визуализации будут полезны в широком спектре областей научной визуализации, где возрастает потребность в масштабируемой и быстрой интерактивной визуализации.

Ключевые слова: графы больших размеров, OpenGL, PyOpenGL, графический конвейер, Python, интерактивная визуализация.

A SYSTEM APPROACH TO VISUALIZING LARGE GRAPHS IN PYTHON USING OPENGL: DESIGNING A CPU–GPU PIPELINE AND SCALABILITY ANALYSIS

M.E. Ushkov*, V.L. Burkovsky**

**Voronezh State Technical University, Voronezh, Russia, ushkov.maksim@mail.ru*

***Voronezh State Technical University, Voronezh, Russia, bvl@vorstu.ru*

Abstract. In science, experimental methods are developing rapidly, and large datasets are encountered increasingly often. Obtaining interpretable results from raw data can sometimes be performed automatically; however, in many situations interactive data visualization is required at all stages of processing. This enables researchers to gain insight into ongoing processes, identify unexpected patterns, and establish reference points for subsequent stages of analysis. Existing visualization tools are primarily focused on producing high-quality static images for publication and do not support interactive visualization of large datasets. While developing Python software for large-scale data visualization, methods were created that leverage the computational power of

modern graphics cards to achieve high-performance interactive visualization. Despite the interpreted and dynamic nature of Python, very high performance was achieved through the use of modern and efficient libraries such as NumPy and PyOpenGL. The paper presents examples of applying these methods to the visualization of large-scale graphs. The demonstrated visualization techniques can be useful across a wide range of scientific visualization fields where the demand for scalable and fast interactive visualization continues to grow.

Keywords: large-scale graphs, OpenGL, PyOpenGL, graphics pipeline, Python, interactive visualization.

Во многих научных областях объем данных для визуализации графов растет все быстрее. Большие графы применяются для моделирования и анализа сложных связей в социальных сетях, интернете и телекоммуникациях, биологических системах (геномика, нейронные сети), рекомендательных системах, логистике и транспорте, кибербезопасности, а также в поисковых системах и анализе знаний. Получение данных, хотя и является технически очень сложной задачей, – это лишь первый шаг в научном процессе. Для получения полезной информации необходимы эффективные методы анализа и визуализации данных.

Визуализация исходных данных сразу после их получения зачастую оказывается чрезвычайно полезной, потому что это позволяет ученым делать интуитивные выводы о данных или находить неожиданные закономерности и т. д. Однако большинство существующих инструментов визуализации (таких как matplotlib [1], Chaco [2], PyQwt [3], Bokeh [4], ruytgraph [5]) либо ориентированы на статистическую визуализацию, либо плохо масштабируются для очень больших наборов данных (т.е. содержащих более миллиона точек).

Одна из главных причин этих ограничений заключается в том, что данные инструменты традиционно разрабатываются для центральных процессоров (ЦП). Однако все современные компьютеры включают в себя выделенную электронную схему для графики, называемую графическим процессором (ГП). ГП обычно используются в видеоиграх и 3D-моделировании, но редко – в традиционных научных приложениях визуализации (за исключением областей, связанных с 3D-моделями). При этом ГП не только намного мощнее ЦП с точки зрения вычислительной производительности, но и специально разработаны для приложений визуализации в реальном времени.

В данной статье описывается, как использовать открытый стандарт рендеринга графики OpenGL [6] для аппаратного ускорения интерактивной визуализации в научных задачах с применением языка программирования Python. Отмечается роль программируемого конвейера и шейдеров OpenGL в этом процессе. Приводятся методы, позволяющие достичь очень высокой производительности, несмотря на интерпретируемость Python. Наконец, представлен инструмент на Python для интерактивной визуализации графов больших размеров.

Интерактивная визуализация графов больших размеров

Графический процессор обладает массово-параллельной архитектурой, что делает его особенно эффективным для одновременной обработки большого количества вершин и пикселей. OpenGL – открытый стандарт для рендеринга графики в реальном времени, поддерживаемый большинством платформ и видеокарт, что делает его предпочтительным решением по сравнению с проприетарными технологиями. Спецификация OpenGL применяется для обработки 2D-данных (с использованием ортогональной проекции) или 3D-данных (с использованием перспективной проекции).

Краткий обзор конвейера OpenGL. Конвейер OpenGL включает несколько этапов: преобразование исходных данных в вершины (вершинный шейдер), сборку примитивов (точек, линий, треугольников), растеризацию и обработку фрагментов (пикселей) во фрагментном шейдере. Ключевую роль играет программируемый конвейер, введенный начиная с OpenGL 2.0. Вершинные и фрагментные шейдеры, написанные на GLSL, позволяют гибко управлять преобразованием и визуализацией данных, а каждый элемент данных обрабаты-

вается параллельно по принципу «одна инструкция – несколько данных». В Python доступ к OpenGL осуществляется через PyOpenGL [7].

Стандартные графы в двумерном пространстве можно естественным образом описать с помощью примитивов OpenGL: пара треугольников, образующих квадрат, с обработкой во фрагментном шейдере – для узлов; ребра состоят из линий, а также из двух треугольников, которые при обработке во фрагментном шейдере образуют наконечник; для подписей возможно использование текстур. Однако для обеспечения производительности при визуализации больших объемов данных необходимо использовать определенные методы. Основные из этих методов будут описаны ниже.

Повышение производительности графических приложений напрямую связано с сокращением количества вызовов API OpenGL, поскольку каждый вызов сопровождается накладными расходами, обусловленными передачей управления драйверу и изменением состояния графического конвейера. Данное требование приобретает особую актуальность при использовании интерпретируемых языков, таких как Python, где дополнительное время интерпретации усиливает отрицательное влияние частых обращений к графической библиотеке. Это обуславливает необходимость применения пакетного рендеринга и минимизации числа вызовов OpenGL. OpenGL предоставляет несколько способов пакетного рендеринга.

Одним из ключевых методов достижения высокой производительности при визуализации больших графовых моделей является рациональная организация структур данных на стороне центрального процессора с учетом особенностей передачи данных в графический процессор через OpenGL. Для минимизации накладных расходов при передаче данных предпочтительно использовать структуры, обеспечивающие непрерывное размещение элементов в памяти.

В контексте языка Python оптимальным инструментом для представления подобных структур являются массивы библиотеки NumPy [8]. Они обеспечивают:

- хранение данных в непрерывном блоке памяти;
- строгую типизацию элементов;
- совместимость с буферным интерфейсом, необходимым для эффективной передачи данных в OpenGL.

Использование стандартных контейнеров Python, например списков, нежелательно по причине их фрагментированного размещения в памяти и дополнительных накладных расходов при преобразовании в формат, пригодный для передачи на ГП. Формат хранения в непрерывном блоке позволяет выполнять загрузку данных в видеопамять одним вызовом OpenGL, например через буферные объекты вершин – VBO [6].

При выборе способа организации хранения массивов, например один многомерный массив или несколько одномерных, на стороне ЦП следует разделять данные по частоте обновления. Часто обновляемые данные, например координаты узлов при динамическом перемещении, целесообразно хранить в отдельных массивах и буферных объектах. Это позволяет передавать на ГП только изменяемую часть информации. Цвета и размеры узлов обычно меняются редко и могут быть объединены в структурированный массив, содержащий поля различных типов. Разделение данных по частоте обновления уменьшает объем передаваемых данных и снижает нагрузку на шину передачи, что особенно важно при визуализации больших графов.

На стороне центрального процессора не стоит хранить полный набор данных для рендеринга примитивов, который требует графический конвейер OpenGL. Чаще всего можно ограничиться минимальным набором данных, который описывает примитив и использовать метод рисования экземпляров (*instanced rendering*) [6]. Он позволяет хранить единственный шаблон геометрии, передавая на графический процессор лишь массив параметров экземпляров (координаты, цвета), что существенно сокращает объём передаваемых данных. Например, для рендеринга линии с использованием конвейера можно хранить цвет для ка-

ждой вершины, но лучше ограничиться одним цветом и использовать метод рисования экземпляров.

Корректный выбор типов данных на стороне ЦП также оказывает существенное влияние на производительность и потребление памяти. Современные графические процессоры оптимизированы для работы с форматом float32, который обеспечивает баланс между точностью и пропускной способностью. Однако в ряде случаев использование float32 может быть избыточным, например при хранении z-координаты или булевого значения. В таких случаях возможно применение следующих методов:

- использование целочисленных типов меньшей разрядности;
- упаковка нескольких логических параметров в одно значение посредством побитовых операций.

В последнем случае в вершинном шейдере выполняется обратное декодирование перед использованием значения. Такой подход позволяет повысить плотность упаковки данных без увеличения объема передаваемой информации, что особенно важно при масштабных сценах.

Отдельно следует отметить способ хранения данных цвета. С целью уменьшения объема передаваемой информации цветовые атрибуты вершин целесообразно хранить на стороне ЦП в виде 32-битного беззнакового целого числа, в котором компоненты RGBA упакованы по байтам. Такое представление обеспечивает компактность (4 байта на вершину) и эффективную передачу в буфер вершин. При загрузке в OpenGL атрибут может быть задан как целочисленный с включённой нормализацией, что позволяет графическому процессору автоматически преобразовывать значения диапазона [0,255] в нормализованный формат [0,1], интерпретируемый в шейдере как вес4. Данный подход позволяет сократить объем передаваемой информации без дополнительных вычислительных затрат.

Основным узким местом при визуализации могут быть расчеты на стороне ЦП. Решением этой проблемы может стать перенос части вычислительных операций с центрального процессора на графический процессор. В частности, при реализации операций панорамирования и масштабирования сцены целесообразно передавать в шейдер лишь два параметра: коэффициент масштабирования и вектор смещения. Эти коэффициенты обновляются в соответствии с действиями пользователя, а непосредственное аффинное преобразование координат выполняется в вершинном шейдере. Для визуального выделения узлов модификация цветовых атрибутов может осуществляться на стороне ГП; при этом изменение цвета может выполняться в вершинном шейдере с последующей передачей значения во фрагментный шейдер. Для отображения мультиграфов вычисление геометрии рёбер, например формирование трапецеидальных примитивов на основе координат инцидентных узлов, также может быть реализовано в шейдерной программе, что позволяет избежать предварительной генерации сложной геометрии на стороне ЦП.

Ключевым фактором оптимизации является корректное распределение вычислений между вершинным и фрагментным шейдерами. Поскольку фрагментный шейдер выполняется для каждого пикселя и вызывается значительно чаще, все вычисления, которые могут быть выполнены на уровне вершин с последующей интерполяцией, следует реализовывать в вершинном шейдере. Перенос таких операций во фрагментный шейдер приводит к избыточному повторению вычислений и снижению производительности.

В случае перехода от частных визуальных свойств примитивов к одному общему, например при установке одного цвета всем ребрам, целесообразно выполнять перекомпиляцию шейдера в OpenGL с фиксацией общего значения на этапе компиляции, что позволяет исключить передачу повторяющихся атрибутов от ЦП к ГП и уменьшить объем передаваемых данных.

Результаты

Описанные в статье методы реализованы в программном проекте «GraphViewSurface», предназначенном для визуализации графов большого размера с использованием OpenGL на языке Python. Основной целью разработки является обеспечение высокой производительности при работе с крупными объемами графовых данных.

Для достижения требуемой производительности архитектура приложения построена в соответствии с модельно-ориентированным подходом. Такой подход позволяет минимизировать операции преобразования и копирования данных. Все интерактивные действия пользователя интерпретируются как изменения состояния модели данных, а не как модификация данных представления. Формат хранения данных модели изначально согласован с требованиями визуализации, что исключает необходимость дополнительных преобразований перед отображением.

Важной особенностью архитектуры является применение принципа ленивых вычислений. Вычислительные операции и преобразования выполняются только при наличии фактической необходимости и исключительно для той части данных, которая требуется для текущего отображения. Данный подход особенно эффективен при визуализации нескольких порожденных подграфов, поскольку позволяет ограничить объем обрабатываемой информации.

В работе проведена сравнительная оценка производительности GraphViewSurface по отношению к библиотекам matplotlib, pyqtgraph и pyqtgraph с поддержкой OpenGL на примере задачи динамической визуализации. В качестве показателей производительности использовались среднее время рендеринга кадра, среднее время отображения первого кадра и средний объем потребляемой памяти. Экспериментальное исследование выполнено на двух аппаратных платформах с различными техническими характеристиками при одинаковом разрешении экрана 1920×1080 пикселей и плотности около 90 пикселей на дюйм. Первая платформа построена на базе актуальных аппаратных решений, вторая – на базе значительно более ранних и менее производительных компонентов. В рамках тестирования формировались идентичные графовые структуры в GraphViewSurface и в сравниваемых библиотеках. Доля узлов составляла 10% от общего числа объектов сцены, доля ребер – 90%. Полученные результаты представлены в таблицах 1, 2, 3, 4, 5, 6 и на рисунке 1.

Таблица 1. Среднее время рендеринга кадра на первой платформе

Кол-во элементов	matplotlib	pyqtgraph	pyqtgraph + opengl	opengl
1000	28,3	7,0	1,0	0,2
10000	257,6	31,2	5,7	0,4
100000	2487,6	288,2	50,4	1,9
1000000		2868,6	483,9	17,3

Таблица 2. Среднее время рендеринга первого кадра на первой платформе

Кол-во элементов	matplotlib	pyqtgraph	pyqtgraph + opengl	opengl
1000	134,8	282,2	441,9	209,3
10000	398,0	450,7	495,3	238,3
100000	2891,6	614,5	615,1	234,3
1000000		3859,0	1658,2	257,8

Таблица 3. Средний объем потребляемой памяти на первой платформе

Кол-во элементов	matplotlib	pyqtgraph	pyqtgraph + opengl	opengl
1000	93,8	101,7	158,2	89,0
10000	97,5	102,9	164,4	90,8
100000	133,7	107,4	204,8	104,3
1000000		155,4	498,3	155,3

Таблица 4. Среднее время рендеринга кадра на второй платформе

Кол-во элементов	matplotlib	pyqtgraph	pyqtgraph + opengl	opengl
1000	51,4	14,0	13,6	5,5
10000	502,2	75,4	40,8	25,5
100000	4696,5	629,6	312,6	106,8
1000000		7018,0	3611,3	878,3

Таблица 5. Среднее время рендеринга первого кадра на второй платформе

Кол-во элементов	matplotlib	pyqtgraph	pyqtgraph + opengl	opengl
1000	231,8	211,7	808,9	251,3
10000	858,7	681,1	777,5	364,0
100000	5900,6	1314,9	1635,7	707,1
1000000		8701,4	6141,6	925,5

Таблица 6. Средний объем потребляемой памяти на второй платформе

Кол-во элементов	matplotlib	pyqtgraph	pyqtgraph + opengl	opengl
1000	87,2	88,5	115,5	53,4
10000	91,2	90,1	124,0	54,1
100000	127,3	95,4	181,3	68,5
1000000		143,5	472,2	119,1

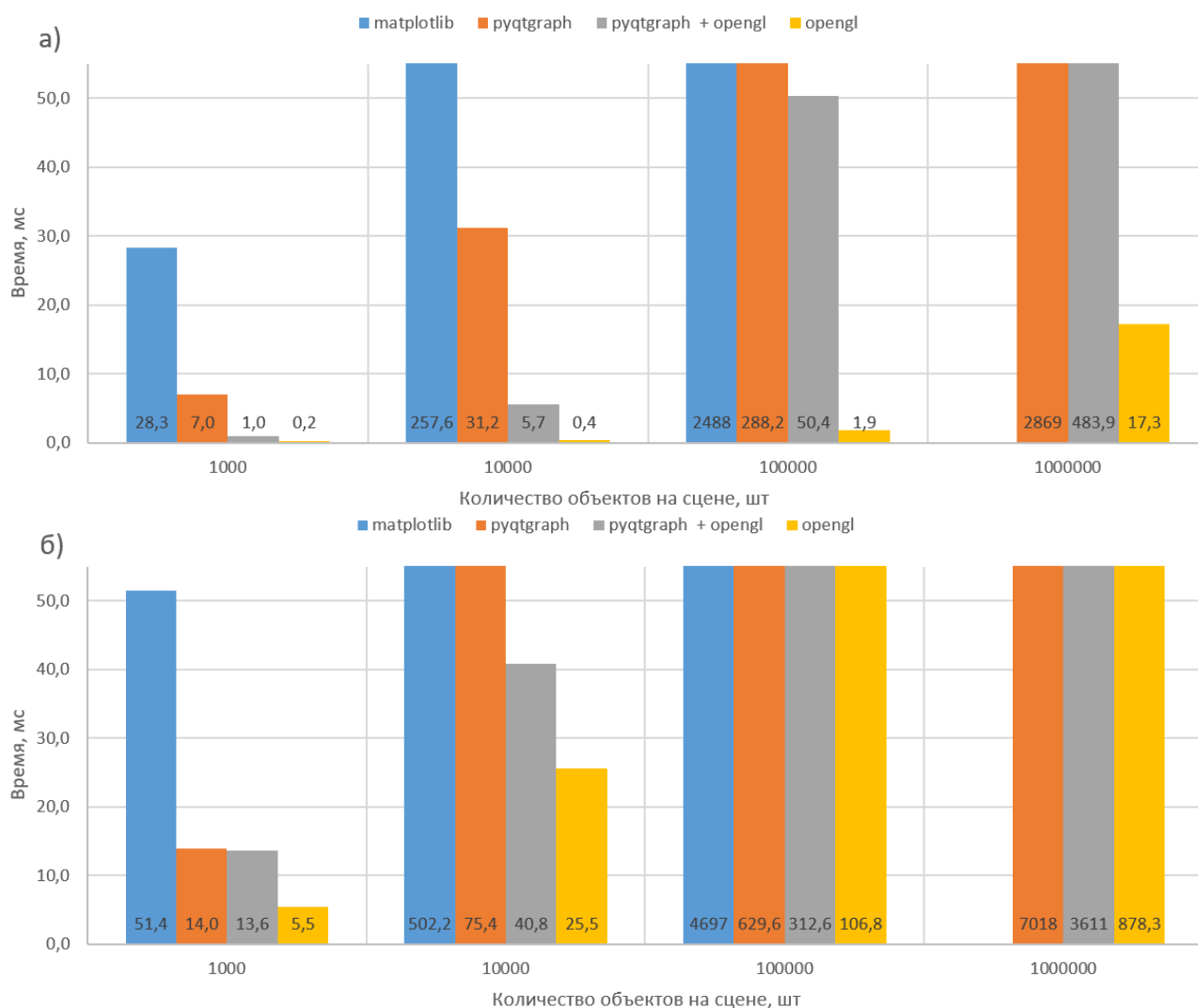


Рис. 1. Диаграммы среднего времени рендеринга кадра (а – на первой платформе, б – на второй платформе)

По результатам экспериментов установлено, что среднее время кадра в GraphViewSurface во всех случаях ниже, чем у остальных рассматриваемых библиотек. Это объясняется тем, что matplotlib и pyplotgraph преимущественно используют центральный процессор для формирования изображения, включая режим работы pyplotgraph с поддержкой OpenGL. В отличие от них, GraphViewSurface осуществляет передачу данных непосредственно из оперативной памяти в память графического процессора и выполняет растеризацию средствами графического конвейера OpenGL, что снижает нагрузку на центральный процессор и повышает эффективность рендеринга.

Потребление оперативной памяти GraphViewSurface сопоставимо с аналогичным показателем у сравниваемых библиотек. Существенное увеличение объема используемой памяти наблюдается при включении поддержки OpenGL в pyplotgraph. Время отображения первого кадра при небольших объемах данных может незначительно превышать соответствующие показатели других библиотек, что связано с накладными расходами на инициализацию контекста OpenGL. Однако при увеличении объема визуализируемых данных преимущество архитектурных решений GraphViewSurface становится более выраженным.

Выводы

В данной работе продемонстрировано, что OpenGL – широко известный стандарт для аппаратного ускорения графики – может использоваться для быстрой интерактивной визуализации больших графов в научных приложениях. Были описаны способы достижения высокой производительности в Python за счёт передачи данных на графический процессор и использования шейдеров для их преобразования и рендеринга. Эти методы минимизируют накладные расходы, связанные с использованием Python, вызовами API OpenGL и передачей данных между ЦП и ГП. Наконец, было представлено приложение для визуализации больших графов и сравнительный тест производительности с другими библиотеками. Показано, что по мере масштабирования объёма данных для визуализации достижение высокой производительности становится возможным только при задействовании графического процессора.

Библиографический список

1. Библиотека для научной визуализации matplotlib [Электронный ресурс]. – Режим доступа: <https://matplotlib.org/> – Дата доступа: 15.03.2026.
2. Библиотека визуализации Chaco [Электронный ресурс]. – Режим доступа: <https://chaco.readthedocs.io/en/latest/> – Дата доступа: 15.03.2026.
3. Библиотека визуализации PyQwt [Электронный ресурс]. – Режим доступа: <https://github.com/PyQwt/PyQwt> – Дата доступа: 15.03.2026.
4. Библиотека визуализации Bokeh [Электронный ресурс]. – Режим доступа: <https://github.com/bokeh/bokeh> – Дата доступа: 15.03.2026.
5. Библиотека визуализации pyplotgraph [Электронный ресурс]. – Режим доступа: <https://github.com/pyqtgraph/pyqtgraph> – Дата доступа: 15.03.2026.
6. Кроссплатформенный программный интерфейс OpenGL [Электронный ресурс]. – Режим доступа: <https://www.opengl.org/> – Дата доступа: 15.03.2026.
7. Библиотека визуализации PyOpenGL [Электронный ресурс]. – Режим доступа: <https://pyopengl.sourceforge.net/> – Дата доступа: 15.03.2026.
8. Пакет для научных вычислений NumPy [Электронный ресурс]. – Режим доступа: <https://numpy.org/> – Дата доступа: 15.03.2026.

УДК 004.421; ГРНТИ 20.53

СОВРЕМЕННЫЕ ИТ-РЕШЕНИЯ ДЛЯ ОПТИМИЗАЦИИ ПРОЦЕССОВ ЗАПИСИ И ПРОВЕДЕНИЯ ПРИЁМА ПАЦИЕНТА В МЕДИЦИНСКИХ УЧРЕЖДЕНИЯХ

С.В. Лыкин, А.В. Крошилин

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, s-lykin05@mail.ru*

Аннотация. В статье рассматриваются современные информационные технологии, направленные на оптимизацию ключевых процессов медицинских организаций – записи на приём и самого амбулаторного приёма. Анализируются возможности электронных регистратур, интеграции с порталами государственных услуг, а также специализированных медицинских информационных систем. Отдельное внимание уделяется значению сбора и анализа статистических данных как фундаментального элемента оптимизации. На примере внедрения BI-систем и аналитических платформ показано, как трансформация первичных данных в управленческую информацию позволяет не только выявлять проблемные моменты в загрузке врачей и потоках пациентов, но и формировать стратегию повышения доступности и качества медицинской помощи.

Ключевые слова: цифровизация здравоохранения, оптимизация процессов, запись на приём к врачу, медицинские информационные системы, сбор статистики, управление на основе данных, эффективность работы медучреждения, мониторинг показателей.

MODERN IT SOLUTIONS FOR OPTIMIZING PATIENT REGISTRATION AND ADMISSION PROCESSES IN MEDICAL INSTITUTIONS

S.V. Lykin, A.V. Kroshilin

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, s-lykin05@mail.ru*

The summary. This article examines modern information technologies aimed at optimizing key processes in healthcare organizations—appointment scheduling and outpatient visits. It analyzes the capabilities of electronic registration systems, integration with government service portals, and specialized medical information systems. Particular attention is paid to the importance of collecting and analyzing statistical data as a fundamental element of optimization. Using the implementation of BI systems and analytical platforms as an example, it demonstrates how transforming raw data into management information not only helps identify bottlenecks in physician workloads and patient flows but also formulates a strategy for improving the accessibility and quality of medical care.

Keywords: digitalization of healthcare, process optimization, appointment booking, medical information systems, BI analytics, statistics collection, data-driven management, healthcare facility efficiency, dashboards, and performance monitoring.

Современное здравоохранение вступило в эпоху глубокой цифровой трансформации. Одним из наиболее чувствительных для пациентов и сложных для организации аспектов работы медицинских учреждений является процесс записи на приём и последующего проведения амбулаторного приёма. Очереди, неудобное расписание, нерациональная нагрузка на врачей и, как следствие, снижение доступности медицинской помощи – эти проблемы остаются актуальными для многих клиник, несмотря на внедрение базовых информационных систем.

Решение этих проблем лежит не только в плоскости автоматизации рутинных операций (замена бумажной карты электронной), но и в применении комплексных ИТ-решений, способных анализировать внутренние процессы учреждения. Цель данной статьи – классифицировать современные ИТ-инструменты, используемые для оптимизации записи и приёма пациентов, и обосновать тезис о том, что именно сбор и глубокая аналитика статистических данных являются ключевым драйвером и основой для эффективной оптимизации [1].

Современные ИТ-решения для оптимизации записи пациентов

Процесс записи на приём – первая точка контакта пациента с медицинской организацией, и от его эффективности зависит общее впечатление и доступность помощи [2]. Традиционные регистратуры, работающие с бумажными журналами, уступают место многоканальным цифровым системам. Далее будут приведены два решения для улучшения процесса записи на приём пациента.

Первое из них – электронная регистратура и интегрированные сервисы записи. Основой оптимизации записи является внедрение модуля «Электронная регистратура» в составе Медицинской информационной системы (МИС). Современные решения предлагают интеграцию с различными каналами:

- Портал государственных услуг и региональные порталы здравоохранения. Это решение обеспечивает гражданам удобный удаленный доступ к расписанию врачей. Настройка корректного взаимодействия региональной МИС с Госуслугами позволяет не только увеличить долю успешных онлайн-записей, но и получать объективные данные о востребованности тех или иных специалистов.

- Информационные табло и терминалы самообслуживания. Размещенные в холлах клиники, они позволяют пациентам без помощи регистратора выбрать удобное время, снижая нагрузку на персонал.

- Голосовые роботы и кол-центры. Современные голосовые меню и чат-боты на базе ИИ способны обрабатывать до 70-80% стандартных вызовов (запись, отмена, перенос), маршрутизируя сложные вопросы на операторов.

Вторым ИТ-решением является управление расписанием на основе данных. Ключевое преимущество цифровых систем – возможность гибко настраивать расписание. ИТ-решения позволяют внедрять временные слоты для разных типов приема (первичный/повторный), срочные записи и автоматическое распределение нагрузки. Однако, настройка расписания становится по-настоящему эффективной только при наличии аналитики по фактической загрузке, позволяющей выявлять дефицит узких специалистов или нерационально длинные талоны [3].

ИТ-инструменты в процессе проведения приёма

Оптимизация не заканчивается на записи. Сам процесс приёма также требует цифрового сопровождения. Современные МИС трансформируют работу врача, смещая фокус с бумажной работы на взаимодействие с пациентом. В данном аспекте было выделено три основных инструмента, помогающих получить желаемый результат.

1. *Электронная медицинская карта (ЭМК) и голосовой ввод.* ЭМК является ядром внутрибольничных процессов. Она обеспечивает мгновенный доступ к истории болезни, результатам анализов и предыдущим назначениям. Для ускорения процесса заполнения данных внедряются шаблоны, протоколы и, на передовых площадках, технологии голосового заполнения, что позволяет врачу не отвлекаться на набор текста во время беседы с пациентом.

2. *Интеграция с диагностическими службами.* Передача данных из лабораторных и диагностических систем напрямую в ЭМК исключает потерю результатов и сокращает время врача на их поиск и вставку в анамнез. Это напрямую влияет на скорость и качество приёма.

3. *Мониторинг загруженности кабинетов и потоков пациентов.* Современные платформы управления здравоохранением позволяют в режиме реального времени видеть распределение потоков пациентов, загрузку конкретных специалистов и оборачиваемость кабинетов. Такие системы дают возможность не только наблюдать, но и прогнозировать пиковые нагрузки, предлагая перенаправить потоки или скорректировать расписание в моменте [4].

Трансформация данных в управленческие решения

Перечисленные выше инструменты автоматизации генерируют огромные массивы данных. Без их осмысления цифровизация рискует остаться лишь неэффективной. Именно сбор статистики и её последующий анализ превращают ИТ-инструменты в механизм непрерывного улучшения [5, 6].

Для руководителя медицинской организации (главного врача) [7, 8] интерфейс системы должен предоставлять сводную информацию по ключевым показателям эффективности (KPI), позволяя перейти от интуитивного управления к объективному, основанному на метриках (таблица 1).

Таблица 1. Описание ключевых показателей эффективности медицинской организации

<i>Метрика</i>	<i>Описание</i>
Динамика потока пациентов	Отображение количества успешных записей, отмен и неявок в динамике (день/неделя/месяц) с возможностью сегментации по отделениям.
Доступность помощи	Расчет среднего времени ожидания приема (по каждой специальности и конкретному врачу) и анализ «глубины» свободной записи.
Нагрузка на персонал	Коэффициент фактической занятости врачебного времени (отношение времени, проведенного с пациентами, к общему времени работы).

Для внешних контролирующих органов и руководства регионального/федерального уровня разрабатываемая система должна обеспечивать формирование сводной статистики на основе унифицированных классификаторов. Это достигается путем создания системы витрин данных, где информация из первичных документов (талонов амбулаторного пациента, карт вызова) агрегируется в показатели для мониторинга целевых индикаторов государственных программ (например, охват профилактическими осмотрами, доля пациентов, записавшихся онлайн). Встроенные алгоритмы системы должны позволять провести такой анализ. Этапы цикла оптимизации системы и их описание представлены в таблице 2.

Таблица 2. Описание этапов цикла оптимизации

<i>Название этапа</i>	<i>Описание этапа</i>
Верификация проблемы	Система фиксирует отклонение KPI «Среднее время ожидания» от нормативного значения. Анализ данных хронометража показывает, что фактическое время, затрачиваемое врачом на одного пациента, превышает плановое на 5 минут.
Корреляционный анализ	Сопоставление временных меток в ЭМК выявляет, что увеличение длительности приема коррелирует с ручным вводом больших объемов текстовой информации (отсутствие шаблонов или протоколов осмотра).
Выработка и внедрение решения	На основе этих данных принимается управленческое решение о доработке интерфейса системы: разработке и внедрении специализированных шаблонов осмотра для кардиологов с полями автозаполнения и справочниками.
Оценка эффективности (обратная связь)	После внедрения доработок система продолжает сбор статистики. Сравнение «слепков» данных «до» и «после» позволяет объективно подтвердить или опровергнуть эффективность изменений. Если средняя длительность приема снизилась, а очередь сократилась – цикл оптимизации успешен. Если нет – требуется дальнейший анализ (например, проверка юзабилити внедренных шаблонов).

Такие циклы являются основой концепции непрерывного улучшения качества (CQI) в здравоохранении, они невозможны без встроенных инструментов статистики и аналитики.

Заключение

Подводя итог, можно констатировать, что современные ИТ-решения для оптимизации процессов записи и приёма пациентов выходят далеко за рамки простой автоматизации рабочих мест. Сегодня это комплексные экосистемы, включающие внешние сервисы для пациента (Госуслуги, информационные автоматы) и мощные аналитические платформы для управления учреждением.

Внедрение МИС и инструментов электронной записи решает задачу операционной эффективности – снижения очередей и ускорения доступа. Однако только внедрение систем сбора и анализа статистики (BI-решений) позволяет перейти к стратегической оптимизации. Превращая накопленные данные о времени приёма, загрузке специалистов, потоках пациентов и отказах в наглядную информацию для принятия решений, руководители здравоохранения получают возможность не просто реагировать на текущие проблемы, а прогнозировать их и предотвращать.

Таким образом, сбор статистики с помощью ИТ-инструментов является не просто важной частью, а ключевым условием успешной и непрерывной оптимизации работы любого современного медицинского учреждения, ориентированного на качество и доступность помощи.

Библиографический список

1. Крошилин А.В., Предметно-ориентированные информационные системы: учебное пособие / Крошилин А.В., Крошилина С.В., Овечкин Г.В. // – Москва: КУРС, 2023. – С. 176.
2. Перепелкин Д.А., Попова А.А., Крошилин А.В., Крошилина С.В., Компьютерное моделирование процессов поддержки принятия решений врача-стоматолога на основе семантических сетей // Вестник РГРТУ. №89 - Рязань: РГРТУ, 2024. – С. 155.
3. Жулева С.Ю., Крошилин А.В., Крошилина С.В., Саморукова О.Д., Задачи разработки систем медицинского назначения при выборе схемы медикаментозного лечения // Вестник РГРТУ. №88 - Рязань: РГРТУ, 2024. – С. 142.
4. Илларионова И.В., Полховская Н.М., Минаев Ю.Л., Определение экономической эффективности использования электронного реестра в работе медицинского отделения // Вестник медицинского института «РЕАВИЗ», №2, 2017 – С. 131.
5. Ивановский Б.Г., Инновации в здравоохранении: проблемы эффективности и внедрения // Экономические и социальные проблемы России, №2, 2021. – С. 144.
6. Овечкин Г.В., Крошилин А.В., Крошилина С.В., Саморукова О.Д. Моделирование процессов управления в организационных системах на основе теории нечетких когнитивных карт // Вестник РГРТУ. №91 - Рязань: РГРТУ, 2025. – 210 с.
7. Крошилина С.В., Жулева С.Ю., Крошилин А.В. Программный комплекс для управления медицинскими материальными потоками на основе нечеткой логики // Биомедицинская радиоэлектроника. 2025. Т. 28. № 4. С. 85-92. DOI: <https://doi.org/10.18127/j15604136-202504-10>
8. Жулев В. И., Крошилин А. В., Крошилина С. В. Проектирование систем поддержки принятия решений. Учебное пособие для вузов. –М.: Горячая линия – Телеком, 2023. – 180 с.: ил.

УДК 004.92; ГРНТИ 50.41.25

МЕТОД ФРАКТАЛЬНЫХ ОПЕРАТОРОВ И АЛГОРИТМЫ РОСТА ДИНАМИЧЕСКИХ ФРАКТАЛЬНЫХ СТРУКТУР

А.А. Теплов

Московский государственный технический университет им. Н.Э. Баумана,
Российская Федерация, Москва, teploff.aa@gmail.com

Аннотация. Рассматривается метод фрактальных операторов для формализации динамических фрактальных структур и алгоритмы их формирования процессом роста. Предложенный метод преодолевает ограничения известных статических подходов (L-систем, систем итерируемых функций) и позволяет моделировать сложные динамические процессы. Разработана классификация алгоритмов роста по сложности: классический, однофазный, многофазный и стохастический. Представлены результаты численного моделирования для двумерных и трехмерных фрактальных структур с анализом метрических характеристик и вычислительной сложности. Экспериментально подтверждена эффективность предложенных алгоритмов для моделирования динамических фрактальных систем в вычислительных комплексах.

Ключевые слова: фрактальные операторы, динамические фрактальные структуры, алгоритмы роста фракталов, моделирование процессов роста, стохастические фракталы, вычислительная сложность, программное обеспечение вычислительных систем.

METHOD OF FRACTAL OPERATORS AND GROWTH ALGORITHMS FOR DYNAMIC FRACTAL STRUCTURES

A.A. Teplov

Bauman Moscow State Technical University,
Russia, Moscow, teploff.aa@gmail.com

The summary. The paper considers a method of fractal operators for formalizing dynamic fractal structures and algorithms for their formation by growth process. The proposed method overcomes the limitations of known static approaches (L-systems, iterated function systems) and allows modeling complex dynamic processes. A classification of growth algorithms by complexity has been developed: classical, single-phase, multi-phase, and stochastic. Results of numerical modeling for two-dimensional and three-dimensional fractal structures with analysis of metric characteristics and computational complexity are presented. The effectiveness of the proposed algorithms for modeling dynamic fractal systems in computing complexes has been experimentally confirmed.

Keywords: fractal operators, dynamic fractal structures, fractal growth algorithms, growth process modeling, stochastic fractals, computational complexity, software for computing systems.

Введение

Проблемы моделирования динамических фрактальных структур являются актуальными в технологиях виртуальной и дополненной реальности, системах компьютерной графики и визуализации [4-5]. Известные подходы, основанные на L-системах [3] и системах итерируемых функций (СИФ) [1], имеют принципиальные ограничения: они не способны адекватно описывать процессы формирования структуры из элементарных форм, внутреннюю трансформацию элементов и стохастичность динамических явлений [8].

L-системы характеризуются дискретностью генерации, фиксированными точками ветвления и отсутствием механизма обратной связи. СИФ работают с готовыми элементами, не содержат механизма внутреннего развития и являются детерминированными системами. Это делает их неэффективными для моделирования сложных динамических процессов с вычислительной сложностью $O(n^d)$ для L-систем и $O(m \cdot n)$ для СИФ, где требуется полный пересчет структуры на каждом шаге.

В настоящей работе предлагается метод фрактальных операторов и алгоритмы для моделирования динамических фрактальных структур процессом роста с оптимизацией вычислительной сложности в программном обеспечении вычислительных систем [8-10].

Метод фрактальных операторов

Предлагается подход к формализации фрактальных структур через введение понятия фрактального оператора [8]. Рассмотрим построение геометрического фрактала типа «Кривой Коха» [1-2]. Известный алгоритм состоит в делении элементарных отрезков на m частей и замене их на n отрезков. Введем фрактальный оператор:

$$f = f(m, n, l), \quad (1)$$

который определяется числами m , n и множеством l , на котором действует [8]. Для геометрических фракталов каждому оператору соответствует геометрический примитив.

Применению оператора (1) в k циклах соответствует обобщенный оператор:

$$F_k = f^k. \quad (2)$$

В общем случае, обобщенный оператор k циклов содержит произведение k различных операторов:

$$F_{ks} = \prod_{i=1}^k f_i = f_1 f_2 \cdots f_k. \quad (3)$$

При циклическом применении обобщенного оператора (3) z раз получаем z -периодический фрактал:

$$(F_{ks})^z = (\prod_{i=1}^k f_i)^z = (f_1 f_2 \cdots f_k)^z. \quad (4)$$

Размерность периодического фрактала определяется выражением:

$$D = \frac{\ln(n_1 n_2 \cdots n_k)}{\ln(m_1 m_2 \cdots m_k)}. \quad (5)$$

Стохастизация фрактальных операторов. Для моделирования реалистичных динамических систем необходимо придать вероятностный характер базовым операторам. Структура стохастического оператора имеет вид:

$$f_i = \begin{cases} f_{i1}, & \text{if } p_1 \leq p \leq p_2 \\ f_{i2}, & \text{if } p \geq p_2 \\ \dots \end{cases}, \quad (6)$$

где f_{ir} – возможные значения оператора, условия их реализации определяются вероятностями p_k . При циклическом применении стохастических операторов получаются сложные фрактальные структуры с возрастающим числом участков различной структуры и размерности [8].

Преимущества операторного подхода: в отличие от известных методов [6-7], где на каждом шаге выполняется полный пересчет, предложенный метод позволяет вычислять фракталы динамически на основе результатов предыдущих циклов. Это обеспечивает значительное сокращение представления фрактала за счет замены тождественных операций символами операторов [8].

Алгоритмы роста фрактальных структур

На основе метода фрактальных операторов разработана классификация алгоритмов формирования фракталов процессом роста по сложности и функциональности.

Классический алгоритм

Соответствует известной схеме построения геометрических фракталов. Принцип преобразования для трехмерных фракталов:

$$a_k = 2^{-k} a, \quad (7)$$

обеспечивает неизменный размах фрактала $l = a$. Детали одного порядка образуются и растут одновременно. Вычислительная сложность составляет $O(n^k)$ для k итераций.

Однофазный алгоритм

Рост начинается из точки [9]. Когда отрезок достигает размера a , в его середине формируется подструктура. Скорости роста всех элементов подбираются так, что они одновременно достигают размера a . Характерная особенность: одновременное образование новых примитивов и их синхронный рост – все примитивы находятся в одной фазе. Для трехмерных фракталов принцип преобразования:

$$a_k = a, \quad (8)$$

обеспечивает показательный рост размаха $l_k = 2^k a$. Вычислительная сложность $O(n \cdot k)$, что существенно эффективнее классического алгоритма.

Многофазный алгоритм

Обобщение однофазного алгоритма путем устранения «жесткой синхронизации»: рост каждого элементарного отрезка выполняется независимо с шагом Δa в каждом цикле [9]. Фазы роста элементов различаются. Преимущество – более реалистичное моделирова-

ние процессов роста естественных фракталоподобных структур. Вычислительная сложность $O(n \cdot k + m)$, где m – число независимых элементов.

Стохастический алгоритм

Влияние стохастичности моделируется путем случайных вариаций длины фрактальных отрезков с амплитудой Δa_s . Стохастичность вносится не только по фазе роста, но и по размерам растущих элементов [9]. Алгоритм имеет сложность $O(n \cdot k + m + s)$, где s – число стохастических вариаций, но обеспечивает более реалистичное моделирование динамических систем.

Обобщенный алгоритм роста фрактальных структур

На основе метода фрактальных операторов разработан обобщенный алгоритм формирования двумерных и трехмерных фрактальных структур. Алгоритм включает следующие этапы:

1. Инициализация начальной конфигурации (точка или элементарный примитив).
2. Итерационное применение фрактальных операторов согласно выбранному принципу роста (однофазный, многофазный или стохастический).
3. Вычисление метрических характеристик (длина L , размах l , площадь S или объем V).
4. Проверка критерия остановки (достижение заданного порядка фрактала или метрических характеристик).
5. Визуализация результата.

Алгоритм реализован в виде программного комплекса для моделирования динамических фрактальных систем. Схемы алгоритмов представлены на рисунках 1 и 2.

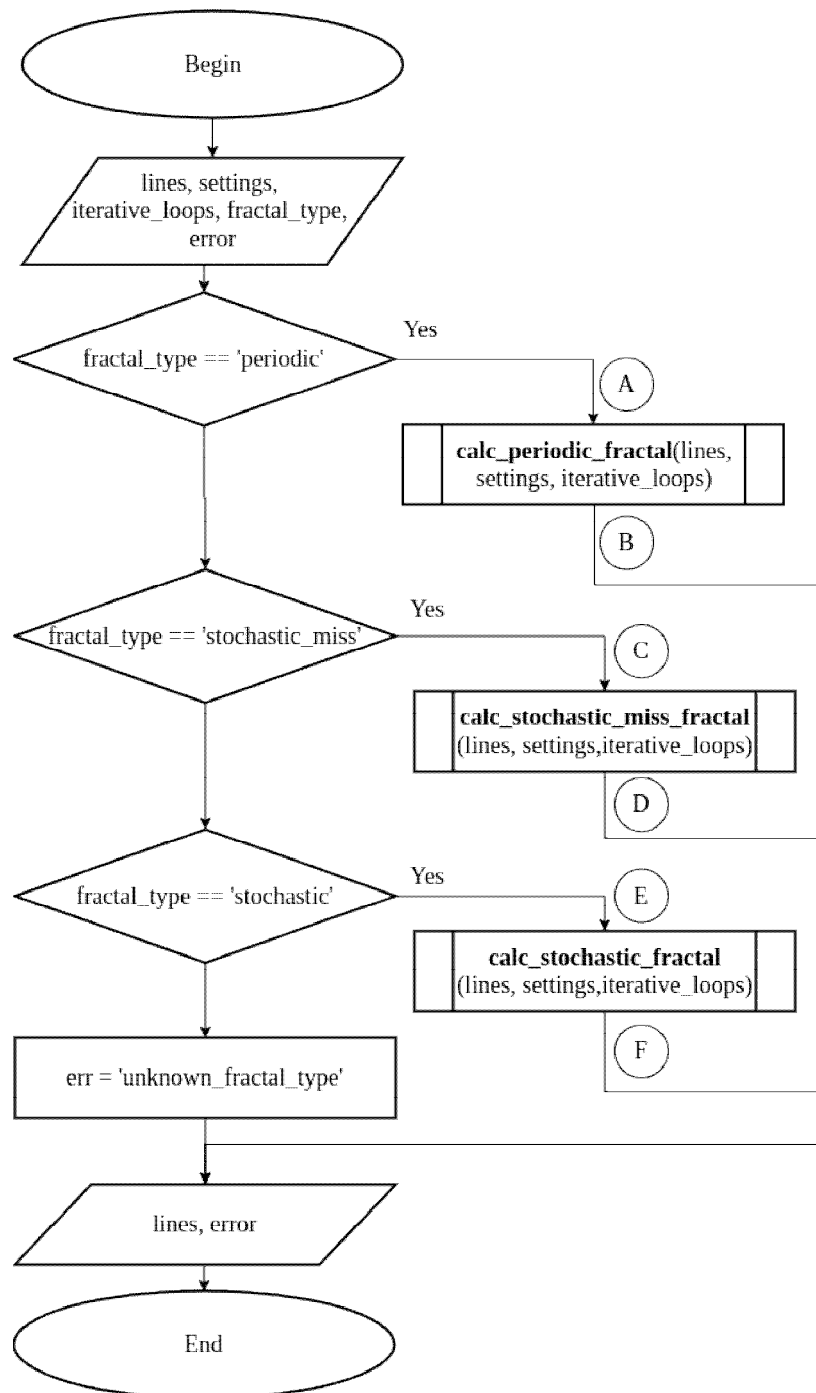


Рис. 1. Схема алгоритма формирования двумерных фрактальных структур

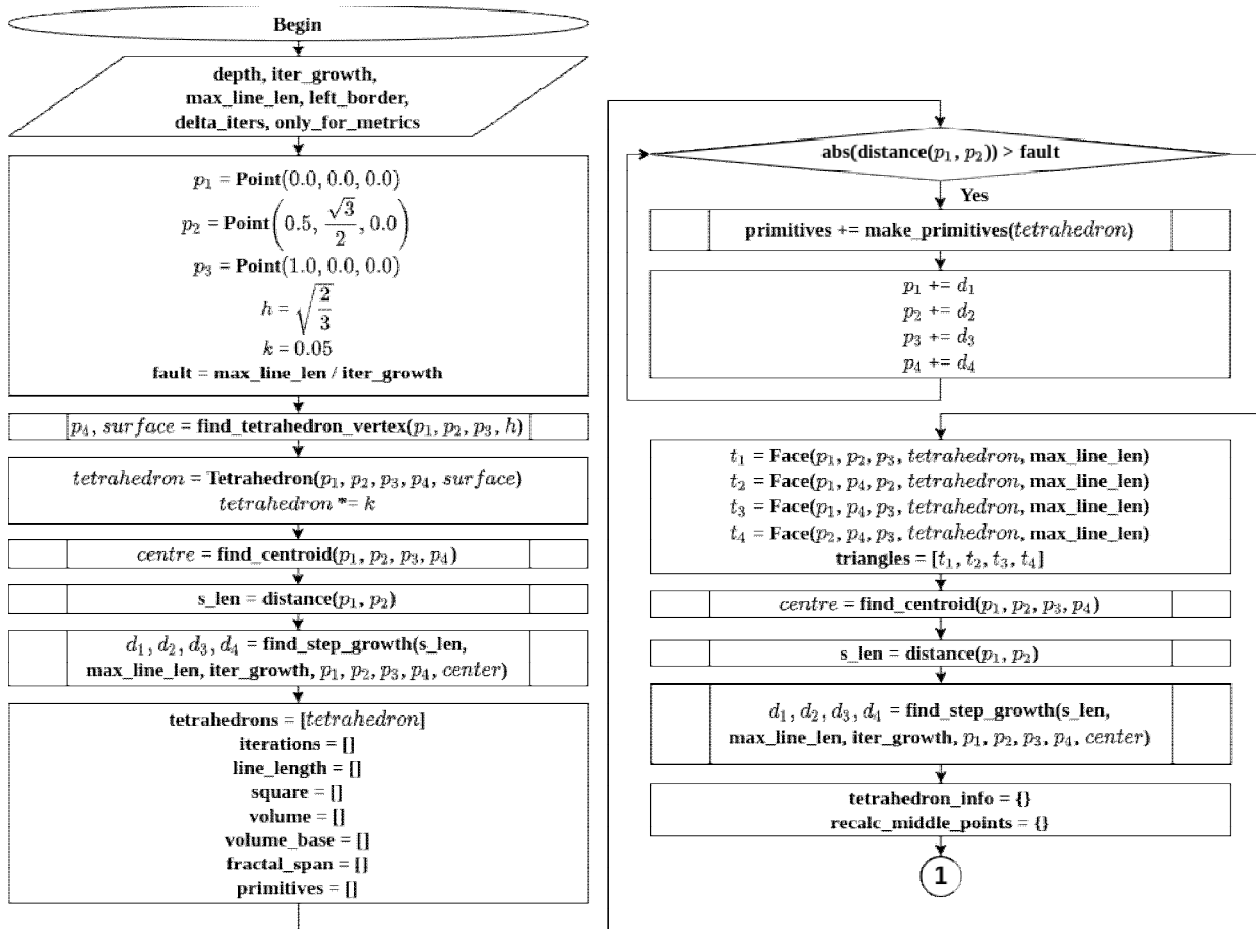


Рис. 2. Схема алгоритма формирования трехмерных фрактальных структур

Результаты численного моделирования

Экспериментальные исследования проведены для двумерных и трехмерных фрактальных структур [9-10]. Входные данные: последовательность от 100 до 1000 итерационных циклов роста, порядок фрактала от 2 до 10. Проведено более 100 серий вычислений для каждой конфигурации. Затраты машинного времени составили более 200 часов на платформе x86-64 Linux.

Размах l и длина L фрактальной линии для однофазной и многофазной моделей представлены на рисунке 3. Для однофазного алгоритма зависимости описываются ломаной линией, вершины которой аппроксимируются показательными функциями:

$$L(k) = a \cdot n^k, \quad (9)$$

$$l(k) = a \cdot m^k, \quad (10)$$

Линейные участки между вершинами соответствуют масштабному росту фрактальных отрезков, а в узловых точках формируются новые отрезки.

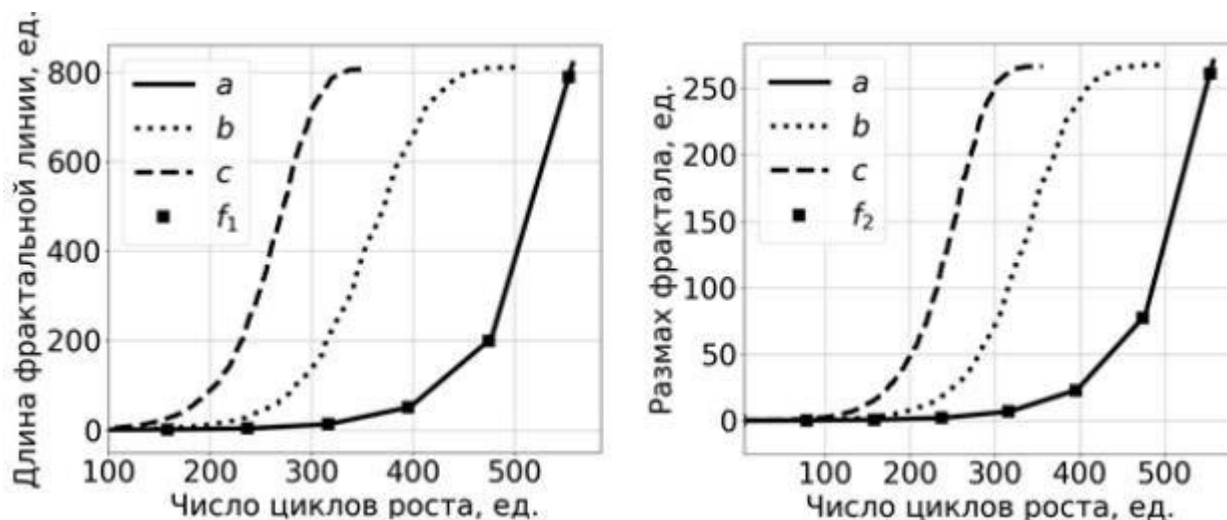


Рис. 3. Длина и размах фрактальной линии для однофазной (а) и многофазной (b, c) моделей роста. Квадратные маркеры соответствуют теоретическим зависимостям

Для многофазного алгоритма зависимости аппроксимируются логистической кривой:

$$L(t) = \frac{4^k a}{1 + 4^k e^{-\varepsilon t}}, \quad (11)$$

где ε – относительная скорость роста, определяемая как параметр аппроксимации. Смена алгоритма с однофазного на многофазный приводит к существенной трансформации закономерностей изменения метрических характеристик фрактала.

Результаты для трехмерных фракталов

Моделирование роста трехмерных тетраэдрических фрактальных поверхностей (рисунк 4) показало эффективность многофазного алгоритма. Метрические характеристики — площадь S и размах l — демонстрируют закономерности, аналогичные двумерному случаю.

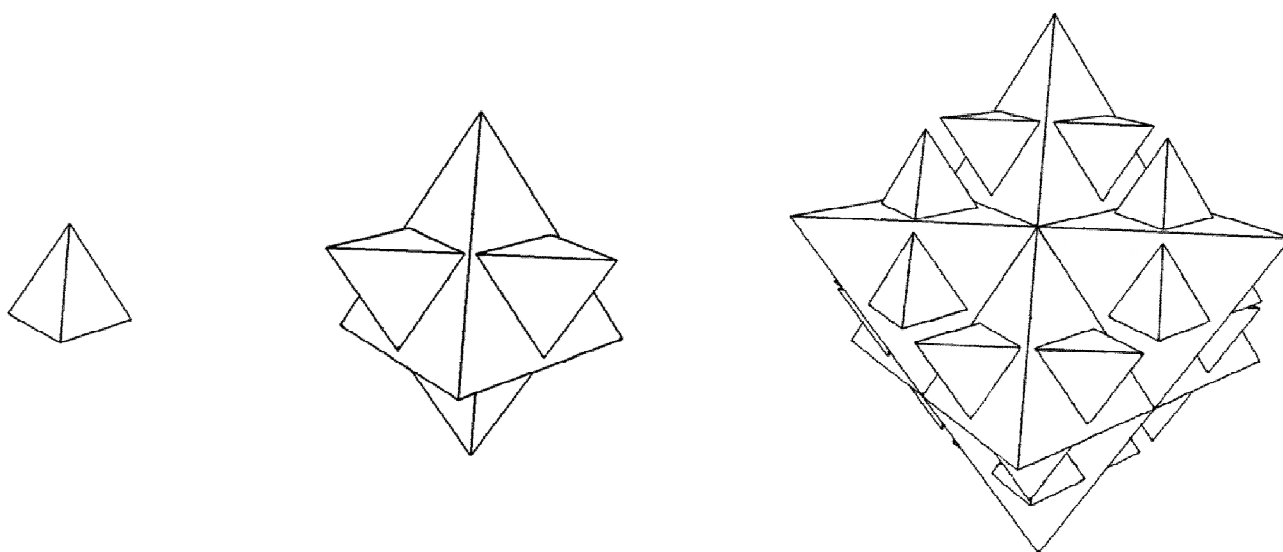


Рис. 4. Моделирование процесса роста трехмерного фрактального комплекса (этапы роста слева направо)

Результаты численного моделирования для многофазного алгоритма представлены на рисунке 5. Экспериментально подтверждено, что при однофазном и многофазном алгоритмах результат дает один и тот же фрактал [10]. Это свидетельствует о сохранении общей формы растущего объекта независимо от выбранного алгоритма роста.

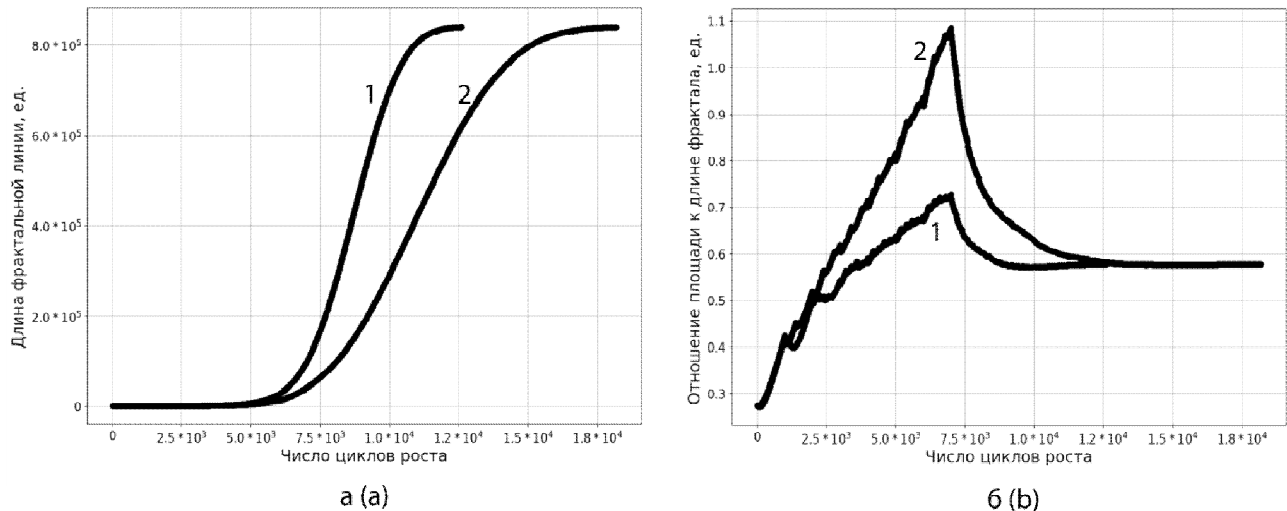


Рис. 5. Метрические характеристики роста трехмерного фрактала (многофазный алгоритм)

Вычислительная сложность алгоритмов

Проведен сравнительный анализ вычислительной сложности разработанных алгоритмов. Для классического алгоритма сложность составляет $O(n^k)$, где n – число элементов на одном уровне, k – порядок фрактала. Однофазный алгоритм имеет сложность $O(nk)$, что обеспечивает существенный выигрыш при больших k . Многофазный алгоритм характеризуется сложностью $O(nk + m)$, где m – число независимых элементов. Стохастический алгоритм имеет сложность $O(nk + m + s)$, где s – число стохастических вариаций.

Экспериментальное сравнение времени выполнения алгоритмов показало, что для фракталов порядка $k = 8$ однофазный и многофазный алгоритмы в 15-20 раз эффективнее классического по времени вычислений. Стохастический алгоритм увеличивает время на 30-40% относительно многофазного, но обеспечивает качественно более реалистичное моделирование.

Заключение

В работе представлен операторный подход к формализации фрактальных структур, позволяющий вычислять фракталы динамически на основе результатов предыдущих итераций. Введенное понятие фрактального оператора и его обобщений (периодических и стохастических) обеспечивает сокращение объема вычислений за счет замены тождественных операций символами операторов.

Численное моделирование показало, что для двумерных фракталов при однофазном алгоритме метрические характеристики описываются показательными функциями $L(k) = a \cdot nk$ и $l(k) = a \cdot mk$, а при многофазном – логистической кривой. Для трехмерных тетраэдрических структур экспериментально установлено сохранение итоговой формы фрактала при различных алгоритмах роста. Сравнительный анализ вычислительной сложности выявил существенное преимущество однофазного и многофазного алгоритмов: выигрыш в 15-20 раз по времени вычислений для фракталов порядка $k = 8$ относительно классического подхода.

Разработанные алгоритмы реализованы в программном комплексе и применимы для задач моделирования в технологиях виртуальной и дополненной реальности, где требуется динамическое формирование сложных геометрических структур с оптимальными вычислительными затратами.

Библиографический список

1. Мандельброт Б. Фрактальная геометрия природы. М.: Институт компьютерных исследований, 2002. 656 с.
2. Федер Е. Фракталы: Пер. с англ. М.: Мир, 1991. 254 с.
3. Окорочков В.А., Сандракова Е.В. Фракталы в фундаментальной физике. М.: МИФИ, 2009. 460 с.
4. Акулич М. Дополненная, виртуальная, смешанная реальность и маркетинг // Создано в Интеллектуальной здательской системе Ridero, 2021
5. Ланье Д. На заре новой эры. Автобиография «отца» виртуальной реальности / ООО «Издательство «Эксмо», 2019 г. 421 с.
6. Новиков В.У., Козлов Г.В. Фрактальный анализ макромолекул // Успехи химии. 2000. Т. 69, № 4. С. 378-399.
7. Потапов А.А. Фракталы в радиофизике и радиолокации: Топология выборки. 2-е изд. М.: Университетская книга, 2005. 848 с.
8. Теплов А.А. Метод фрактальных операторов в задачах моделирования динамических систем // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2020. № 3. С. 45-58.
9. Теплов А.А. Алгоритмы формирования фрактальных структур процессом роста // Современные технологии в науке и образовании СТНО-2021: сб. тр. междунар. науч. конф. Рязань, 2021. Т. 2. С. 112-116.
10. Теплов А.А. Численное моделирование роста трехмерных фрактальных структур // Программные системы: теория и приложения. 2022. Т. 13, № 2. С. 89-104.

УДК 656.073; ГРНТИ 55.43.35

DEVELOPMENT OF MULTIMODAL TRANSPORT SYSTEMS BASED ON THE INTEGRATION OF UNMANNED TECHNOLOGIES AND DIGITAL NAVIGATION

T.V. Zhashkova*, L.N. Pepel*, M.A. Terekhin*

*Penza State Technological University,
Russia, Penza, tzhaskova@gmail.com

Abstract. The prospects for the development of multimodal transport systems using unmanned aerial vehicles (UAVs) and intelligent navigation systems are considered. The possibilities of integrating air and ground transport into a single logistics network are analyzed. Special attention is paid to improving the reliability, safety, and environmental friendliness of transport transportation through the implementation of digital solutions. Directions for further research in the field of adaptive traffic flow management are proposed.

Keywords: unmanned aerial vehicles, multimodal transport system, navigation system, transport digitalization, logistics, intelligent control.

РАЗВИТИЕ МУЛЬТИМОДАЛЬНЫХ ТРАНСПОРТНЫХ СИСТЕМ НА ОСНОВЕ ИНТЕГРАЦИИ БЕСПИЛОТНЫХ ТЕХНОЛОГИЙ И ЦИФРОВОЙ НАВИГАЦИИ

Т.В. Жашкова*, Л.Н. Пепел*, М.А. Терехин*

*Пензенский государственный технологический университет,
Россия, Пенза, tzhaskova@gmail.com

Аннотация. Рассмотрены перспективы развития мультимодальных транспортных систем с применением беспилотных летательных аппаратов (БПЛА) и интеллектуальных систем навигации. Проанализированы возможности интеграции воздушного и наземного транспорта в единую логистическую сеть. Особое внимание уделено вопросам повышения надежности, безопасности и экологичности транспортных перевозок за счет внедрения циф-

ровых решений. Предложены направления дальнейших исследований в области адаптивного управления транспортными потоками.

Ключевые слова: беспилотные летательные аппараты, мультимодальная транспортная система, навигационная система, цифровизация транспорта, логистика, интеллектуальное управление.

In the context of digital transformation of the modern information and communication environment, the development of integrated transport systems capable of ensuring fast, safe, and efficient delivery of goods and passengers becomes particularly important. As noted in speeches at the state level, the creation of a "single digital control circuit for the transport industry" is one of the priority tasks of national technological development [1]. This requires the implementation of innovative solutions based on the use of unmanned technologies, modern navigation systems, and big data analytics, which together form the basis of the digital transport ecosystem.

Multimodal transport systems, combining various modes of transport into a single network, make it possible to optimize logistics routes, reduce delivery time, and decrease infrastructure load. A key role in such integration is played by unmanned transport systems, including aerial vehicles (UAVs), which can be used for monitoring traffic flows, delivering small-sized goods to hard-to-reach areas, and responding promptly to road incidents [2, pp. 67-72]. The efficiency of their operation directly depends on the degree of integration into the digital ecosystem, which ensures unified data exchange and management standards.

Navigation systems based on GLONASS and GPS, supplemented by computer vision and artificial intelligence (AI) technologies, provide high positioning accuracy and real-time motion control. This allows for the creation of adaptive routes that take into account the current traffic situation, weather conditions, and other factors [3, pp. 45-50]. The integration of satellite navigation with local sensor networks forms a digital framework for managing multimodal flows within a unified digital ecosystem.

Figure 1 shows a diagram of the interaction of key elements of an intelligent multimodal transport system using UAVs.

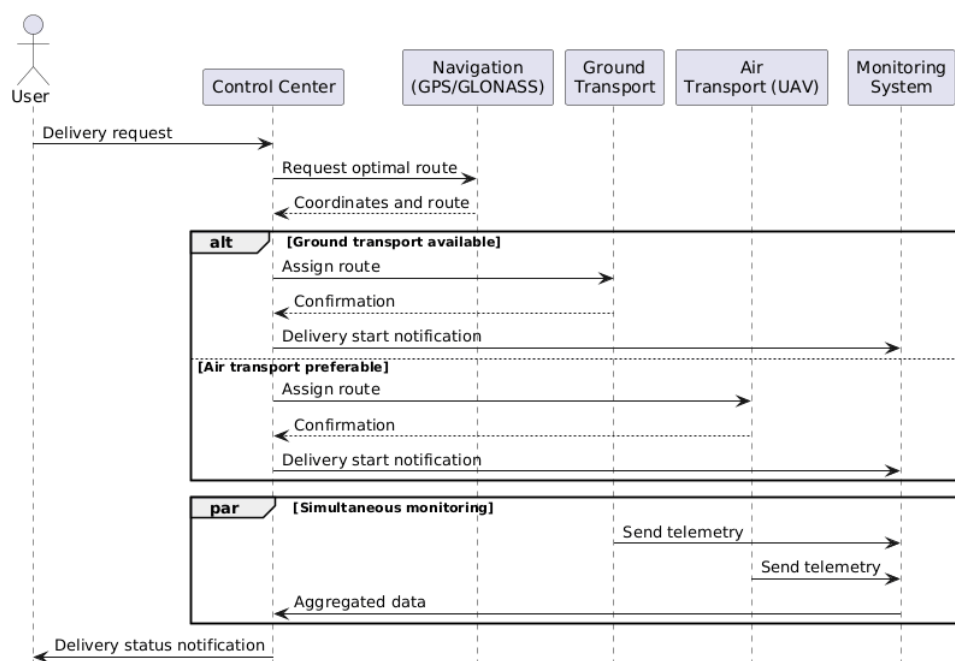


Fig. 1. Sequence diagram of elements of a multimodal transport system using UAVs

The control center, having received a request, accesses the navigation subsystem (GPS/GLONASS) to calculate the optimal route and, based on this data, selects the most efficient type of transport – ground or air (UAV). A key stage is parallel monitoring: both types of transport

transmit telemetry in real time to the monitoring system, which allows data aggregation, prompt response to changes in the situation, and informing the user about the delivery status.

This diagram shows the principle of adaptive, data-driven selection of a logistics solution within a single digital circuit. The process is initiated by a user's delivery request [4, p. 112].

The effectiveness of implementing such systems is confirmed by calculations based on the model of minimizing total transportation costs [2, pp. 98-102]:

$$C = \sum_{i=1}^n (c_{1i}x_i + c_{2i}y_i) \rightarrow \min, \quad (1)$$

где c_{1i} – costs for ground transportation;

c_{2i} – costs for UAV delivery;

x_i and y_i – volumes of transportation by the respective modes of transport.

The use of UAVs in combination with intelligent navigation systems can reduce delivery time by an average of 25-30%, as well as reduce the number of road accidents by improving control over the traffic situation [2, p. 45].

Development of a multimodal system based on digital navigation and UAVs requires the convergence of telecommunications, robotics, and artificial intelligence technologies. Pilot projects will allow for the formation of a technological and regulatory foundation for the transformation of the transport industry [5-9].

The conducted analysis confirms the high relevance and significant potential of integrating unmanned technologies and digital navigation for the development of multimodal transport systems. The proposed model of adaptive traffic flow management provides a comprehensive solution to the problems of optimizing logistics routes, improving the safety and environmental friendliness of transportation.

However, the implementation of these technologies faces a number of challenges, including the need to improve the regulatory framework, increase the reliability of control systems, and ensure information security. An important direction for further research is the development of universal standards for the interaction of various modes of transport within a single digital ecosystem [3, p. 187].

A promising direction is also the integration of artificial intelligence systems for predicting traffic flows and preventing congestion, which will increase the capacity of the transport network by 15-20% [2, p. 112]. Additional research in the field of energy efficiency of unmanned vehicles will expand their application in logistics operations, including long-distance cargo delivery.

The implementation of the approaches presented in the article involves the development of interaction and the creation of infrastructure that supports the digital transformation of the transport industry. The introduction of multimodal transport systems based on unmanned technologies and digital navigation will be an important step towards creating a sustainable and competitive transport infrastructure of the future that meets the requirements of the digital ecosystem [8, 9].

The article is published with the support of RSF grant No. 25-21-00445, <https://rscf.ru/project/25-21-00445/> "Conducting research and development of a mathematical apparatus minimizing deviations in the location of 3D models of reference objects synthesized by aviation simulator simulators for their observation in multispectral mode."

References

1. Kozlov P.N. Unmanned Systems in Transport Infrastructure. – St. Petersburg: Nauka, 2021. – 210 p.

2. Address of the President of the Russian Federation to the Federal Assembly dated February 28, 2025 // Official Internet Portal of Legal Information. – URL: <http://publication.pravo.gov.ru/document/0001202502280001>, (дата обращения: 27.01.2026).
3. Sidorov A.V. Digital Technologies in Logistics. – Moscow: Transport, 2022. – 180 p.
4. Petrova E.L. Intelligent Transport Systems: Principles and Architecture. – Moscow: INFRA-M, 2023. – 254 c.
5. Gromov A.I., Kiselyova T.V. Environmental Aspects of UAV Usage in Logistics // Транспорт: наука, техника, управление. – 2024. – № 5. – pp. 28-35. (In Russ.).
6. Vasiliev R.S. Adaptive Control Algorithms in Multimodal Systems // Автоматика и телемеханика. – 2023. – № 12. – pp. 89-101. (In Russ.).
7. Zhashkova T.V., Parshin A.A., Renskova A.Yu. Mathematical and Information-Structural Models for Identifying the Time-Frequency Structure of Complex Signals under Conditions of A Priori Uncertainty // Современные информационные технологии. – 2025. – № 42 (42). – pp. 123-127. (In Russ.).
8. Roganov V.R., Mikheev M.Yu., Dolgovosov B.S., Asmolova E.A. Basic concepts of building modern software and hardware systems of aviation training simulators to solve the tasks of training pilots// In the collection: II International Scientific Forum on Sustainable Development and Innovation (WFSDI 2023). Conference Proceedings. Ekaterinburg. 2024. – pp. 2159-2170.
9. Kuvshinova O.A., Kuvshinova E.V. Application of a Neural Network for Exclusion of Active Segments from Processing in Real Time // Образование и наука в современном мире. Инновации. 2025. – № 1 (56). – pp. 205-211. (In Russ.).

УДК 004.94; ГРНТИ 50.41.25

ПОСТРОЕНИЕ НЕЧЕТКИХ КОГНИТИВНЫХ МОДЕЛЕЙ ГРУППОЙ ЭКСПЕРТОВ: АКТУАЛЬНОСТЬ ЗАДАЧИ И ПЛАН ИССЛЕДОВАНИЙ

Р.А. Исаев, В.А. Игнатенко

*Брянский государственный технический университет,
Российская Федерация, Брянск, Ruslan-Isaev-32@yandex.ru*

Аннотация. Статья посвящена вопросам построения нечетких когнитивных моделей с участием группы экспертов. Рассмотрены особенности процесса группового когнитивного моделирования, которые обосновывают актуальность разработки специальных алгоритмов группового построения нечетких когнитивных карт. Сформулирован перечень задач, подлежащих решению: устранение терминологической неоднозначности, формирование согласованного множества концептов, определение отношений влияния между концептами и определение знаков и интенсивностей этих влияний. Представлен план дальнейших исследований, включающий разработку соответствующих алгоритмов структурной и параметрической идентификации нечетких когнитивных карт.

Ключевые слова: когнитивное моделирование, нечеткая когнитивная карта, структурная идентификация, параметрическая идентификация, экспертная группа, согласование экспертных мнений.

CONSTRUCTION OF FUZZY COGNITIVE MODELS BY A GROUP OF EXPERTS: RELEVANCE OF THE PROBLEM AND RESEARCH PLAN

R.A. Isaev, V.A. Ignatenko

*Bryansk State Technical University,
Russia, Bryansk, Ruslan-Isaev-32@yandex.ru*

Abstract. The article is devoted to the issues of constructing fuzzy cognitive models with the participation of a group of experts. The features of the group cognitive modeling process are considered, which substantiate the relevance of developing special algorithms for the group construction of fuzzy cognitive maps. The list of tasks to be solved is formulated: elimination of terminological ambiguity, formation of a consistent set of concepts, definition of influence relations between concepts and definition of signs and intensities of these influences. A plan for further research is presented, including the development of appropriate algorithms for the structural and parametric identification of fuzzy cognitive maps.

Keywords: cognitive modeling, fuzzy cognitive map, structural identification, parametric identification, expert group, coordination of expert opinions.

Введение

Когнитивное моделирование – это инструмент исследования сложных систем, основанный на построении их формальных моделей, называемых когнитивными моделями. Анализ такой модели позволяет понять внутреннюю структуру системы и принципы взаимодействия составляющих ее факторов [1], а также определить оптимальные сценарии целенаправленного воздействия на конкретные факторы для перевода системы в требуемое состояние.

В основе когнитивного моделирования лежит математический аппарат когнитивных карт – в частности, одной из известных их разновидностей являются нечеткие когнитивные карты (НKK) В.Б. Силова [2]. Модель анализируемой системы, основанная на НKK Силова, может быть представлена в виде взвешенного ориентированного графа. Вершины такого графа (называемые концептами) описывают факторы системы, а дуги задают отношения влияния между ними. Интенсивность каждого влияния определяется весовым параметром дуги, что позволяет формализовать не только качественные, но и количественные оценки характера связей между концептами.

Данная работа посвящена рассмотрению круга вопросов, возникающих при переходе от индивидуального к коллективному построению нечетких когнитивных карт. Основное внимание уделяется обоснованию актуальности данной предметной области и формированию программы дальнейших исследований. В статье намечаются ключевые этапы работы и подходы, которые далее будут положены в основу соответствующего математического аппарата.

Актуальность группового построения нечетких когнитивных моделей

Процесс когнитивного моделирования на практике имеет ряд особенностей, которые обуславливают необходимость перехода от индивидуальных методов построения когнитивных моделей к коллективным (групповым).

Во-первых, построение и исследование НKK реальных систем, как правило, не может быть результатом работы лишь одного специалиста. Данный процесс требует интеграции знаний и согласования позиций как минимум трех категорий участников: экспертов в предметной области, специалистов по когнитивному моделированию и лиц, принимающих решения. При исследовании относительно простых систем возможна ситуация, когда один специалист совмещает две или даже все три роли, однако при достаточно высокой сложности моделируемой системы такое совмещение становится плохо осуществимым, и функции участников процесса разделяются. В частности, сам процесс построения НKK, в общем случае, предполагает наличие множества субъектов с различными компетенциями и точками зрения на моделируемую систему, что делает задачу согласования их мнений неотъемлемой частью группового когнитивного моделирования.

Во-вторых, по мере усложнения моделируемой системы количество экспертов, привлекаемых к построению когнитивной модели, закономерно возрастает. Сложные социально-экономические, организационно-технические, экологические и иные системы требуют учета множества факторов различной природы, находящихся в компетенции специалистов из разных предметных областей. При этом следует отметить, что процесс построения НKK предполагает последовательное решение двух задач: идентификацию структуры модели (т.е. определение состава концептов и наличия влияний между ними) и идентификацию ее параметров (оценку весов дуг, отражающих силы влияний) [3]. На каждом из указанных этапов неизбежно возникают расхождения в позициях экспертов, обусловленные различиями в их специализации и субъективных восприятиях моделируемой системы. Это порождает потребность в разработке формализованных процедур, позволяющих не просто агрегировать индивидуальные оценки, но также выявлять, анализировать и согласовывать содер-

жательные противоречия в позициях специалистов как на уровне структуры, так и на уровне параметров когнитивной модели.

В-третьих, при построении когнитивных моделей сложных систем, находящихся «на стыке» различных предметных областей, эксперты, как правило, не обладают равными и полными компетенциями относительно всех аспектов моделируемой системы. Каждый специалист может уверенно судить, как правило, лишь о некоторой части концептов и связей, соответствующих его специализации (например, только об экологических или только об экономических факторах). Это означает, что процесс группового построения НКК не должен предполагать обязательного участия каждого эксперта в оценке всех элементов модели. Напротив, необходимы алгоритмы, поддерживающие частичное построение НКК, когда разные фрагменты структуры и параметров модели заполняются разными (возможно, пересекающимися) подгруппами экспертов, а в частном случае – даже одним компетентным специалистом. Необходимость учета данной особенности усложняет задачу согласования мнений экспертов, поскольку для различных элементов НКК состав экспертной группы (в том числе ее размер) может меняться, что требует разработки гибких механизмов согласования получаемой экспертной информации.

Таким образом, рассмотренные особенности процесса построения нечетких когнитивных карт сложных систем позволяют рассматривать задачу группового когнитивного моделирования как актуальное направление исследований.

Этапы группового построения нечетких когнитивных моделей

Процесс построения нечеткой когнитивной модели с участием группы экспертов предполагает последовательное прохождение ряда следующих этапов, частично соответствующих этапам идентификации модели одним экспертом.

1. Устранение неоднозначности терминологии, возникающей при описании предметной области разными экспертами. При этом, поскольку один и тот же фактор может быть обозначен различными понятиями и вербальными формулировками, для обеспечения поддержки данного этапа требуется разработка процедуры приведения наименований концептов к единому словарю.

2. Формирование согласованного множества концептов (первый подэтап структурной идентификации модели). Для этого каждый эксперт предлагает собственный набор факторов, значимых, с его точки зрения, для описания системы. Полученные множества факторов могут частично пересекаться, однако в общем случае не являются идентичными и даже могут не содержать общих элементов, что отражает различия в специализации экспертов. Необходимо на основе этих множеств осуществить построение единого набора концептов когнитивной карты, который будет использован при дальнейшем моделировании.

3. Определение отношений влияния между концептами (второй подэтап структурной идентификации модели). После фиксации согласованного набора концептов каждый эксперт указывает наличие и направление влияний между ними. Индивидуальные матрицы смежности, полученные от разных специалистов, подлежат согласованию для построения единого графа, отражающего коллективное представление о структуре взаимосвязей в исследуемой системе.

4. Определение знаков и интенсивности влияний между концептами (параметрическая идентификация). На данном этапе эксперты указывают знаки и оценивают силы влияний, и требуется согласование индивидуальных оценок для получения итоговых параметров когнитивной модели, с контролем степени согласованности экспертных суждений.

Перечисленные этапы составляют содержание процесса группового построения нечеткой когнитивной модели и определяют направления разработки соответствующих алгоритмов.

Предлагаемый план исследований

Дальнейшая работа в рамках исследуемой задачи предполагает реализацию следующих процессов.

Первым из них является анализ степени исследованности изучаемого вопроса в отечественных и зарубежных научных работах. Предполагается провести обзор существующих подходов к структурной и параметрической идентификации нечетких когнитивных карт в условиях привлечения группы экспертов, а также методов согласования экспертных оценок, потенциально применимых в задачах группового когнитивного моделирования. Результаты данного обзора будут положены в основу процессов разработки алгоритмов групповой структурной и параметрической идентификации НКК.

В рамках разработки алгоритмов структурной идентификации предполагается создание двух взаимосвязанных алгоритмов:

- алгоритма формирования согласованного множества концептов;
- алгоритма определения отношений влияния между концептами на фиксированном (по результатам выполнения предыдущего алгоритма) множестве концептов, обеспечивающего согласование индивидуальных матриц смежности, полученных от разных специалистов.

Конкретный математический аппарат, который ляжет в основу алгоритма параметрической идентификации НКК группой экспертов, будет определен по результатам анализа литературы. Ключевыми компонентами такого алгоритма должны являться механизмы оценки степени согласованности экспертных мнений и, при необходимости, их корректировки для достижения требуемого уровня согласованности перед агрегацией индивидуальных оценок в итоговые параметры когнитивной модели.

Выводы

В рамках работы выявлены особенности процесса построения нечетких когнитивных моделей с участием группы экспертов, которые обуславливают актуальность исследований в данной области. К таким особенностям относятся наличие множества участников с различными компетенциями, необходимость согласования их позиций на этапах структурной и параметрической идентификации модели, а также возможность фрагментарного построения модели, когда разные эксперты описывают лишь отдельные ее участки в соответствии со своими специализациями. Совокупность указанных факторов позволяет рассматривать задачу группового построения НКК как актуальное и значимое направление в рамках когнитивного моделирования, требующее разработки соответствующего алгоритмического обеспечения и специализированных процедур согласования экспертных суждений.

Библиографический список

1. Захарова А.А., Подвесовский А.Г., Исаев Р.А. Нечеткие когнитивные модели в управлении слабоструктурированными социально-экономическими системами // Информационные и математические технологии в науке и управлении. 2020. № 4 (20). С. 5-23. DOI:10.38028/ESI.2020.20.4.001
2. Силов В.Б. Принятие стратегических решений в нечеткой обстановке. – М.: ИНПРО-РЕС, 1995.
3. Подвесовский А.Г., Исаев Р.А. Идентификация структуры и параметров нечетких когнитивных моделей: экспертные и статистические методы // International Journal of Open Information Technologies. 2019. Т. 7. № 6. С. 35-61.

УДК 621.396; ГРНТИ 47.47

МОДЕЛЬ КОНТРОЛЯ ДИНАМИЧЕСКОЙ СИСТЕМЫ

М.А. Посохов

Московский технический университет связи и информатики,
Российская Федерация, Москва, mikhail.posokhov@yandex.ru

Аннотация. В работе приводятся результаты предыдущих исследований, выполненных автором данной статьи. На основе них проводится подробное изучение свойств разработанной модели контроля динамической системы. Также приводятся результаты расчетов в зависимости от входных параметров таких, как длина интервала проверки, линейный множитель затрат, количество изменений состояния динамической системы. На основе полученных результатов строятся выводы о свойствах оптимальности разработанной модели и их причинах. Исследуется модель оптимальных средних суммарных затрат на контроль в зависимости от постановки фиксированной по времени точки контроля.

Ключевые слова: случайные процессы, теория надежности, качество данных, математическое моделирование, постановка контрольных точек, задача нелинейного программирования.

THE MODEL OF DYNAMIC SYSTEM CONTROL

M.A. Pososkhov

Moscow Technical University of Communications and Informatics,
Russia, Moscow, mikhail.posokhov@yandex.ru

The summary. This paper presents the results of previous research conducted by the author. Based on these studies, a detailed examination of the properties of the developed dynamic system control model is conducted. Calculation results are also presented depending on input parameters such as the length of the control interval, the linear cost multiplier, and the number of state changes in the dynamic system. Based on these results, conclusions are drawn regarding the optimality properties of the developed model and their causes. A model of optimal average total inspection costs is examined depending on the setting of a fixed-time checkpoint.

Keywords: stochastic processes, reliability theory, data quality, mathematical modeling, checkpoint set, nonlinear programming.

В работах [1-2] исследована проблема обеспечения правильного функционирования системы, параметры которой меняются со временем, и рассматривался пример системы, состоящей из одного объекта, на который воздействовали внешние и внутренние возмущения, меняющие его состояние. Таким образом, система имела одно «правильное» состояние, а все воздействия меняли это состояние на «неправильное».

Авторы работ [3-5] предложили свои варианты алгоритмов и технологий контроля динамических систем, которые были основаны на конкретном математическом аппарате. Здесь представлено решение задачи определения моментов контроля состояния системы, при которых система находится в «неправильном» состоянии минимальное время. Контроль проводится в течение интервала времени $(0, B)$. Контроль заключался в том, чтобы обнаружить «неправильное» состояние и мгновенно привести в «правильное», если это необходимо. В момент времени B заканчивается процесс контроля. Суммарные затраты на контроль состоят из затрат: на отсутствие, опоздание и проведение контроля. В данной работе фокус делается на отсутствие и опоздание.

Общие результаты

Пусть на интервале $(0, B)$ происходит n изменений состояния системы на «неправильные» в моменты, которые являются элементами вектора $t = [t_1, t_2, \dots, t_n]$, где $t_i > t_j$ при $i > j$ и $t_j < B$, $i = 1, 2, \dots, n$.

Рассматривается вариант однократной проверки. Это означает, что на интервале $(0, B)$ ставится только одна точка контроля z , $(0 < z < \infty)$. Очевидно, что, если момент

контроля совпадает с моментом изменения состояния системы, то затраты на пребывания в «неправильном» состоянии равны 0. Ввиду этого, была предложена методика расчета суммарных затрат на контроль в виде формулы ниже (1).

$$S_0(t, z, B, a) = a \left[\sum_{\substack{i=1 \\ z \geq t_i}}^k (z - t_i) + \sum_{\substack{i=1 \\ z < t_i}}^m (B - t_i) \right] = a \left[kz + mB - \sum_{i=1}^n t_i \right], \quad (1)$$

где k – число моментов изменения состояний объекта, меньших или равных z ;

m – число моментов изменения состояний объекта больших z и всегда справедливо равенство: $k + m = n$, a – коэффициент затрат.

Так как величина затрат меняется в зависимости от момента контроля, требуется определить такую точку z^* , при которой глобально минимизируется функция $S_0(t, z, B, a)$. Данную задачу можно рассматривать как задачу оптимизации:

Для заданных t , a найти z^* , такую, что $S(t, z^*, B, a) = \min_z [S(t, z, B, a)]$.

Функция $S_0(t, z, B, a)$ имеет несколько полезных свойств: значения суммарных затрат в точках $z = 0$ и $z = B$ имеют одинаковые максимальные значения, расчет затрат инвариантен относительно положения элементов вектора t , значения суммарных затрат распределяются по уникальным элементам вектора состояний, за счет чего отсутствует необходимость определения суммарных затрат для повторных элементов t . Но при этом, в точках $z = t_i$ ($i = 1, 2, \dots, n$), исходя из формулы (1), непрерывная по z функция $S_0(t, z, B, a)$ имеет скачкообразные разрывы первого рода. Рассмотрим произвольную точку t_i .

Пусть $z = t_i - 0$, тогда $S_0(t, z, B, a) = a \sum_{j=1}^{i-1} (z - t_j) + a \sum_{j=1}^n (B - t_j)$,

если $z = t_i$, то $S_0(t, z, B, a) = a \sum_{j=1}^{i-1} (z - t_j) + a \sum_{j=i+1}^n (B - t_j)$,

наконец, если $z = t_i + 0$, то $S_0(t, z, B, a) = a \sum_{j=1}^i (z - t_j) + a \sum_{j=i+1}^n (B - t_j)$.

Пример: $t = (t_1, t_2, t_3)$, где $t_1 = 3$, $t_2 = 6$, $t_3 = 9$, $B = 20$, $a = 1$. Ниже приведены результаты расчетов:

1) при $z_1 = t_2 - s$ и $S_0(t, z, 20, 1) = (6 - s - 3) + (20 - 6) + (20 - 9) = 28 - s$;

2) при $z_2 = t_2$ и $S_0(t, z, 20, 1) = (6 - 3) + (20 - 6) = 17$;

3) при $z_3 = t_2 + s$ и $S_0(t, z, 20, 1) = (6 - 3) + (6 + s - 6) + (20 - 9) = 14 + s$;

В результате, при $s \rightarrow 0$ получается $S_0(t, z_1, 20, 1) = 28$, $S_0(t, z_2, 20, 1) = 17$, $S_0(t, z_3, 20, 1) = 14$.

Далее проведены исследования зависимости минимума функции $S_0(t, z, B, a)$ от значений параметров n , B , a и вектора t .

В этом случае, для заданного вектора t при любом значении его размера задача поиска z^* (2) решается при помощи разработанного программного комплекса моделирования контроля динамического объекта.

В таблице 1, приведены результаты расчетов z^* , при $n = 6$ или $n = 7$, для различных значений a при $B = 24$ и $B = 36$.

Таблица 1. Результаты расчетов

$B = 24, n = 6$				$B = 24, n = 7$			
a	t	z^*	$S_0(t, z, B, a)$	a	t	z^*	$S_0(t, z, B, a)$
0.1	(1,3,4,7,9,10)	10	2.6	0.1	(1,3,4,7,9,10, 23)	10	2.7
0.9	(1,15,16,20,21,22)	16	22.5	0.9	(1,15,16,20,21,22,23)	16	23.4
1.8	(3,7,9,10,14,17)	10	50.4	1.8	(3,7,9,10,14,17,19)	10	59.4
2.3	(17,18,19,20,21,22)	20	25.3	2.3	(17,18,19,20,21,22,23)	20	27.6
3.1	(1,9,10,14,19,20)	10	89.9	3.1	(1,9,10,14,19,20,23)	10	93.0
5.1	(11,14,15,19,20,21)	15	86.7	5.1	(11,14,15,19,20,21,22)	15	96.9
7.0	(1,3,5,7,9,11)	11	210.0	7.0	(1,3,5,7,9,11,13)	11	287.0
$B = 36, n = 7$				$B = 36, n = 7$			
0.1	(1,3,4,7,9,10)	10	2.6	0.1	(1,3,4,7,9,10, 23)	10	3.9
0.9	(10,15,16,20,21,22)	22	33.3	0.9	(1,15,16,20,21,22,23)	23	38.7
1.8	(3,7,9,10,14,17)	17	75.6	1.8	(3,7,9,10,14,17,19)	19	97.2
2.3	(17,18,19,20,21,22)	22	34.5	2.3	(17,18,19,20,21,22,23)	23	48.3
3.1	(1,9,10,14,19,20)	20	145.7	3.1	(1,9,10,14,19,20,23)	20	186.0
5.1	(11,14,15,19,20,21)	21	132.6	5.1	(11,14,15,19,20,21,22)	22	163.2
7.0	(1,3,5,7,9,11)	11	210.0	7.0	(1,3,5,7,9,11,13)	13	294.0

Приведенные в таблице 1 результаты показывают, что на величину затрат в глобально оптимально выбранном моменте контроля, в основном, влияют длительность функционирования объекта (B) и количество изменений (n).

Было проведено моделирование для различных значений a , $n = 50$, $B = 120$ для различных N вариантов вектора t . По результатам моделирования были вычислены среднее значение $E[z^*]$, значение $S_0(z^*, B, a)$ и среднее значение $S_0(E[z^*], B, a)$ по всем вариантам z^* . Результаты моделирования приведены в таблице 2.

Таблица 2. Результаты моделирования

N	a	$E[z^*]$	$S_0(z^*, B, a)$	$S_0(E[z^*], B, a)$
10000	0.1	57.063489179712896	136.85779953062976	150.38852086852197
10000	0.5	57.21846216112892	684.8666273442763	752.3913220647346
10000	0.9	57.10725455130851	1233.661497931484	1353.2278217173655
50000	0.1	57.03376017125462	136.908602923376	150.39622467673252
50000	0.5	57.16275782938732	684.740942013988	751.9479773097805
50000	0.9	57.04868652012312	1231.9569917087476	1352.9245141513748

Из результатов, приведенных в таблице 2 следует, что среднее значение величины затрат $S_0(z^*, B, a)$ не совпадает со значением величины затрат при среднем значении оптимальной точки контроля $S_0(E[z^*], B, a)$, что связано с нелинейностью функции затрат. Отсюда следует, что задача оптимизации должна решаться для каждого конкретного варианта вектора t . Если модель должна в режиме реального времени «учиться» автоматически выбирать оптимальную точку контроля, то алгоритм обучения должен иметь возможность адаптации к изменяющимся значениям моментов «неправильного» состояния системы. В качестве примера такого алгоритма можно привести проблему многорукого бандита.

Результаты в таблице 2 были получены при моделировании вектора t , состоящего из элементов равномерно распределенных на $(0, B)$, и отсортированных в порядке возрастания. Далее исследована зависимость средних суммарных затрат от значений параметров a , n и B .

Для этого модель затрат преобразована в модель средних затрат, зависящую от положения точки z . Учитывая особенности равномерного распределения сделать можно следующим образом (3):

$$k + m = n, kz \rightarrow \frac{z}{B}z \rightarrow \frac{z^2}{B}, mB \rightarrow \left(1 - \frac{z}{B}\right)B, \sum_{i=1}^n t_i \rightarrow \frac{nB}{2} \quad (3)$$

$$M[S(t, z, B, a)] = a \left[\frac{nz^2}{B} + \left(n - \frac{nz}{B}\right)B - \frac{nB}{2} \right] = an \left[\frac{z^2}{B} + \left(1 - \frac{z}{B}\right)B - \frac{B}{2} \right] \quad (4)$$

Для определения точки минимума получившейся функции (4), решается уравнение (5):

$$M'[S(t, z, B, a)] = an \left[\frac{2z}{B} - 1 \right] = 0 \rightarrow \frac{2z}{B} - 1 = 0 \rightarrow \frac{2z}{B} = 1 \rightarrow z = \frac{B}{2} \quad (5)$$

Таким образом, средние суммарные затраты достигаются в точке $z^* = \frac{B}{2}$, и равны:

$$M \left[S \left(t, \frac{B}{2}, B, a \right) \right] = an \left[\frac{B^2}{4B} + \left(1 - \frac{B}{2B}\right)B - \frac{nB}{2} \right] = an \left[\frac{B}{4} + B - \frac{B}{2} - \frac{B}{2} \right] = \frac{anB}{4} \quad (6)$$

Из таблицы 2 видно, что средняя точка оптимальности $E[z^*]$ сходится к половине интервала проверки $B/2$. Теперь сравним ожидаемые значения средних суммарных затрат в средней точке глобальной оптимальности с полученными $S_0(E[z^*], B, a)$.

Таблица 3. Сравнение ожидаемых значений средних суммарных затрат с полученными

N	a	$E[z^*]$	$B/2$	$S_0(E[z^*], B, a)$	$anB/4$
10000	0.1	57.063489179712896	60	150.38852086852197	150
10000	0.5	57.21846216112892	60	752.3913220647346	750
10000	0.9	57.10725455130851	60	1353.2278217173655	1350
50000	0.1	57.03376017125462	60	150.39622467673252	150
50000	0.5	57.16275782938732	60	751.9479773097805	750
50000	0.9	57.04868652012312	60	1352.9245141513748	1350

Из таблицы 3 видно, что между ожидаемыми средними суммарными затратами в средней точке оптимальности и полученными не наблюдается сильных отклонений.

Анализ результатов проведенных исследований, позволил сделать следующие выводы:

- оптимальное значение z^* наиболее сильно зависит от вектора t и величины B ;

- значение $z^* = B/2$, часто предлагаемое в качестве решения, не позволяет получить реальное минимальное значение суммарных затрат на контроль, так как отсутствует сходимость величины $S_0(z^*, B, a)$ к ожидаемым средним $anB/4$;

- величина B может сильно влиять на результаты оптимизации;

Следовательно, при известном векторе t и параметрах B и a требуется каждый раз решать задачу оптимизации.

Заключение

На практике, для многих объектов вектора t может быть либо детерминированным (элементы известны заранее) или стохастическим образом (генерация случайного вектора с элементами конкретного распределения).

Решение задачи оптимизации может проводиться методом простого перебора компонент вектора t , однако, при большом его размере потребуется время. Алгоритм перебора имеет квадратичную сложность.

В более общем случае целесообразно проведение исследований, когда компоненты t являются случайными величинами.

В этом случае требуется провести сбор статистики, что, во-первых, не всегда возможно, а, во-вторых, требует много времени и проверки выполнения таких условий, например, как однородность выборки.

Если оптимизируется средняя величина затрат, возможно применение методов стохастической аппроксимации [6]. Такой подход позволяет сократить затраты на сбор статистики, однако требует большего времени для получения результата, поскольку скорость сходимости к результату может быть достаточно малой.

Библиографический список

1. Посохов М.А. Обеспечение актуальности данных в системах больших данных / М.А. Посохов, Е.А. Саксонов // Технологии информационного общества. — 2024. — № 18. — 122-124 с.
2. Посохов М.А., Саксонов Е.А. Детерминированная модель контроля динамических объектов // Телекоммуникационные и вычислительные системы. Сборник трудов тридцать второго международного научно-технического форума. Москва, 2024 С. 122-124.
3. Леонтьев А.С., Тимошкин М.С. Методика выбора контрольных точек при решении функциональных задач // Форум молодых ученых. — 2023 — № 4 (80).
4. Леонтьев А.С., Тимошкин М.С. Разработка методики выбора контрольных точек в вычислительных комплексах со структурной избыточностью // Наукосфера. — 2023 — № 4 (2). — DOI: 10.5281/zenodo.7860606.
5. Марон, М. А., Модели, методы и алгоритмы управления контролем качества в динамике сложных процессов. Москва, 2019 — 146 с.
6. Вазан М. Стохастическая аппроксимация. М.: Мир, 1972, 296с.

УДК 621.396; ГРНТИ 47.47

ПРОГРАММНЫЙ КОМПЛЕКС МОДЕЛИРОВАНИЯ ПРОЦЕССА КОНТРОЛЯ ДИНАМИЧЕСКИХ СИСТЕМ

Е.А. Саксонов, М.А. Посохов

*Московский технический университет связи и информатики,
Российская Федерация, Москва, Estu.d.379@mail.ru, mikhail.posokhov@yandex.ru*

Аннотация. В работе рассматривается важность обеспечения контроля динамических объектов, а именно своевременным отслеживанием моментов плохого состояния системы. Приводятся функциональные требования к программному комплексу моделирования процесса контроля динамических объектов для обозначения задач, которые должен он уметь решать. Приводится графическая структурная схема программного комплекса и подробно описываются составляющие его части. Рассмотрены примеры применения программного комплекса к входным значениям векторов моментов плохого состояния и моментов контроля и параметров динамической системы.

Ключевые слова: случайные процессы, теория надежности, качество данных, математическое моделирование, постановка контрольных точек, математическое обеспечение вычислительных систем

THE SOFTWARE COMPLEX OF MODELLING OF DYNAMIC SYSTEM CONTROL

E.A. Saksonov, M.A. Pososkhov

*Moscow Technical University of Communications and Informatics,
Russia, Moscow, Estu.d.379@mail.ru, mikhail.posokhov@yandex.ru*

The summary. This paper examines the importance of management of control of dynamic objects, specifically the timely detection of system failures. Functional requirements for a software package for modeling the dynamic object control process are presented, outlining the tasks it must be able to solve. A graphical structural diagram of the software package is provided, along with a detailed description of its components. Examples of the software package's application to

input values of failure and control moment vectors and dynamic system parameters are discussed.

Keywords: stochastic processes, reliability theory, data quality, mathematical modeling, checkpoint set, mathematical software for computing systems

Обеспечение качества функционирования динамических систем является актуальной проблемой [1]. Особенностью таких систем является изменение состояния объектов системы в течение времени [2]. В связи с этим требуется создание средств для исследования изменения состояния систем, которые позволят быстро обнаруживать изменения, сокращая время пребывания системы в недопустимом состоянии.

Сложность задачи заключается в том, что изменения состояния носят случайный характер, имеют различные причины и могут быть взаимосвязаны. Применение аналитических методов требует сбора и обработки больших объемов данных для разработки моделей, обладающих требуемой точностью.

Одним из возможных решений является применение метода статистического моделирования для анализа принимаемых решений. Появляется возможность изучать поведение виртуальной системы без её реального аналога, обеспечиваются, в целом, удовлетворительные результаты, которые могут уточняться при проведении испытаний и сборе дополнительной статистики.

Программные комплексы моделирования, как правило, являются достаточно универсальными, ориентированными на решение задач формирования и обработки требуемой статистики, поэтому их построению требует проведение предварительных исследований для настройки на решаемые задачи.

Функциональные требования к программному комплексу

Функциональные требования – это описание поведения системы в определенных условиях [3]. Функциональные требования определяют конкретные действия, которые система должна выполнять при взаимодействии с пользователем, описывая поведение ПО по принципу «если пользователь делает X, система выдает результат Y» [4].

Применение метода статистического моделирования требует разработки соответствующего программного обеспечения, которое целесообразно объединить в специализированный программный комплекс. В связи с этим возникает задача формирования программного комплекса включает определение состава и структуры, методов взаимодействия программных модулей (подсистем), средств программирования, отладки и тестирования.

Отличительной особенностью данного комплекса является возможность сбора и обработки данных о работе реальной системы для формирования требуемых статистических моделей контроля состояний системы.

Также комплекс должен моделировать процесс контроля системы и вычислять характеристики при заданных параметрах контроля.

Описание процесса контроля

Рассматривается динамическая система, функционирующая на интервале $(0, B)$, $0 < B < \infty$. Состояние системы в каждый момент времени может быть либо «правильным» либо «неправильным». Возможны переходы из «правильного» состояния в «неправильное», которые происходят в случайные моменты времени, упорядоченное множество таких моментов $t(n) = (t_1, t_2, \dots, t_n)$. Для обнаружения «неправильных» состояний в определенные моменты времени проводится контроль состояния системы, упорядоченное множество моментов (точек) контроля $z(k) = (z_1, z_2, \dots, z_k)$. Затраты, связанные с пребыванием системы в «не-

правильном» состоянии в течение единицы времени, равны a единицам. Суммарные затраты за время $B - S_0(B, t(n), a, z(k))$.

Затраты на создание и эксплуатацию точки контроля с учетом мгновенного приведения системы при обнаружении «неправильного» состояния в «правильное», равны с единицам. Суммарные затраты за время $B - S_1(B, b, k)$, где b - затраты на эксплуатацию точки контроля.

Суммарные затраты на контроль $S(t(n), z(k), a, b)$ складываются из затрат на пребывание системы в «неправильном» состоянии, которые пропорциональны длительности пребывания, и затрат на проведение контроля, которые зависят от количества точек (моментов) контроля, т.е. $S(t(n), z(k), a, b) = S_0(B, t(n), a, z(k)) + S_1(B, b, k)$.

Очевидно, что с увеличением количества точек контроля k , задаваемых вектором $z = (z_1, z_2, \dots, z_k)$, возможно снижение затрат $S_0(B, t(n), a, z(k))$, однако увеличиваются затраты $S_1(B, b, k)$.

Возможны различные варианты формирования множества $t(n)$ и различные варианты определения моментов контроля, связанные с особенностями динамической системы. В связи с этим построение универсальной и достаточно точной математической модели вычисления затрат на контроль, представляет значительные трудности и метод моделирования представляется приемлемым вариантом решения задачи определения моментов контроля и оценки затрат на контроль.

Задачи комплекса

1. Сбор и обработка статистики о работе реальных систем для определения состояний системы, формирования множества $t(n)$, определения значений параметров a и b , возможного количества точек контроля. Множество $t(n)$ является множеством моментов появления событий заданного потока «неправильных» состояний на интервале $(0, B)$. Множество $t(n)$ – упорядоченное множество случайных величин на интервале $(0, B)$ с заданным законом распределения.
2. Генерация множества моментов контроля $z(k)$. Данное множество является множеством событий (независимых или связанных) на интервале $(0, B)$. Оно является множеством случайных величин (возможно коррелированных) с заданными распределением на интервале $(0, B)$.
3. Расчет суммарных затрат на проведение контроля при заданных множествах $t(n)$ и $z(k)$.
4. Проведение цикла моделирования из M испытаний и сбор статистики по затратам при различных вариантах множеств $t(n)$ и $z(k)$.
5. Обработка статистики: вычисление среднего значения и дисперсии величины суммарных затрат за цикл моделирования. Исследование зависимости величины суммарных затрат от параметров множеств $t(n)$ и $z(k)$.
6. Получение рекомендаций для практического применения результатов.

Состав комплекса программного обеспечения

В соответствии с поставленными ранее функциональными требованиями к разрабатываемому программному комплексу он состоит из следующей системы взаимодействующих модулей:

1. **Модуль API генерации данных.** Используются три функции генерации данных. Библиотека языка *Python numpy* в качестве алгоритма генерации псевдослучайных чисел использует Вихрь Мерсена (англ. *MersenneTwister*), который обладает перио-

дом, равным $2^{19937} - 1$, что является достаточно огромным значением, описанный в работах [3-4]. Реализует функциональные требования

2. **API расчета суммарных затрат на контроль системы и эксплуатацию точек контроля.** Используются функции для расчета суммарных затрат на контроль системы и затрат на эксплуатацию точек контроля, поиска точки/точек оптимальности.
3. **API проведения статистических испытаний методом.** Содержит функции для проведения испытаний с целью оценки и анализа параметров наиболее важных показателей обеспечения качества данных.
4. **API визуализации результатов экспериментов.** Содержит функции для проведения графической визуализации результатов экспериментов и вычислений.
5. **Форма для проведения расчета суммарных затрат на контроль системы.** В этой форме пользователь вводит предварительное значение вектора t или генерирует его случайным образом, а также значения параметров точки контроля z , предела интервала проверки B и параметров затрат a и b .
6. **Форма для проведения расчета затрат на контроль системы и относительных/фиксированных затрат на эксплуатацию точек контроля.** Здесь пользователь вводит предварительное значение векторов t и z или генерирует их случайным образом в случае неавтоматического поиска точек оптимальности. В противном случае, пользователь вводит количество подынтервалов в случае автоматического поиска точек глобальной оптимальности.
7. **Формы для проведения статистических испытаний.** В этих формах пользователь вводит количество испытаний, предварительное значение векторов t и z или генерирует их случайным образом в случае неавтоматического поиска точек глобальной оптимальности. В противном случае, пользователь вводит количество подынтервалов в случае автоматического поиска точек глобальной оптимальности. Также вводятся значения предела интервала проверки B и множителя затрат a .
8. **Форма по сбору и обработке статистики.** В этой форме проводятся сбор и обработка статистики, необходимые для проведения подробного анализа затрат на контроль динамической системы с определенными параметрами B , a и b .

На рисунке 1 приведена структурная схема программного комплекса.



Рис. 1. Структурная схема программного комплекса

Примеры применения программного комплекса

Таблица 1. Для заданных вектора t , и значения параметров B и a . вычислены значения значения суммарных затрат в точке Z .

Вектор t	Параметры Z , B и a	Средние суммарные затраты результат
[1, 3.5, 7, 11, 20]	$z = 8$, $B = 25$, $a = 2$	63.0
[1, 3.5, 7, 11, 20]	$z = 8$, $B = 25$, $a = 4$	126.0
[3.1, 4.6, 7.9, 11.1, 13.3]	$z = 4.5$, $B = 14.5$, $a = 1.5$	33.75
[]	$z = 3$, $B = 100$, $a = 1$	0
[1.5, 3.2, 4.1, 7.8, 9.1]	$z = 9.5$, $B = \infty$, $a = 5$	109.0

Таблица 2. Для заданных векторов t и Z и значений параметров B , a , b вычислены значение суммарных затрат для вектора Z .

Векторы t и Z	Другие параметры	Выходной результат
$t = [5, 6.1, 9, 11, 14.5, 17, 21, 25.1]$ $z = [1, 6.5, 7, 17, 20]$	$B = 30$, $a = 2$, $b = 1$	76.6
$t = [5, 6.1, 9, 11, 14.5, 17, 21, 25.1]$ $z = [0, 1, 6.5, 7, 17, 20, 25]$	$B = 30$, $a = 2$, $b = 1$	61.6
$t = [11, 12, 20.1, 30.4, 40.2, 57.4, 65.2]$ $z = [8, 12, 43.2, 54.3, 66]$	$B = 66$, $a = 10$, $b = 4$	513.0

Заключение

Созданный комплекс позволяет проводить предварительные испытания различных вариантов расстановки контрольных точек, выбирая оптимальны по средним суммарным затратам.

Комплекс может быть адаптирован для исследования различных динамических систем.

Результаты могут быть полезны администраторам при выборе политики контроля состояний систем.

Библиографический список

1. Посохов М.А. Обеспечение актуальности данных в системах больших данных / М.А. Посохов, Е.А. Саксонов // Технологии информационного общества. — 2024. — № 18. — 122 -124 с.
2. Посохов М.А., Саксонов Е.А. Детерминированная модель контроля динамических объектов // Телекоммуникационные и вычислительные системы. Сборник трудов тридцать второго международного научно-технического форума. Москва, 2024 С. 122-124.
3. Вигерс К. Разработка требований к программному обеспечению, М. : Русская редакция, 2004. 576 с.
4. Леффингуэлл Д., Уидриг Д. Принципы работы с требованиями к программному обеспечению. Унифицированный подход. М.: Вильямс, 2002. 448 с.
5. Makoto Matsumoto, Takuji Nishimura. 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. 8, 1 (Jan. 1998), 3–30. <https://doi.org/10.1145/272991.272995>.
6. Паньков М.А., Вавилов В.А. Исследование генераторов псевдослучайных чисел // Материалы IX-й Международной научной конференции "Математическое и программное обеспечение информационных, технических и экономических систем", Томск, 26-28 мая 2022 г. Томск: Изд-во Том. ун-та, 2022. С. 295–298. DOI: 10.17223/978-5-907572-27-0-2022-44

УДК 004.9; ГРНТИ 28.23

НАСТРОЙКА ПАРАМЕТРОВ АЛГОРИТМА РАСМАР В ПРОСТРАНСТВЕ ПРОЕКТИРОВАНИЯ ПЕРЕМЕННОЙ РАЗМЕРНОСТИ

Н.А. Демидов

*МИРЭА – Российский технический университет,
Российская Федерация, Москва, nick.a.demidov@rambler.ru*

Аннотация. В работе рассматривается подход к решению задачи настройки параметров алгоритма PaCMAР. Эта задача относится к задачам смешанной оптимизации в пространстве проектирования переменной размерности. Предлагается для решения задачи настройки параметров алгоритма PaCMAР использовать генетический алгоритм, кодирующий в хромосоме гены, описывающие параметры разных типов, с использованием кода Грея. Кроме того, предлагается ввести в использование дополнительный параметр, называемый размерной переменной, и использовать его для кодирования в гене информации о размерности пространства проектирования. Значение, хранящееся в гене размерной переменной, определяет, какие гены в хромосоме являются активными, при этом остальные гены считаются скрытыми. Такой подход к кодированию хромосомы в генетическом алгоритме позволяет переключаться в процессе оптимизации между пространствами проектирования разной размерности.

Ключевые слова: алгоритм PaCMAР, пространство проектирования переменной размерности, размерная переменная, генетический алгоритм.

PARAMETER TUNING FOR THE PACMAP ALGORITHM IN A VARIABLE-SIZE DESIGN SPACE

N.A. Demidov

*MIREA – Russian technological university,
Russia, Moscow, nick.a.demidov@rambler.ru*

The summary. This paper examines an approach to solving the parameter turning problem for the PaCMAР algorithm. This problem belongs to the class of mixed optimization problems in a variable-size design space. It is proposed to use a genetic algorithm to solve the PaCMAР parameter tuning problem, encoding genes describing parameters of different types in a chromosome using Gray code. Additionally, it is proposed to introduce an additional parameter, called the dimensional variable, and use it to encode information about the dimensionality of the design space within a gene. The value stored in the dimensional variable gene determines which genes in the

chromosome are active, while the remaining genes are considered hidden. This chromosome encoding approach in the genetic algorithm allows for switching between design spaces of different sizes during the optimization process.

Keywords: PaCMAP algorithm, variable-size design space, dimensional variable, genetic algorithm.

В настоящее время предложено большое количество метаэвристических алгоритмов оптимизации, позволяющих эффективно решать практические задачи оптимизации в пространстве высокой размерности с целевыми функциями, характеризующимися сложным ландшафтом [1]. При этом зачастую классические методы оптимизации могут быть не применимы к решению таких задач, например, если целевая функция в явном виде не зависит от оптимизируемых параметров, и, следовательно, не может быть по ним продифференцирована.

Большинство метаэвристических алгоритмов оптимизации устроены так, что работают с переменными только одного типа, например, только с непрерывными или только с дискретными переменными, в то время как реальные практические задачи оптимизации предполагают одновременную работу с переменными разных типов: в таких задачах одни переменные могут быть, например, непрерывными, а другие – дискретными и категориальными. Кроме того, возможны ситуации, когда задачи оптимизации должны решаться в пространстве проектирования переменной размерности [variable-size design space (VSDS)]. Такие задачи, в частности, могут возникать в сфере машинного и глубокого обучения, например, при проектировании моделей нейронных сетей. В связи с этим предлагаются различные модификации базовых метаэвристических алгоритмов, позволяющих решать задачи смешанной оптимизации в пространстве переменной размерности [2]. В частности, известны подходы к оптимизации, в которых значения переменных разных типов оптимизируются одним метаэвристическим алгоритмом, например, генетическим алгоритмом, и подходы, в которых за оптимизацию переменных разных типов отвечают разные метаэвристические алгоритмы. При этом предложены различные подходы к решению задачи оптимизации в пространстве проектирования переменной размерности, в частности, в ряде работ предложено использовать генетический алгоритм, в котором в хромосому вводится дополнительный ген, называемой размерной переменной. Этот ген используется для кодирования информации о размерности пространства проектирования. Значение, хранящееся в гене размерной переменной, определяет, какие гены в хромосоме являются активными, при этом остальные гены считаются скрытыми. Такой подход к кодированию хромосомы в генетическом алгоритме позволяет переключаться в процессе оптимизации между пространствами проектирования разной размерности.

В предлагаемой работе предлагается решение задачи настройки параметров алгоритма PaCMAP в пространстве проектирования переменной размерности с применением генетического алгоритма, в котором в хромосому вводится размерная переменная, а кодирование хромосом выполняется с использованием кода Грея с целью обеспечения возможности работы с параметрами разных типов.

Аспекты работы алгоритма PaCMAP

Алгоритм PaCMAP (Pairwise Controlled Manifold APproximation) [3] является одним из современных алгоритмов нелинейного снижения размерности, который, в отличие от других, таких как, например, t-SNE (t-distributed Stochastic Neighbor Embedding) [4], UMAP (Uniform Manifold Approximation and Projection) [5], TriMAP (Triplets Manifold Approximation Projection) [6], решает, как утверждают его авторы, проблему одновременно сохранения локальных и глобальных структур многомерных наборов данных при выполнении их вложения в пространство низкой размерности.

Алгоритм PaСМАР оперирует в своей работе парами точек трёх типов, такими как близкие, средне-близкие и дальние пары точек, и использует функцию потерь *Loss*, минимизация которой выполняется с применением алгоритма Adam в три этапа с соответствующей настройкой весовых коэффициентов слагаемых функции потерь, соответствующих трем типам пар точек.

Авторы алгоритма PaСМАР применяют алгоритм PCA (Principal Component Analysis) для инициализации вложения многомерных данных в пространство низкой размерности. Кроме того, авторы алгоритма PaСМАР применяют алгоритм PCA к точкам данных перед построением графа k ближайших соседей, считая, что это позволяет без существенной потери точности значительно ускорить алгоритм PaСМАР. При этом авторы алгоритма PaСМАР обеспечивают возможность применением многомерного нормального распределения для случайной инициализации начального вложения. Авторы алгоритма PaСМАР считают, что алгоритм достаточно (но не полностью) устойчив к выбору инициализации благодаря использованию пар средне-близких точек и динамической структуры графа.

В работе [7] выполнено исследование различных сценариев предварительной обработки данных и инициализации вложений при реализации алгоритма PaСМАР, показавшее перспективность применения, и других алгоритмов нелинейного снижения размерности. В этой работе применение каждого алгоритма нелинейного снижения размерности выполнялось индивидуально, при этом одни алгоритмы требовали настройки каких-либо параметров, а другие – нет. В связи с этим задачу исследования различных сценариев предварительной обработки данных и инициализации вложений при реализации алгоритма PaСМАР можно сформулировать как задачу оптимизации в пространстве проектирования переменной размерности.

Кодирование хромосомы для задачи оптимизации в пространстве проектирования переменной размерности

В настоящей работе в качестве алгоритмов, перспективных (с учётом работы [7]) для использования в составе модифицированного алгоритма PaСМАР, выбраны алгоритмы PCA, MSD (Multidimensional Scaling), Isomap (Isometric Feature Mapping), SE (Spectral Embedding), UMAP (Uniform Manifold Approximation and Projection), t-SNE (t-distributed Stochastic Neighbor Embedding) и TriMAP (Triplets Manifold Approximation Projection).

При включении этих алгоритмов в составе алгоритма PaСМАР необходимо учитывать, что:

- алгоритмы PCA и MSD не требуют настройки параметров;
- алгоритмы Isomap, SE и UMAP требуют настройки параметра $n\_neighbors$, задающего число ближайших соседей и равного соответственно 5 , $\max(\text{int}(n/10), 1)$ (n – число объектов в наборе данных) и 15 для указанных алгоритмов;
- алгоритм t-SNE требует настройки параметра *perplexity*, сопоставимого с числом ближайших соседей $n\_neighbors$ и равного 30 ;
- алгоритм TriMAP требует настройки параметра $n\_inliers$, задающего число ближайших соседей для формирования триплетов ближайших соседей и равного 12 .

При этом все вышеназванные параметры являются целочисленными.

Пусть при работе с алгоритмом PaСМАР необходимо настроить такие его параметры, как параметр $n\_neighbors$, задающий число ближайших соседей (по умолчанию $n\_neighbors = 10$), а также параметры MN_{ratio} , определяющий отношение числа средне-близких пар к числу близких пар (по умолчанию $MN_{ratio} = 0.5$) и FP_{ratio} – отношение числа дальних пар к числу близких пар (по умолчанию $FP_{ratio} = 2$). Параметр $n\_neighbors$ является целочисленным, параметры MN_{ratio} и FP_{ratio} являются вещественными. Кроме

того, в модифицированном алгоритме PaСMAP необходимо вместо логического параметра *apply_PCA* [8], отвечающего за применение алгоритма PCA к данным перед построением графа k-ближайших соседей, настроить логический параметр *apply_algorithm*, который принимает значение *False* или *True*. Следует отметить, что авторы алгоритма PaСMAP предлагают применять алгоритм PCA, если размерность пространства больше, чем 100. Параметр *apply_algorithm* можно в дальнейшем рассматривать как целочисленный, принимающий значение 0 или 1. Пусть параметр *init* в алгоритме PaСMAP, отвечающий за выбор алгоритма инициализации, всегда инициализируется названием одного из алгоритмов PCA, MSD, Isomap, SE, UMAP, t-SNE и TriMAP. Названия этих алгоритмов будут ассоциированы со значением размерной переменной в хромосоме.

С учетом вышесказанного, кодирование хромосомы может быть выполнено следующим образом.

Первый ген хромосомы кодирует логический параметр *apply_PCA*.

Второй ген хромосомы кодирует логический параметр *apply_MDS*.

Третий и четвертый гены хромосомы кодируют соответственно логический параметр *apply_Isomap* и параметр $n\_neighbors_{Isomap}$.

Пятый и шестой гены хромосомы кодируют соответственно логический параметр *apply_SE* и параметр $n\_neighbors_{SE}$.

Седьмой и восьмой гены хромосомы кодируют соответственно логический параметр *apply_UMAP* и параметр $n\_neighbors_{UMAP}$.

Девятый и десятый гены хромосомы кодируют соответственно логический параметр *apply_t – SNE* и параметр $perplexity_{t-SNE}$.

Одиннадцатый и двенадцатый гены хромосомы кодируют соответственно логический параметр *apply_TRIMAP* и параметр $n\_inliers_{TRIMAP}$.

Тринадцатый, четырнадцатый и пятнадцатый гены хромосомы кодируют соответственно параметры MN_{ratio} , FP_{ratio} и $n\_neighbors_{PaСMAP}$.

Шестнадцатый ген хромосомы кодирует размерную переменную *dimensional_variable*. При этом при кодировании используются целочисленные значения из множества {0, 1, 2, 3, 4, 5, 6}, сопоставленные алгоритмам PCA, MSD, Isomap, SE, UMAP, t-SNE и TriMAP.

При кодировании логических параметров используются целочисленные значения из множества {0, 1}, сопоставленные *False* и *True*.

При кодировании параметров, определяющих $n\_neighbors$ и $n\_inliers$, используются целочисленные значения из множества {5, ..., 20}. При кодировании параметра, определяющего *perplexity*, используются целочисленные значения из множества {10, ..., 40}.

При кодировании параметров MN_{ratio} и FP_{ratio} используются вещественные значения соответственно из диапазонов [0.2, 0.9] и [1.5, 2.5].

Следует отметить, что при таком способе кодирования тринадцатый, четырнадцатый и пятнадцатый гены хромосомы являются *общими*: они относятся непосредственно к алгоритму PaСMAP и используются во всех реализациях алгоритма независимо от того, какой дополнительный алгоритм (PCA, MSD, Isomap, SE, UMAP, t-SNE или TriMAP) еще задействован. Остальные гены хромосомы связаны с конкретными алгоритмами, т.е. являются *индивидуальными (персональными)*.

Выбор указанных границ диапазонов изменения для вещественных и целочисленных параметров обусловлен стремлением избежать переобучения алгоритма PaСMAP, т.к. этап валидации в этом алгоритме не предусмотрен, в связи с чем можно получить очень маленькое значение функции потерь (в результате применения алгоритма Adam), однако результаты вложения в пространство низкой размерности будут плохо отображать реальную ситуацию (будут неинтерпретируемыми). При оценке качества вложения многомерного набора данных в пространство низкой размерности можно использовать дополнительно визуаль-

ный анализ. При наличии известных меток классов объектов результаты вложения набора данных в пространство низкой размерности могут быть приняты (в случае высокого соответствия меток классов объектов результатам вложения) или отвергнуты (в противном случае).

Сформированная таким образом хромосома определяет задачу смешанной оптимизации в пространстве проектирования переменной размерности. Кодирование генов хромосомы может быть выполнено с использованием кода Грея. При этом кодирование вещественных параметров должно выполняться с некоторой точностью ε (например, с точностью $\varepsilon = 10^{-5}$), а для целочисленных параметров при реализации алгоритма оптимизации необходимо выполнять контроль за тем, чтобы значения параметров не выходили за границы диапазона.

Для решения предлагаемой задачи оптимизации может быть использован генетический алгоритм, предполагающий использование кода Грея для кодирования хромосом и адаптированный для работы с переменными разных типов. В данной работе был использован адаптированный к решению задач смешанной оптимизации алгоритм SelfCSHAGA [9], являющийся самоконфигурируемым генетическим алгоритмом оптимизации с адаптацией на основе истории успеха и доказавший своё преимущество над рядом других современных метаэвристических алгоритмов.

Экспериментальные исследования

Экспериментальные исследования были выполнены на языке Python 3.10 в среде Google Colab с применением программной реализации алгоритма PaCMAP [8] и библиотек машинного обучения, реализующих различные дополнительно используемые алгоритмы снижения размерности наряду с алгоритмом PCA, предлагаемым к использованию в алгоритме PaCMAP. В ходе экспериментальных исследований были использованы 3 практических набора данных results_2023, results_2024 и results_2025 [7].

Практические наборы данных results_2023, results_2024 и results_2025 сформированы на основе результатов решения задач в сфере онлайн-обучения студентов 2 курса ИИТ РТУ МИРЭА программированию на языке Python [7]. Эти наборы данных содержат уже обработанную информацию, извлеченную из лог-журналов событий системы «Цифрового ассистента преподавателя» (ЦАП) по завершении весенних семестров в 2023, 2024 и 2025 годах. Число признаков в этих наборах данных соответственно равно 34, 37 и 34, а число точек данных – 1056, 1111 и 1080.

На рисунке 1 представлены результаты визуализации в двухмерном пространстве наборов данных results_2023, results_2024 и results_2025 после применения к ним алгоритма PaCMAP с настройками по умолчанию и алгоритма PaCMAP, настройки которого подобраны с применением метаэвристического алгоритма.

Анализ результатов вложения наборов данных results_2023, results_2024 и results_2025 в двухмерное пространство (рисунок 1) позволяет сделать вывод о целесообразности применения предлагаемого подхода к настройке параметров алгоритма PaCMAP, обеспечивающего одновременный поиск лучших значений параметров в пространствах проектирования разной размерности. Уточненные на основе метаэвристического алгоритма результаты вложения наборов данных results_2023, results_2024 и results_2025 в двухмерное пространство могут быть использованы для выявления компактных групп точек данных с целью более подробного изучения их свойств и принятия решений о наличии сходства поведения в группах студентов, соответствующих выявленным группам точек.

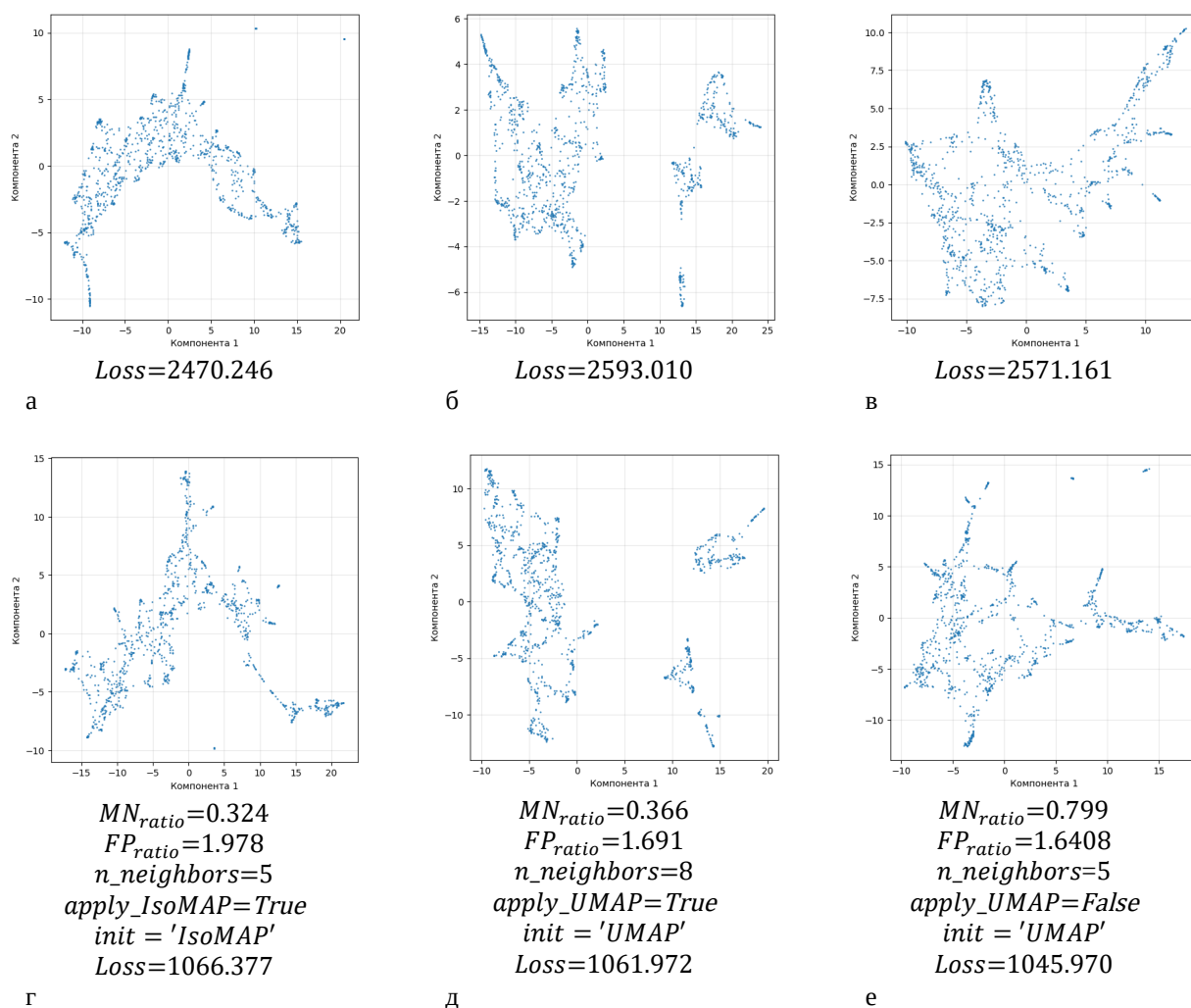


Рис. 1. Результаты визуализации в двухмерном пространстве:

- а – набора данных results_2023 с настройками параметров алгоритма PaCMAP по умолчанию;
 б – набора данных results_2024 с настройками параметров алгоритма PaCMAP по умолчанию;
 в – набора данных results_2025 с настройками параметров алгоритма PaCMAP по умолчанию;
 г – набора данных results_2023 с настройками параметров алгоритма PaCMAP на основе метаэвристического алгоритма;
 д – набора данных results_2024 с настройками параметров алгоритма PaCMAP на основе метаэвристического алгоритма;
 е – набора данных results_2025 с настройками параметров алгоритма PaCMAP на основе метаэвристического алгоритма

Библиографический список

1. Карпенко А. П. Современные алгоритмы поисковой оптимизации: алгоритмы, вдохновленные природой: учебное пособие / А.П. Карпенко. – 2-е изд. – Москва : МГТУ им. Н.Э. Баумана, 2017. – 448 с.
2. Gamot J., Balesdent M., Tremolet A., Wuilbercq R., Melab N., Talbi E.-G. Hidden-variables genetic algorithm for variable-size design space optimal layout problems with application to aerospace vehicles // Engineering Applications of Artificial Intelligence, 2023. – Vol. 121. – P. 105941. <https://doi.org/10.1016/j.engappai.2023.105941>.
3. Wang Y., Huang H., Rudin C., Shaposhnik Y. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization // Journal of Machine Learning Research, 2021. – Vol. 22(201). – P. 1–73.
4. van der Maaten L., Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008. – Vol. 9. – P. 2579–2605.
5. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction // arXiv:1802.03426, 2018. <https://doi.org/10.48550/arXiv.1802.03426>.
6. Amid E., Warmuth M.K. TriMap: Large-scale Dimensionality Reduction Using Triplets // arXiv:1910.00204, 2019. <https://doi.org/10.48550/arXiv.1910.00204>.

7. Андрианова Е.Г., Демидов Н.А. Исследование сценариев предварительной обработки данных и инициализации вложений при реализации алгоритма PaCMAP // ИТ-Стандарт, 2025. – № 4 (45). – С. 102-123.
8. PaCMAP. [Электронный ресурс]. – Режим доступа: <https://github.com/YingfanWang/PaCMAP/tree/master>, свободный (дата обращения: 07.02.2026).
9. Шерстнев П.А., Семенкин Е.С. SelfCSHAGA: Самоконфигурируемый генетический алгоритм оптимизации с адаптацией на основе истории успеха // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение, 2025. – № 2 (151). – С. 122 – 139.

УДК 004.932; ГРНТИ 28.23.15

ОБЗОР СУЩЕСТВУЮЩИХ СИСТЕМ ДЛЯ РАСПОЗНАВАНИЯ ДОРОЖНЫХ ЗНАКОВ И УКАЗАТЕЛЕЙ

Д.Д. Зангин, А.В. Крошилин

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань

Аннотация. В работе рассматриваются существующие коммерческие и исследовательские системы распознавания дорожных знаков и указателей, применяемые в автомобилях и мобильных приложениях. Проанализированы архитектурные особенности решений, используемые технологии компьютерного зрения и искусственного интеллекта, а также их преимущества и ограничения. Рассмотрены встроенные автомобильные системы Traffic Sign Recognition (TSR), решения для автономного транспорта и мобильные приложения. Проведён сравнительный анализ по критериям точности, устойчивости к внешним воздействиям, вычислительным требованиям и области применения.

Ключевые слова: распознавание дорожных знаков, интеллектуальные транспортные системы, ADAS, автономный транспорт, мобильные приложения, компьютерное зрение.

REVIEW OF EXISTING SYSTEMS FOR RECOGNIZING ROAD SIGNS AND INDICATORS

D.D. Zangin, A.V. Kroshilin

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan

The summary. This paper discusses existing commercial and research traffic sign recognition systems used in vehicles and mobile applications. Architectural features, applied computer vision technologies, advantages and limitations are analyzed. Built-in automotive TSR systems, autonomous driving platforms and mobile applications are considered. Comparative characteristics are provided.

Keywords: traffic sign recognition, intelligent transportation systems, ADAS, autonomous driving, mobile applications, computer vision.

Автоматическое распознавание дорожных знаков является важным элементом современных интеллектуальных транспортных систем. Данная технология применяется в системах помощи водителю (ADAS), автономных транспортных средствах и мобильных навигационных сервисах. В отличие от теоретических алгоритмов, практические системы должны обеспечивать высокую точность работы в реальных условиях: при изменении освещения, погодных воздействиях, частичном перекрытии знаков и высокой скорости движения автомобиля [1].

Системы распознавания дорожных знаков являются критически важной составляющей для развития современных технологий помощи водителю и беспилотных автомобилей. Эти системы призваны повысить безопасность дорожного движения, обеспечивая своевременное информирование и поведение автономных систем. Ключевые задачи системы распознавания дорожных знаков и указателей – локализация потенциальных областей с дорожными знаками на изображении и идентификация конкретного типа знака (например, "ограничение скорости", "главная дорога", "стоп") [2].

Сфера распознавания дорожных знаков на основе компьютерного зрения является особенно актуальной в области автономного транспорта. Существует необходимость моделировать, анализировать и осуществлять управление различными элементами предметной области в реализуемой информационной системе [5, 6]. Происходит интенсивное развитие указанной области и смещение фокуса с задач детектирования в идеальных условиях к созданию всепогодных, надежных и энергоэффективных информационных систем, интегрированных в общую архитектуру. Дальнейший прогресс напрямую связан с развитием алгоритмов, способных к непрерывному обучению и адаптации в изменчивой дорожной обстановке [3].

На современном рынке представлено множество решений, различающихся по архитектуре, функциональным возможностям, стоимости и областям применения. С одной стороны, существуют закрытые аппаратно-программные комплексы от крупных производителей автокомпонентов, прошедшие сертификацию и оптимизированные для работы в реальном времени. С другой стороны, активно развиваются открытые программные решения, позволяющие исследователям и разработчикам создавать собственные системы распознавания на базе общедоступных библиотек и фреймворков. Кроме того, формируется отдельный класс инфраструктурных систем, предназначенных для стационарного монтажа и решения задач городского хозяйства.

В данной статье рассматриваются реально существующие системы и программные решения, применяемые в промышленности и потребительских продуктах.

Автомобильная система Traffic Sign Recognition (TSR)

Система Traffic Sign Recognition (TSR) является составной частью комплекса ADAS современных автомобилей. Она основана на использовании фронтальной видеокамеры, установленной в верхней части лобового стекла. Камера непрерывно фиксирует дорожную обстановку, после чего видеопоток поступает в электронный блок управления, где осуществляется анализ изображения с применением алгоритмов компьютерного зрения и нейросетевых моделей.

Распознанные знаки, в частности ограничения скорости, запреты обгона и временные предписания, выводятся на цифровую приборную панель или проекционный дисплей. В ряде автомобилей данные TSR интегрируются с адаптивным круиз-контролем, что позволяет автоматически корректировать скорость движения.

Преимущества системы:

- высокая точность распознавания в реальных дорожных условиях;
- автономность работы без подключения к сети Интернет;
- интеграция с другими системами активной безопасности;
- стабильность функционирования при движении на высокой скорости.

Недостатки системы TSR:

- зависимость качества распознавания от чистоты камеры и лобового стекла;
- снижение точности при сильных осадках и недостаточной освещённости;
- высокая стоимость внедрения, так как система является частью комплектации автомобиля

Приложение Google Maps

Google Maps представляет собой навигационное приложение, которое отображает ограничения скорости и другую дорожную информацию на основе картографических данных и пользовательских отчётов. В большинстве регионов информация о допустимой скорости формируется на основе базы данных цифровых карт, а не прямого распознавания знаков камерой устройства.

Приложение использует данные GPS для определения местоположения транспортного средства и сопоставляет его с информацией о дорожной инфраструктуре. Обновление сведений осуществляется через облачные сервисы, что позволяет поддерживать актуальность данных при наличии подключения к сети.

Преимущества Google Maps:

- широкая географическая доступность;
- регулярное обновление картографической информации;
- интеграция с навигацией и маршрутным планированием;
- отсутствие необходимости в специализированном оборудовании.

Недостатки Google Maps:

- зависимость от интернет-соединения;
- возможная неактуальность информации при временных дорожных изменениях;
- отсутствие полноценного анализа видеопотока в большинстве версий приложения.

Приложение RoadAR

RoadAR является специализированным мобильным приложением, ориентированным непосредственно на распознавание дорожных знаков с использованием камеры смартфона. В отличие от навигационных сервисов, данное приложение анализирует видеопоток в реальном времени и накладывает распознанную информацию поверх изображения в режиме дополненной реальности.

Обработка данных может выполняться как локально на мобильном устройстве, так и с использованием облачных вычислений. Приложение информирует водителя голосовыми и визуальными уведомлениями о действующих ограничениях.

Преимущества RoadAR:

- использование реального видеопотока для распознавания знаков;
- режим дополненной реальности;
- доступность для пользователей без встроенных автомобильных систем.

Недостатки RoadAR:

- зависимость от производительности смартфона;
- чувствительность к вибрациям и положению устройства;
- снижение точности при высокой скорости движения и неблагоприятных погодных условиях;
- ограниченная интеграция с системами безопасности автомобиля.

Обзор существующих систем показал, что автомобильная система TSR представляет собой наиболее технологически интегрированное решение, обеспечивающее высокую точность и стабильность работы благодаря специализированному аппаратному обеспечению и тесной связи с другими системами автомобиля. Она функционирует автономно и рассчитана на эксплуатацию в широком диапазоне дорожных условий.

Google Maps выполняет преимущественно информационную функцию, предоставляя данные о скоростных ограничениях на основе цифровых карт. Приложение отличается масштабируемостью и доступностью, однако не всегда способно оперативно реагировать на временные изменения дорожной обстановки.

RoadAR занимает промежуточное положение между встроенными системами и навигационными сервисами. Приложение действительно осуществляет визуальное распознавание знаков, однако ограничено вычислительными возможностями мобильного устройства и не обладает глубокой интеграцией с транспортным средством.

Сравнительная оценка проанализированных систем распознавания дорожных знаков представлена в таблице 1.

Таблица 3. Сравнительная оценка систем распознавания дорожных знаков

<i>Система</i>	<i>Точность распознавания</i>	<i>Работа в реальном времени</i>	<i>Устойчивость к погодным условиям</i>	<i>Автономность работы</i>	<i>Интеграция с системами автомобиля</i>
Автомобильная система TSR	9	+	7	10	10
Google Maps	7	+	9	6	3
RoadAR	6	+	5	8	1

Согласно проведенному анализу сделаны следующие выводы: TSR является наиболее надёжным решением для обеспечения активной безопасности движения; Google Maps – универсальным навигационным инструментом с функцией информирования; RoadAR – мобильным средством визуального анализа дорожной среды, пригодным для вспомогательного использования.

Библиографический список

1. Winner H., Hakuli S., Lotz F., Singer C. (eds.). Handbook of Driver Assistance Systems: Basic Information, Components and Systems for Active Safety and Comfort. – Cham: Springer, 2016.
2. Соколов М. О., Лобазев Н. А., Федоркова А. О. Проектирование и реализация системы распознавания дорожных знаков с использованием компьютерного зрения // Сборник трудов СПбГЭТУ «ЛЭТИ». – 2023.
3. Зангин Д.Д., Крошила С.В. Обзор методов компьютерного зрения для распознавания дорожных знаков и указателей // Новые информационные технологии в научных исследованиях: материалы XXX Всероссийской научно-технической конференции студентов, молодых ученых и специалистов. Издательство ИП Коняхин А.В. (Book Jet). 2025., т.1, 280 с. (36-38).

УДК 004.891; ГРНТИ 50.41

ГИБРИДНАЯ МОДЕЛЬ ПРОГНОЗИРОВАНИЯ И КЛАССИФИКАЦИИ РИСКОВ ОТКАЗОВ В МУЛЬТИСЕРВИСНЫХ СЕТЯХ НА ОСНОВЕ АНСАМБЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ АЛГОРИТМОВ

С.А. Гушчин

*Воронежский государственный технический университет,
Российская федерация, Воронеж, GuSt-work@yandex.ru*

Аннотация. В работе предлагается гибридная модель прогнозирования и классификации рисков отказов в мультисервисных сетях, основанная на ансамблевом использовании интеллектуальных алгоритмов анализа данных. Рассмотренные ранее методы нейросетевого прогнозирования, классификации рисков на основе алгоритмов машинного обучения и анализа шумов демонстрируют высокую эффективность при решении отдельных задач, однако их изолированное применение ограничивает устойчивость и адаптивность систем управления в условиях зашумлённых и гетерогенных данных. В предлагаемой модели осуществляется интеграция нейросетевых методов анализа временных рядов, ансамблевых алгоритмов классификации и плотностных методов выявления шумов, что позволяет повысить точность и надёжность оценки рисков отказов. Приведена математическая формализация гибридной модели и алгоритм интеграции результатов отдельных модулей.
Ключевые слова. мультисервисные сети, прогнозирование отказов, классификация рисков, гибридные модели, машинное обучение, ансамблевые алгоритмы.

NEURAL NETWORK PREDICTION OF FAILURES IN MULTISERVICE NETWORKS

S.A. Gushchin

*Voronezh State Technical University,
Russian Federation, Voronezh, GuSt-work@yandex.ru*

Abstract. This paper proposes a hybrid model for failure risk prediction and classification in multiservice networks, based on the ensemble application of intelligent data analysis algorithms. Previously investigated methods of neural network-based failure prediction, risk classification using machine learning algorithms, and noise analysis demonstrate high effectiveness in solving individual tasks; however, their isolated application limits the robustness and adaptability of network management systems under conditions of noisy and heterogeneous data. The proposed model integrates neural network methods for time series analysis, ensemble classification algorithms, and density-based noise detection techniques, which makes it possible to improve the accuracy and reliability of failure risk assessment. A mathematical formalization of the hybrid model and an algorithm for integrating the outputs of individual modules are presented.
Keywords: multiservice networks, failure prediction, risk classification, hybrid models, machine learning, ensemble algorithms.

Введение

Современные мультисервисные сети представляют собой сложные гетерогенные системы, обеспечивающие одновременную передачу разнородных видов трафика и функционирующие в условиях динамического изменения нагрузки, структуры и состава сервисов. Повышение сложности таких сетей приводит к росту вероятности возникновения отказов различной природы, включая аппаратные сбои, программные ошибки, деградацию качества обслуживания и инциденты информационной безопасности.

В работах [1–2] были рассмотрены методы нейросетевого прогнозирования отказов, алгоритмизация процедуры классификации рисков отказов, применение ансамблевых алгоритмов машинного обучения, а также интеллектуализация анализа сетевых данных на основе плотностных методов выявления шумов. Полученные результаты подтверждают высокую эффективность указанных подходов при решении частных задач прогнозирования и классификации.

Однако изолированное применение отдельных интеллектуальных алгоритмов не позволяет в полной мере учитывать зашумлённость и неоднородность данных, характерных для мультисервисных сетей, а также снижает устойчивость оценки рисков при изменении эксплуатационных условий. В этой связи актуальной является задача разработки гибридной модели, объединяющей преимущества различных методов анализа данных в рамках единой архитектуры прогнозирования и классификации рисков отказов.

Постановка задачи прогнозирования и классификации рисков отказов

Рассмотрим мультисервисную сеть, состояние которой в дискретные моменты времени $t \in \{1, 2, \dots, T\}$ описывается вектором наблюдаемых метрик:

$$\mathbf{x}_t = (x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(n)}),$$

где $x_t^{(i)}$ – значение i -й метрики (задержка, пропускная способность, загрузка ресурсов, потери пакетов и др.).

На основе временной последовательности наблюдений

$$X_t = \{x_{t-L+1}, \dots, x_t\},$$

где L – длина временного окна, требуется решить две основные задачи:

1. Спрогнозировать вероятность возникновения отказа в последующий момент времени $p^t = P(y_{t+1} = 1 | X_t)$, где $y_{t+1} \in \{0, 1\}$ – индикатор отказа;

2. Классифицировать уровень риска отказа

$$r^t \in R = \{r_1, r_2, \dots, r_K\},$$

где R – множество классов риска (низкий, средний, высокий и др.).

Дополнительно необходимо учитывать наличие шумов и выбросов в исходных данных, которые могут приводить к искажению результатов прогнозирования.

Архитектура гибридной модели

Предлагаемая гибридная модель включает три функциональных модуля и модуль интеграции решений:

- Модуль нейросетевого прогнозирования, реализующий анализ временных зависимостей в данных;
- Модуль классификации рисков, основанный на ансамблевых алгоритмах машинного обучения;
- Модуль анализа шумов, предназначенный для выявления аномальных и недостоверных наблюдений;

- Модуль интеграции, осуществляющий объединение выходов отдельных моделей.

Функционально гибридную модель можно представить в виде отображения:

$$H : X_t \rightarrow (\hat{p}^t, \hat{r}^t),$$

где результат формируется на основе согласованного решения всех модулей.

Математическая формализация гибридной модели

Нейросетевой модуль прогнозирования. Нейросетевой модуль реализует отображение вида:

$$\hat{p}_{\text{NN}}^t = f_{\text{NN}}(X_t; \theta),$$

где $f_{\text{NN}}(\cdot)$ – нейросетевая модель (LSTM или TCN), а θ – вектор обучаемых параметров.

Нейросетевой модуль предназначен для прогнозирования вероятности отказа на следующий временной шаг $t+1$ с использованием исторических значений метрик сети X_t . Модель учитывает временные зависимости в данных и позволяет выявлять скрытые паттерны, которые сложно формализовать с помощью классических статистических методов. Обучение модели осуществляется путём минимизации функции потерь бинарной кросс-энтропии:

$$L = - \sum_{t=1}^T [y_{t+1} \log \hat{p}_t^{\text{NN}} + (1 - y_{t+1}) \log(1 - \hat{p}_t^{\text{NN}})].$$

Функция потерь обеспечивает корректировку весов нейросети таким образом, чтобы предсказанные вероятности максимально соответствовали фактическим событиям отказов, что критически важно для высоконагруженных мультисервисных сетей.

Модуль классификации рисков. Для классификации уровня риска используется ансамблевый алгоритм Gradient Boosting, формирующий модель:

$$\hat{r}_t^{\text{GB}} = \arg \max_k \sum_{m=1}^M \alpha_m h_m(z_t),$$

где $h_m(\cdot)$ – базовые решающие деревья;

α_m – веса моделей;

z_t – вектор агрегированных признаков сети.

Ансамблевый модуль позволяет объединить предсказания множества слабых моделей в один согласованный прогноз уровня риска. Применение Gradient Boosting обеспечивает адаптивное обучение на ошибках предыдущих моделей, повышая точность классификации и снижая вероятность переобучения. Модуль эффективно работает с числовыми и категориальными признаками сети, включая параметры загрузки ресурсов, пропускную способность и задержку, что делает его универсальным инструментом для анализа мультисервисных сетей.

Модуль анализа шумов. Для выявления шумов и выбросов используется плотностной алгоритм кластеризации, формирующий бинарную функцию доверия:

$$w_t = \begin{cases} 1, & x_t \in \text{кластер}, \\ 0, & x_t \in \text{шум} \end{cases}.$$

Коэффициент w_i позволяет корректировать предсказания нейросетевого и ансамблевого модулей, снижая влияние аномальных точек на итоговый прогноз. Такой подход обеспечивает устойчивость модели к зашумлённым и гетерогенным данным, что особенно важно в реальных условиях работы мультисервисных сетей, где выбросы и нестабильные сигналы могут существенно исказить результаты анализа.

Модуль интеграции результатов. Итоговая оценка вероятности отказа формируется как взвешенная комбинация:

$$\hat{p}^i = w_i \cdot \hat{p}_{NN}^i + (1 - w_i) \cdot \bar{p},$$

где $\bar{p}$ – априорная вероятность отказа.

Класс риска определяется на основе совместного анализа $\hat{p}_i$ и $\hat{r}_i^{(GB)}$

Заключение

Разработанная гибридная модель интегрирует преимущества нейросетевого прогнозирования, ансамблевой классификации и плотностного анализа шумов, что позволяет существенно повысить устойчивость и точность оценки рисков отказов в мультисервисных сетях. Использование нейросетевого модуля обеспечивает моделирование сложных временных зависимостей и выявление скрытых закономерностей в сетевых метриках, ансамблевый модуль Gradient Boosting повышает точность классификации рисков за счёт последовательного уменьшения ошибок слабых моделей, а модуль анализа шумов минимизирует влияние аномалий и выбросов на итоговые прогнозы.

Математическая формализация гибридной модели создаёт строгую основу для экспериментальных исследований, позволяя количественно оценивать вклад каждого модуля в общую производительность системы. Предложенный подход открывает возможности для разработки адаптивных систем управления сетевой инфраструктурой, способных динамически перенастраивать параметры прогнозирования и классификации в условиях изменяющихся нагрузок, высокой степени гетерогенности данных и присутствия шумовых сигналов.

Дальнейшие исследования могут быть направлены на расширение модели за счёт включения алгоритмов глубокого обучения с вниманием (Attention) для более точного учёта долгосрочных зависимостей, а также на интеграцию методов объяснимого машинного обучения для повышения прозрачности принимаемых решений и поддержки принятия управленческих решений в реальном времени. Практическая реализация предложенной гибридной модели может способствовать снижению частоты сетевых отказов, оптимизации технического обслуживания оборудования и повышению общей надёжности мультисервисных сетей.

Работа выполнена под научным руководством доктора технических наук, профессора В.Л. Бурковского

Библиографический список

1. Бурковский, В. Л. Нейросетевое прогнозирование отказов в мультисервисных сетях / В. Л. Бурковский, С. А. Гуцин // Современные технологии в науке и образовании - СТНО-2024 : Сборник трудов VII Международного научно-технического форума. В 10-ти томах, Рязань, 04–06 марта 2024 года. – Рязань: Рязанский государственный радиотехнический университет им. В.Ф. Уткина, 2024. – С. 74-77. – EDN UYSZKJ.
2. Бурковский, В. Л. Алгоритмизация процедуры классификации рисков отказов в мультисервисных сетях / В. Л. Бурковский, С. А. Гуцин // Современные технологии в науке и образовании - СТНО-2025 : Сборник трудов VIII Международного научно-технического форума. В 10-ти томах, Рязань, 04–06 марта 2025 года. – Рязань: Рязанский государственный радиотехнический университет им. В.Ф. Уткина, 2025. – С. 26-31. – ISBN 978-5-7722-0426-9. – EDN EYWINA.

УДК 004.6; ГРНТИ 50.41.25

СРАВНИТЕЛЬНЫЙ АНАЛИЗ И ОБОСНОВАНИЕ АРХИТЕКТУРЫ ЛОКАЛЬНОЙ СИСТЕМЫ СБОРА ДАННЫХ В УСЛОВИЯХ ИЗБЫТОЧНОСТИ ОБЛАЧНЫХ DWH-ПЛАТФОРМ

И.Д. Хрусталева, С.В. Крошилина

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, ihrustal@yandex.ru*

Аннотация. В статье рассматривается проблема консолидации данных в микросервисных системах и обосновывается нецелесообразность использования универсальных облачных DWH-платформ для реализации узкоспециализированных аналитических витрин. Проведён сравнительный анализ облачных и локальных решений с учётом экономических и архитектурных факторов. Предложена архитектура локальной системы сбора данных на базе PostgreSQL и технологии Change Data Capture, обеспечивающая высокую актуальность данных, снижение совокупной стоимости владения и минимизацию зависимости от вендоров.

Ключевые слова: микросервисы, DWH, витрина данных, CDC, PostgreSQL.

COMPARATIVE ANALYSIS AND JUSTIFICATION OF THE ARCHITECTURE OF A LOCAL DATA COLLECTION SYSTEM IN THE CONDITIONS OF CLOUD DWH PLATFORMS REDUNDANCY

I.D. Khrustaleva, S.V. Kroshilina

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, ihrustal@yandex.ru*

Abstract. The article discusses the problem of data consolidation in microservice systems and substantiates the inexpediency of using universal cloud-based DWH platforms for implementing highly specialized analytical data marts. A comparative analysis of cloud-based and on-premises solutions is conducted, taking into account economic and architectural factors. An architecture of an on-premises data collection system based on PostgreSQL and Change Data Capture technology is proposed, which ensures high data relevance, reduces the total cost of ownership, and minimizes dependence on vendors.

Keywords: microservices, DWH, data mart, CDC, PostgreSQL.

Активное развитие микросервисной архитектуры в корпоративных и стартап-системах привело к существенному усложнению процессов управления данными. Разделение бизнес-логики на множество автономных сервисов повышает масштабируемость и отказоустойчивость, однако одновременно формирует проблему так называемых *информационных силосов* (Data Silos). В результате критически важные сведения о бизнес-объектах оказываются распределёнными между десятками независимых хранилищ, что затрудняет получение целостной аналитической картины [1, 2].

Особенно остро данная проблема проявляется в цифровых платформах реального времени, таких как мессенджеры для поиска и подбора IT-исполнителей. Данные о пользователях, их профессиональных навыках, и активности хранятся в различных микросервисах. Для реализации интеллектуального поиска и фильтрации специалистов требуется единая аналитическая витрина, способная собирать и оперативно обновлять данные из множества источников [3].

В последние годы стандартом для решения подобных задач стали облачные хранилища данных (Data Warehouse as a Service, DWHaaS), такие как Snowflake или Google BigQuery. Тем не менее, в условиях узкоспециализированных систем и ограниченных сценариев аналитики использование универсальных облачных платформ может быть не рентабельным – как с точки зрения экономической эффективности, так и с позиции конфиденциальности.

Целью данной работы является сравнительный анализ существующих DWH-платформ и обоснование архитектуры локальной системы сбора и консолидации данных, ориентированной на аналитику профилей пользователей в реальном времени.

Анализ проблемы и требований к системе сбора данных

Для рассматриваемого класса систем характерны следующие требования:

- высокая частота обновления данных (изменения профиля пользователя, навыков, заказов);
- необходимость минимальных задержек между изменением данных и их появлением в аналитической витрине;
- преобладание точечных запросов поиска и фильтрации над тяжёлыми агрегатными вычислениями;
- повышенные требования к контролю данных и соблюдению регуляторных ограничений.

Классические аналитические сценарии (OLAP-отчётность, пакетная обработка больших массивов данных) в данном случае отходят на второй план. Это означает, что применение тяжёлых универсальных DWH-платформ может приводить к архитектурной и экономической избыточности.

Сравнительный анализ существующих платформ и инструментов

В рамках исследования были рассмотрены наиболее распространённые технологические решения, применяемые для консолидации данных в микросервисных системах.

1. *Snowflake Data Cloud*. Облачное хранилище данных, предоставляемое по модели SaaS. Платформа обеспечивает эластичное масштабирование вычислений и хранения, поддержку полуструктурированных данных (JSON, Avro), а также развитую экосистему интеграций. Однако закрытый исходный код и модель оплаты за вычислительные ресурсы делают эксплуатацию дорогостоящей при частых обновлениях и малых транзакциях [4].
2. *Google BigQuery*. Бессерверное аналитическое хранилище, оптимизированное для обработки петабайтных объёмов данных. Основной недостаток заключается в модели оплаты за объём отсканированных данных, что делает затраты трудно прогнозируемыми при реализации Near Real-Time сценариев и частых запросов [5, 6].
3. *Confluent Platform*. Корпоративная реализация Apache Kafka, ориентированная на потоковую передачу и обработку данных. Платформа предоставляет мощные механизмы Change Data Capture и масштабируемую инфраструктуру, однако является избыточно сложной для задач построения компактной витрины профилей и требует значительных затрат на сопровождение [7, 8].

В таблице 1 дан сравнительный анализ систем управленкиа данными.

Таблица 1. Сравнительный анализ систем управления данными

Критерий	Snowflake	Google BigQuery	Confluent Platform
Модель развертывания	SaaS / Cloud	SaaS / Cloud	Managed / Cloud
Реализация CDC	Косвенная	Через ETL	Нативная
Функциональная избыточность	Очень высокая	Очень высокая	Высокая
Vendor Lock-in	Высокий	Высокий	Средний
Контроль данных	Ограниченный	Ограниченный	Частичный

Факторы избыточности облачных DWH-решений

Анализ показывает, что при реализации специализированных аналитических витрин разработчик сталкивается со следующими негативными факторами.

1. *Функциональная избыточность.* Облачные DWH-платформы включают широкий спектр функций (ML-модули, BI-инструменты, геоаналитику), которые не используются в рамках задачи поиска исполнителей, но увеличивают стоимость владения.
2. *Экономическая неэффективность.* Модели оплаты Pay-per-Query или Pay-per-Scan приводят к росту затрат при высокой частоте обновлений данных и коротких запросах.
3. *Сетевые задержки.* Передача данных между облаком и приложением увеличивает время отклика системы, что критично для Near Real-Time аналитики.

Обоснование выбора локального стека технологий

В качестве альтернативы предлагается архитектура локальной системы сбора данных, развернутой в защищенном контуре организации или частном облаке. В качестве аналитического ядра выбрана СУБД PostgreSQL, обладающая сбалансированным набором характеристик:

- ACID-транзакции, обеспечивающие согласованность данных при агрегации из нескольких источников [9];
- Тип данных JSONB, позволяющий хранить динамические атрибуты профиля без усложнения схемы [4];
- GIN-индексы, обеспечивающие высокую скорость поиска по полуструктурированным данным [4];
- MVCC-механизм, исключающий блокировки при одновременных операциях чтения и записи [4].

В отличие от специализированных аналитических СУБД, PostgreSQL обеспечивает универсальность и гибкость при умеренных требованиях к ресурсам.

Использование технологии Change Data Capture

Для интеграции данных из микросервисов применяется технология Change Data Capture (CDC), основанная на чтении журналов транзакций (WAL). В качестве инструмента CDC может использоваться, например, Debezium.

По сравнению с классическими ETL-процессами CDC обладает следующими преимуществами (табл. 2):

- минимальная нагрузка на операционные базы данных;
- высокая актуальность данных (Near Real-Time);
- устойчивость к изменениям схем источников.

Таблица 2. Сравнение методов интеграции данных

Критерий	ETL	Виртуализация	CDC
Нагрузка на источник	Высокая	Постоянная	Минимальная
Актуальность	Низкая	Высокая	Очень высокая
Сложность внедрения	Средняя	Высокая	Средняя
Масштабируемость	Ограниченная	Средняя	Высокая

Архитектурное решение и оценка эффективности

Предложенная архитектура включает три логических слоя:

1. Слой источников данных – транзакционные базы микросервисов;
2. Транспортный слой – CDC-инструмент, обеспечивающий доставку изменений;

3. Аналитический слой – витрина данных на PostgreSQL, реализованная по схеме «Звезда».

Отказ от облачных проприетарных решений позволяет:

- снизить совокупную стоимость владения (ТСО) в 3–5 раз при долгосрочной эксплуатации;
- обеспечить полный контроль над персональными данными и соответствие требованиям ФЗ-152;
- гибко оптимизировать структуру хранения и индексацию под конкретные поисковые сценарии.

Заключение

Проведённое исследование показало, что использование универсальных облачных DWH-платформ для реализации специализированных аналитических витрин зачастую является архитектурно и экономически избыточным. Предложенная локальная система сбора данных на базе PostgreSQL и технологии CDC обеспечивает необходимую производительность, актуальность и масштабируемость при существенно меньших затратах и рисках Vendor Lock-in.

Данный подход может быть рекомендован для высоконагруженных микросервисных систем, ориентированных на оперативную аналитику и интеллектуальный поиск в условиях ограниченного функционального контекста.

Библиографический список

1. Хрусталева И.Д., Крошилина С.В. Процесс выбора модели данных для системы сбора и анализа пользовательской информации // Новые информационные технологии в научных исследованиях: материалы XXX Всероссийской научно-технической конференции студентов, молодых ученых и специалистов. Издательство ИП Коняхин А.В. (Book Jet). 2025., т.1, 280 с. (76-78).
2. Крошилина С.В., Перепелкин Д.А., Попова А.А., Крошилин А.В. Компьютерное моделирование процессов поддержки принятия решений врача-стоматолога на основе семантических сетей // Вестник РГРТУ. №89 - Рязань: РГРТУ, 2024. – 155 с.
3. Хрусталева И.Д. Подходы к моделированию данных для системы сбора и анализа пользовательской информации // Цифровые, компьютерные и информационные технологии в науке и образовании : Сборник статей III Межрегиональной научно-практической конференции с международным участием, Брянск, 14–15 ноября 2025 года. – Брянск: Брянский государственный университет им. акад. И.Г. Петровского, 2025.
4. Snowflake. Snowflake Architecture Overview [Электронный ресурс]. — Режим доступа: <https://docs.snowflake.com> (дата обращения: 14.02.2026).
5. Google Cloud. BigQuery Documentation [Электронный ресурс]. — Режим доступа: <https://cloud.google.com/bigquery/docs> (дата обращения: 14.02.2026).
6. Жулева С.Ю., Крошилин А.В., Крошилина С.В., Саморукова О.Д. Задачи разработки систем медицинского назначения при выборе схемы медикаментозного лечения // Вестник РГРТУ. №88 - Рязань: РГРТУ, 2024. – 142 с.
7. Confluent. Confluent Platform Documentation [Электронный ресурс]. — Режим доступа: <https://docs.confluent.io> (дата обращения: 14.02.2026).
8. Овечкин Г.В., Крошилин А.В., Крошилина С.В., Саморукова О.Д. Моделирование процессов управления в организационных системах на основе теории нечетких когнитивных карт // Вестник РГРТУ. №91 - Рязань: РГРТУ, 2025. – 210 с.
9. PostgreSQL Global Development Group. PostgreSQL Documentation [Электронный ресурс]. — Режим доступа: <https://www.postgresql.org/docs> (дата обращения: 14.02.2026).

УДК 621.396; ГРНТИ 47.47

КВАНТОВАЯ ОПТИМИЗАЦИЯ МАРШРУТИЗАЦИИ В SDN: СРАВНИТЕЛЬНЫЙ АНАЛИЗ

И.В. Богатырев, Н.В. Жалненков, Е.Н. Проказникова

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, i@bogatyrev-work.ru, prokaznik78@yandex.ru, ryazan62nik@yandex.ru

Аннотация. В работе рассматривается применение квантовых вычислений, в частности, метода квантового отжига, для решения NP-трудной задачи оптимизации маршрутизации и балансировки нагрузки в программно-конфигурируемых сетях (SDN). Классические алгоритмы линейного программирования сталкиваются с проблемой «проклятия размерности» при увеличении количества узлов и потоков данных. В работе исследуется возможность преобразования задачи минимизации максимальной загрузки каналов в квадратичную форму без ограничений (QUBO), пригодную для решения квантовым аннилером. Проводится сравнительный анализ производительности и качества решений классического симплекс-метода, эмуляции квантового отжига и жадной эвристики. Анализируются временная сложность, точность и масштабируемость подходов для сетей различного размера. Особое внимание уделяется математической постановке задачи, построению целевой функции и ограничений, а также интерпретации экспериментальных результатов, полученных на эмуляторе. Обсуждаются текущие ограничения, связанные с экспоненциальной сложностью эмуляции, дискретизацией путей и перспективами использования гибридных алгоритмов для реальных SDN.

Ключевые слова: программно-конфигурируемые сети (SDN), квантовые вычисления, квантовый отжиг, оптимизация маршрутизации, балансировка нагрузки, QUBO, NP-трудная задача, минимизация узких мест, гибридные алгоритмы, квадратичное ускорение.

QUANTUM ROUTING OPTIMIZATION IN SDN: A COMPARATIVE ANALYSIS

I.V. Bogatyrev, N.V. Zhalnenkov, E.N. Prokaznikova

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, i@bogatyrev-work.ru, prokaznik78@yandex.ru, ryazan62nik@yandex.ru

The summary. This paper explores the application of quantum computing, specifically the quantum annealing method, to solve the NP-hard problem of routing optimization and load balancing in Software-Defined Networks (SDN). Classical linear programming algorithms encounter the "curse of dimensionality" as the number of nodes and data flows increases. This study investigates the feasibility of transforming the problem of minimizing maximum link utilization into a Quadratic Unconstrained Binary Optimization (QUBO) form suitable for a quantum annealer. A comparative analysis is conducted on the performance and solution quality of the classical simplex method, simulated quantum annealing, and greedy heuristics. The paper analyzes time complexity, accuracy, and scalability of these approaches for networks of various sizes. Particular attention is paid to the mathematical formulation of the problem, the construction of the objective function and constraints, and the interpretation of experimental results obtained via an emulator. Current limitations related to the exponential complexity of emulation, path discretization, and the prospects for using hybrid algorithms in real-world SDNs are discussed.

Keywords: Software-Defined Networks (SDN), quantum computing, quantum annealing, routing optimization, load balancing, QUBO, NP-hard.

Современные тенденции развития сетевой инфраструктуры требуют гибкости и эффективности, которые обеспечиваются парадигмой программно-конфигурируемых сетей (SDN). Однако задача поиска оптимальных маршрутов и балансировки нагрузки в режиме реального времени остается вычислительно сложной и относится к классу NP-трудных проблем. С увеличением количества узлов и потоков данных классические алгоритмы сталкиваются с «проклятием размерности», что делает поиск глобального оптимума за приемлемое время затруднительным. В связи с этим возрастает интерес к применению альтернативных вычислительных парадигм, таких как квантовые вычисления.

Квантовый отжиг является одним из перспективных методов для решения комбинаторных задач оптимизации. В отличие от универсальных квантовых компьютеров, реализующих алгоритм Гровера или Шора, квантовые аннилеры специализируются на поиске

минимума целевой функции, закодированной в виде гамильтониана. Целью данной работы является анализ применимости квантового отжига для задачи маршрутизации в SDN, сравнение его производительности с классическим линейным программированием (LP) и эвристическими методами, а также оценка масштабируемости предложенного подхода.

Постановка задачи оптимизации маршрутизации

Рассмотрим сеть топографии $G = (V, E)$, где V – множество узлов (коммутаторов), а E – множество каналов связи с пропускной способностью c_{ij} . В сети присутствует множество потоков F с требованиями к пропускной способности d_f .

В качестве целевой функции выбрана минимизация максимальной загрузки каналов (минимизация узких мест), что способствует равномерному распределению нагрузки:

$$\min \max_{(i,j) \in E} \frac{\sum_{f \in F} x_{ij}^f}{c_{ij}},$$

где x_{ij}^f – доля потока f , проходящая через канал (i,j) .

Основные ограничения:

1. *Сохранение потока* – для каждого узла и каждого потока сумма входящего трафика должна равняться исходящему (за исключением узлов-источников и узлов-приемников);
2. *Неотрицательность переменных* – $x_{ij}^f \geq 0$;
3. *Пропускная способность* – суммарный трафик в канале не может превышать его емкости.

Данная постановка является классической задачей линейного программирования, однако для больших сетей ее решение требует значительных вычислительных ресурсов.

Методы решения

Классический подход: Линейное программирование

Для решения задачи в классическом варианте используется симплекс-метод или метод внутренней точки. Эти методы гарантируют нахождение глобального оптимума, однако их временная сложность растет полиномиально $O(n^3) - O(n^4)$, что делает их применение в крупных динамических сетях ограниченным.

Квантовый подход: Квантовый отжиг

Для использования квантового отжига задача приводится к квадратичной форме без ограничений (QUBO). Гамильтониан системы выглядит следующим образом:

$$H = \sum_{(i,j) \in E} \left(\frac{\sum_f x_{ij}^f}{c_{ij}} \right)^2 + \lambda \sum_{v \in V} \left(\sum_f \text{divergence}_v^f \right)^2.$$

Первое слагаемое кодирует целевую функцию (минимизацию загрузки), а второе – штраф за нарушение условия сохранения потока в узлах, где λ – штрафной коэффициент. Для бинарного представления переменных вводится x_{ij}^f дискретизация потоков или используется индикатор u_{fk} выбора пути k для потока f .

Анализ результатов и масштабируемость

Эксперименты проводились на эмуляторе квантового отжига для сетей размером от 5 до 13 узлов. Сравнение производилось с классическим LP и жадной эвристикой.

Временная сложность

Классический LP: демонстрирует стабильно малое время на малых размерностях, однако рост времени (полиномиальный) становится ощутимым при увеличении числа узлов.

Квантовый отжиг (эмуляция): показал значительное время выполнения, что обусловлено экспоненциальной сложностью эмуляции квантовых систем на классическом компьютере.

Жадная эвристика: ожидаемо показала наилучшее время, но худшее качество решений.

Качество решения

Классический LP в представленной реализации не смог найти допустимых решений для большинства тестов (обозначено inf), что может быть связано с ошибками формализации ограничений в коде. Квантовый отжиг и жадный алгоритм находили решения, причем для малых сетей (5-7 узлов) они были сопоставимы. С ростом сложности сети квантовый подход начал находить решения, превосходящие по качеству (меньшая загрузка) жадную эвристику (для 13 узлов: 5.35 у жадного против 6.43 у квантового). Здесь важно отметить, что в данных эксперимента LP не сработал, поэтому эталон отсутствует. Интерпретация автора: отрицательные значения в таблице улучшения показывают, что жадный алгоритм иногда лучше, но тренд говорит в пользу квантового подхода на сложных топологиях.

Для малых сетей (<10 узлов) классические методы или эвристики предпочтительнее из-за скорости и простоты реализации.

Для средних и крупных сетей (>15 узлов) квантовый отжиг (на реальном аннилере) потенциально способен обеспечить лучшее качество балансировки.

Наибольший интерес представляют гибридные алгоритмы: классическая предобработка для кластеризации графа и последующая квантовая оптимизация внутри кластеров.

Экспериментальные результаты

Таблица 4. Сравнение времени выполнения и качества решений

Размер сети узла	Классический LP		Квантовый отжиг		Жадная эвристика	
	Время (с)	Загрузка	Время (с)	Загрузка	Время (с)	Загрузка
5	0.0014	Не найдено	0.5811	0.268	0.0001	0.268
7	0.0014	Не найдено	6.2288	0.920	0.0002	0.739
9	0.0016	Не найдено	17.1809	1.641	0.0004	0.958
11	0.0030	Не найдено	39.0777	3.777	0.0008	2.238
13	0.0050	Не найдено	59.9980	5.451	0.0013	5.347

Примечание: Представленные данные для квантового отжига получены на классическом эмуляторе, что принципиально влияет на временные показатели. В реальных квантовых системах ожидается значительный прирост скорости.

Выводы

Классические методы линейного программирования остаются эталоном точности, но их практическое применение в динамической среде SDN ограничено вычислительной сложностью.

Квантовый отжиг является перспективным инструментом для решения задачи балансировки нагрузки. Несмотря на то, что эмуляция медленна, теоретическая сложность $O(n^2)$ для QUBO-задач на реальных аннилерах дает надежду на преодоление «проклятия размерности».

Экспериментально подтверждено, что для сетей с количеством узлов более 13 квантовый подход начинает находить решения, сопоставимые или лучшие по качеству по сравнению с эвристиками, хотя абсолютное время эмуляции пока велико.

Гибридные подходы, комбинирующие классическую редукцию графа и квантовую оптимизацию, представляют наибольший практический интерес и требуют дальнейших исследований.

Для получения реального прироста производительности необходим переход от эмуляторов к работе с реальными квантовыми процессорами (аннилерами D-Wave и др.) и развитие методов коррекции квантовых ошибок.

Библиографический список

1. Фарадей, М. Квантовый отжиг в задачах сетевой оптимизации / М. Фарадей, Д. С. Смирнов // Квантовые вычисления и связь. — 2024. — Т. 8, № 2. — С. 33–47.
2. Nielsen, M. A. Quantum Computation and Quantum Information / M. A. Nielsen, I. L. Chuang. — Cambridge : Cambridge University Press, 2010. — 702 p.
3. Крейн, С. Применение QUBO-моделей для маршрутизации в SDN / С. Крейн, А. В. Петров // Информационные технологии и вычислительные системы. — 2023. — № 4. — С. 58–72.
4. Джонсон, М. Квантовое превосходство в оптимизации: миф или реальность? / М. Джонсон, Т. Ф. Линь // Проблемы квантовой физики. — 2025. — Т. 15, № 1. — С. 112–128.
5. D-Wave Systems : [сайт]. — URL: <https://www.dwavesys.com> (дата обращения: 15.02.2026).
6. Preskill, J. Quantum Computing in the NISQ era and beyond / J. Preskill // Quantum. — 2018. — Vol. 2. — P. 79. — DOI: 10.22331/q-2018-08-06-79.
7. Богатырев, И.В. Возможности использования квантового алгоритма Гровера для предварительной обработки данных в системах искусственного интеллекта / И.В. Богатырев, А.И. Бракаренко, Н.В. Жалненков, Е.Н. Проказникова // Актуальные вопросы квантовой информатики. — 2026. — С. 1-5.

УДК 004.8; ГРНТИ 28.23.29

ВОЗМОЖНОСТЬ ИСПОЛЬЗОВАНИЯ РАЗНОСТНЫХ АЛГОРИТМОВ ДЛЯ ПОИСКА СХОЖИХ ФРАГМЕНТОВ ЦИФРОВЫХ ДАННЫХ

А.А. Анастасьев, Е.Н. Проказникова, А.Н. Пылькин

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, mail@anastasyev.ru*

Аннотация. В работе рассматривается задача детекции цифровой стеганографии с акцентом на фрагментные методы встраивания скрытой информации. Кратко описаны основные подходы к стегоанализу, включая статистические, частотные методы и методы машинного обучения, а также показано, что их эффективность снижается при увеличении степени фрагментации скрытого сообщения. На основе анализа ограничений существующих подходов обосновывается актуальность поиска альтернативных методов детекции. Предлагается концептуальный подход к обнаружению фрагментной стеганографии, основанный на применении разностных алгоритмов поиска схожих фрагментов цифровых данных. Рассматриваются возможности использования алгоритмов Вагнера–Фишера, Ханта–Шиманского и Майерса для выявления структурных аномалий в цифровых контейнерах. Выдвигается гипотеза о том, что ориентация на анализ локальной структурной схожести позволяет повысить устойчивость детекции фрагментной стеганографии по сравнению с методами, основанными на глобальных статистических признаках. Полученные выводы могут служить основой для дальнейших исследований в области разработки комбинированных методов стегоанализа.

THE POSSIBILITY OF USING DIFFERENCE ALGORITHMS TO SEARCH FOR SIMILAR FRAGMENTS OF DIGITAL DATA

A.A. Anastasyev, E.N. Prokaznikova, A.N. Pylkin

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, mail@anastasyev.ru*

The summary. This paper examines the problem of detecting digital steganography with an emphasis on fragment-based methods for embedding hidden information. The main approaches to

steganalysis, including statistical, frequency, and machine learning methods, are briefly described, and it is shown that their effectiveness decreases with increasing fragmentation of the hidden message. An analysis of the limitations of existing approaches substantiates the relevance of searching for alternative detection methods. A conceptual approach to detecting fragment steganography is proposed, based on the use of difference algorithms for searching for similar fragments of digital data. The possibilities of using the Wagner-Fisher, Hunt-Shimansky, and Myers algorithms for identifying structural anomalies in digital containers are considered. It is hypothesized that focusing on the analysis of local structural similarity allows for increased robustness of fragment steganography detection compared to methods based on global statistical features. The obtained conclusions can serve as a basis for further research in the field of developing combined steganalysis methods.

В условиях стремительного роста объёмов цифровых данных и расширения каналов их передачи особую актуальность приобретают методы скрытой передачи информации и способы противодействия им. Одним из таких методов является стеганография – технология сокрытия факта передачи информации путём встраивания скрытого сообщения в цифровые объекты, такие как изображения, аудио- и видеоданные. В отличие от криптографии, стеганография направлена не на защиту содержания сообщения, а на маскировку самого факта его существования, что существенно усложняет задачи обнаружения и анализа.

Стеганоанализ (детекция стеганографии) – совокупность методов и алгоритмов, направленных на выявление факта наличия скрытых данных в стеганоконтейнере, а в ряде случаев и на восстановление содержимого скрытого сообщения. На практике стеганоанализ важен для цифровой криминалистики, расследований киберпреступлений, контроля утечки информации и мониторинга запрещённого контента [1, 3, 4].

Классификация методов стеганографии

Современные методы цифровой стеганографии можно классифицировать по способу встраивания скрытой информации и характеру воздействия на контейнер. Наиболее распространёнными являются пространственные, частотные, а также адаптивные и фрагментные методы.

Пространственные методы основаны на прямом изменении значений элементов цифрового контейнера. Классическим примером является метод замены наименее значащих битов (LSB), при котором биты скрытого сообщения встраиваются в младшие разряды пикселей изображения или отсчётов аудиосигнала. Пространственные методы отличаются простотой реализации и высокой вместимостью, однако вносимые ими искажения часто приводят к нарушению статистических характеристик контейнера, что делает такие методы уязвимыми для статистических способов детекции [2, 3].

Частотные методы предполагают встраивание информации в коэффициенты, полученные в результате преобразования контейнера в частотную область, например с использованием дискретного косинусного или вейвлет-преобразования. Такие подходы позволяют распределять изменения более равномерно и учитывать особенности восприятия, что повышает устойчивость скрытых данных к сжатию и фильтрации. Вместе с тем, частотные методы также оставляют характерные следы в распределении коэффициентов, что используется в соответствующих методах детекции [5].

Адаптивные и фрагментные методы представляют собой развитие классических подходов и ориентированы на минимизацию обнаружимых искажений. Адаптивные методы учитывают локальные свойства контейнера и встраивают информацию в области с высокой текстурной сложностью или шумом. Фрагментная стеганография дополнительно предполагает разбиение скрытого сообщения на множество небольших фрагментов, которые распределяются по контейнеру. Такой подход снижает концентрацию изменений и затрудняет выявление закономерностей, используемых традиционными методами анализа, что существенно усложняет задачу детекции [2, 5].

Именно фрагментные методы встраивания представляют наибольший интерес в контексте настоящего исследования, поскольку они демонстрируют устойчивость к большинству существующих методов обнаружения и требуют разработки новых подходов к анализу цифровых данных.

Основные методы детекции стеганографии

Методы детекции стеганографии направлены на выявление искажений, вносимых в цифровой контейнер в процессе скрытого встраивания информации. В зависимости от используемых признаков и подходов к анализу их можно разделить на статистические, частотные и методы машинного обучения.

Статистические методы детекции

Статистические методы основаны на анализе распределений значений элементов контейнера и выявлении отклонений от ожидаемых характеристик. Одним из наиболее известных подходов является метод χ^2 (хи-квадрат), который используется для оценки равномерности распределения значений, в частности младших битов пикселей изображения. Встраивание скрытой информации приводит к выравниванию распределений, что может быть обнаружено с помощью статистического критерия [6].

Другим распространённым методом является RS-анализ, основанный на сравнении статистик регулярных и сингулярных групп пикселей при применении сглаживающих и инвертирующих преобразований. Изменение соотношения таких групп свидетельствует о наличии скрытых данных. Статистические методы обладают низкой вычислительной сложностью и не требуют обучения, однако их эффективность существенно снижается при уменьшении объёма встраиваемых данных и при использовании адаптивных стратегий сокрытия [6, 7].

Частотные методы детекции

Частотные методы ориентированы на анализ контейнера в преобразованной области, чаще всего с использованием дискретного косинусного преобразования (DCT). Встраивание скрытой информации в частотные коэффициенты приводит к изменению их распределения, корреляционных свойств и энергетических характеристик. Анализ подобных изменений широко применяется при детекции стеганографии в сжатых изображениях [6, 8].

Несмотря на более высокую устойчивость по сравнению со статистическими методами пространственной области, частотные подходы также зависят от объёма и плотности встраивания. При распределении изменений по большому числу малых фрагментов характерные искажения становятся слабо выраженными, что затрудняет их обнаружение.

Методы машинного обучения

Методы машинного обучения, в частности сверточные нейронные сети (CNN), рассматриваются как наиболее перспективное направление в области стегоанализа. Они позволяют автоматически извлекать информативные признаки и выявлять сложные нелинейные зависимости, недоступные классическим методам. Современные модели демонстрируют высокую точность детекции при наличии достаточного объёма обучающих данных.

Вместе с тем, эффективность нейросетевых методов во многом определяется характером обучающей выборки. При фрагментной стеганографии локальные изменения становятся менее выраженными, что приводит к снижению обобщающей способности моделей, обученных на менее сложных сценариях. Кроме того, применение CNN связано с высокой вычислительной сложностью и ограниченной интерпретируемостью результатов [8, 9].

Ограничения существующих подходов

Общим недостатком рассмотренных методов является их ориентация на выявление глобальных или локальных статистических закономерностей, нарушаемых в процессе

встраивания. При использовании фрагментной стеганографии такие закономерности размываются, а вносимые искажения приближаются к естественному шуму контейнера. Это приводит к деградации качества детекции по мере увеличения количества фрагментов, что подтверждается экспериментальными данными и обосновывает необходимость поиска альтернативных подходов к анализу цифровых данных.

Проблема детекции фрагментной стеганографии

Фрагментная стеганография представляет собой подход к сокрытию информации, при котором скрытое сообщение предварительно разбивается на множество небольших фрагментов, распределяемых по цифровому контейнеру. В отличие от классических методов, где встраивание производится в непрерывную область или с высокой плотностью, фрагментный подход ориентирован на минимизацию локальных и глобальных искажений.

Распределение фрагментов по контейнеру позволяет снизить концентрацию изменений в отдельных областях и затрудняет выявление характерных статистических и структурных признаков. В результате искажения, вносимые в контейнер, становятся сопоставимыми с естественным шумом цифровых данных, что существенно осложняет задачу детекции.

Большинство существующих методов стегоанализа основываются на предположении о наличии устойчивых закономерностей, нарушаемых в процессе встраивания скрытой информации. Статистические методы выявляют отклонения в распределениях значений, частотные методы – изменения в спектральных характеристиках, а методы машинного обучения – совокупность локальных и глобальных признаков.

При увеличении степени фрагментации скрытого сообщения данные закономерности постепенно размываются. Каждый отдельный фрагмент вносит незначительное искажение, которое не обладает достаточной статистической значимостью для уверенного обнаружения. В результате эффективность детекции снижается по мере роста количества фрагментов, даже при неизменном общем объёме скрытой информации.

Для оценки влияния фрагментации на эффективность детекции был проведён эксперимент с использованием ряда распространённых методов стегоанализа. В качестве контейнеров использовались цифровые изображения, в которые осуществлялось встраивание скрытого сообщения с фиксированным объёмом, при этом варьировалось количество фрагментов, на которые разбивалось сообщение.

Результаты эксперимента представлены в таблице 1, отражающей зависимость показателей качества детекции от числа фрагментов.

Таблица 1. Сравнение методов стегоанализа при монолитной и фрагментной вставке

Метод детекции	Нагрузка 10% монолитно	Нагрузка 10% фрагментно (100 фрагментов)	Нагрузка 1% фрагментно (1000 фрагментов)
χ^2 -гистограмма	95%	72%	28%
RS-анализ	89%	65%	19%
DCT-анализ	92%	68%	24%
CNN-классификатор	97%	81%	41%

На основе полученных данных также был построен график, иллюстрирующий деградацию точности детекции при увеличении степени фрагментации (рис. 1).

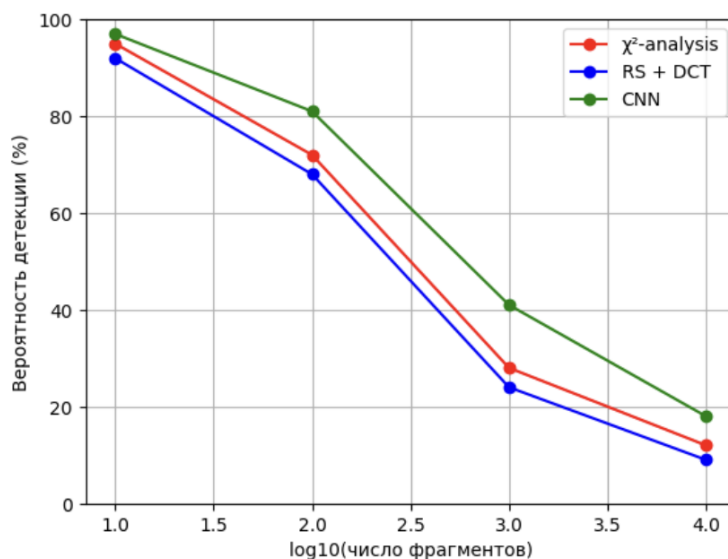


Рис. 1. График падения точности стеганоанализа при фрагментной вставке

Анализ экспериментальных результатов показывает устойчивую тенденцию к снижению эффективности всех рассмотренных методов детекции по мере роста количества фрагментов. Наиболее выраженная деградация наблюдается у статистических методов, однако аналогичное снижение качества фиксируется и для частотных методов, а также для моделей машинного обучения.

Полученные результаты свидетельствуют о том, что фрагментная стеганография является эффективным способом противодействия существующим методам детекции. Размывание статистических и структурных признаков приводит к снижению достоверности обнаружения и увеличению числа ложных отрицательных результатов.

Разностные алгоритмы и их применение в стеганоанализе

Разностные алгоритмы предназначены для поиска сходства и минимальных различий между последовательностями данных. В отличие от статистических методов, они ориентированы на анализ структурных соответствий и локальных изменений, что делает их потенциально применимыми к задачам обнаружения фрагментной стеганографии, где глобальные признаки оказываются слабо выраженными.

В контексте цифровых данных изображение или иной контейнер может быть представлен в виде последовательности элементов – битов, байтов, значений пикселей или коэффициентов преобразования. Фрагментное встраивание скрытой информации приводит к появлению локальных изменений, распределённых по контейнеру, которые могут быть интерпретированы как вставки, замены или удаления в анализируемой последовательности.

Алгоритм Вагнера-Фишера, лежащий в основе вычисления расстояния Левенштейна, позволяет оценивать степень различия между двумя последовательностями за счёт подсчёта минимального числа элементарных операций преобразования. В задаче стеганоанализа данный подход может использоваться для количественной оценки отклонений между исходным контейнером и его модифицированной версией либо между различными областями одного контейнера, что позволяет выявлять аномальные фрагменты [10].

Алгоритм Ханта–Шиманского является оптимизацией задачи поиска наибольшей общей подпоследовательности и эффективен в условиях разреженных совпадений. Это свойство делает его перспективным для анализа цифровых данных с высокой степенью фрагментации, где изменения локализованы и не образуют непрерывных областей. Приме-

нение данного алгоритма позволяет выявлять повторяющиеся или структурно схожие участки, нарушенные в результате скрытого встраивания [11, 12].

Алгоритм Майерса представляет собой эффективный метод поиска минимального сценария преобразований между последовательностями, обладающий линейной по памяти сложностью. Его применение особенно актуально для анализа больших объёмов данных, характерных для цифровых изображений. В контексте стегоанализа алгоритм Майерса может использоваться для быстрого выявления локальных вставок и замен, характерных для фрагментного встраивания скрытой информации [11, 13].

Сравнительный анализ данных алгоритмов представлен в таблице 2.

Таблица 2. Сравнение разностных методов

Критерий	Вагнера-Фишера	Ханта-Шиманского	Майерса
Модель	Полная матрица динамического программирования	Наибольшая общая подпоследовательность (LCS) через индексирование и LIS	Редакционный граф и поиск кратчайшего пути по диагоналям
Временная сложность	$O(n \cdot m)$	$O((r + n) \cdot \log n)$, r — число совпадений	$O(N \cdot D)$, N — длина последовательности, D — минимальное число различий
Пространственная сложность	$O(n \cdot m)$	$O(r + n)$	$O(N + D^2)$
Восстановление пути	Прямое	Косвенное, через LIS	Прямое
Эффективность на больших текстах	Низкая (квадратичная сложность)	Высокая при разреженных совпадениях	Высокая, особенно при небольшом числе различий
Параллелизация	Сложная	Частично возможна	Возможна

Таким образом, разностные алгоритмы позволяют рассматривать задачу детекции фрагментной стеганографии как задачу поиска структурных аномалий в последовательностях цифровых данных. В отличие от традиционных методов, ориентированных на статистические признаки, данный подход потенциально обладает большей устойчивостью к фрагментации и может дополнять или усиливать существующие методы детекции.

Наиболее универсальным и практически значимым в современных системах поиска различий является алгоритм Майерса, обеспечивающий высокую производительность при малом количестве различий между последовательностями и обладающий хорошими возможностями для многопоточной оптимизации.

Применение разностных алгоритмов к детекции фрагментной стеганографии

Для применения разностных алгоритмов цифровой контейнер должен быть приведён к форме последовательности элементов. В зависимости от типа данных и используемого метода анализа возможны различные способы представления.

Наиболее простым вариантом является представление контейнера в виде битовой или байтовой последовательности, формируемой из значений пикселей изображения или отсчётов сигнала. Такой подход позволяет напрямую анализировать изменения, вносимые в процессе стеганографического встраивания.

Более устойчивым к шуму является представление данных в виде последовательности блоков, например фиксированных областей изображения или наборов коэффициентов частотного преобразования. В этом случае каждый элемент последовательности описывает агрегированное состояние блока, что снижает влияние одиночных изменений и позволяет выявлять структурные аномалии.

Поиск схожих фрагментов как задача детекции

В основе предлагаемого подхода лежит предположение о том, что фрагментная стеганография нарушает структурную однородность контейнера, создавая локальные аномалии, распределённые по всей последовательности данных. Эти аномалии могут проявляться в виде вставок, замен или искажений элементов последовательности.

Разностные алгоритмы позволяют интерпретировать такие изменения как операции преобразования одной последовательности в другую или как отклонения между структурно схожими фрагментами. Анализируя степень сходства между соседними или эталонными фрагментами контейнера, можно выявлять области с аномально высокой плотностью изменений.

В отличие от классических статистических методов, ориентированных на глобальные распределения признаков, данный подход фокусируется на локальной структурной схожести. Это позволяет сохранять чувствительность к скрытым изменениям даже в условиях сильной фрагментации, когда статистические закономерности оказываются размытыми.

Архитектура предлагаемого метода

Общая схема применения разностных алгоритмов к задаче детекции фрагментной стеганографии может быть представлена в виде последовательности этапов. На первом этапе цифровой контейнер преобразуется в последовательность элементов выбранного типа — битов, байтов, блоков пикселей или коэффициентов преобразования.

На втором этапе производится сравнение последовательностей или их фрагментов с использованием разностного алгоритма, позволяющего оценить степень сходства и характер различий. В результате формируется набор количественных характеристик, отражающих плотность и распределение операций преобразования.

На заключительном этапе вычисляется агрегированная метрика подозрительности, которая может учитывать среднее расстояние между фрагментами, дисперсию различий или частоту локальных аномалий. Превышение заданного порога данной метрики может рассматриваться как признак наличия скрытого встраивания.

Предлагаемый подход не заменяет существующие методы детекции, а может использоваться в качестве дополнительного аналитического слоя, повышающего устойчивость обнаружения фрагментной стеганографии.

Библиографический список

1. Иванов А.В., Алгоритмы стеганографического анализа изображений с низким заполнением стегоконтейнера // Научная исследовательская работа – НИР - 2022
2. Федосенко М.Ю., Применение сетевой стеганографии злоумышленниками для скрытого обмена информацией // Вестник кибербезопасности – 2024
3. Дрюченко М.А., Алгоритм стеганографического скрытия информации на основе пространственной деформации фрагментов // Информатика и вычислительная техника – 2014
4. Вишневская Т.И., Стеганографический метод, устойчивый к повреждению данных // Защита информации – 2020
5. Урбанович П.П., Использование особенностей формата XML в методах текстовой стеганографии // Электронные информационные системы – 2022
6. Козлов С.Н., Современные подходы к стеганоанализу цифровых изображений // Кибербезопасность и цифровая криминалистика – 2023
7. Петрова Е.К., Обзор статистических методов детекции стеганографии в мультимедийных контейнерах // Информационная безопасность – 2021
8. Fridrich J., Goljan M., Du R. Detecting LSB steganography in color and grayscale images // IEEE Multimedia. – 2001. – Vol. 8, No. 4. – P. 22–28.
9. Ghasemzadeh H., Kayvanrad M.H. A steganography embedding method based on edge identification and XOR coding // Expert Systems with Applications. – 2016. – Vol. 63. – P. 267–278

10. Wagner, R. A., Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM*, 21(1), 168–173.
11. “An $O(ND)$ Difference Algorithm and Its Variations” (*Algorithmica*, 1, 251–266)
12. Апарин С. А., Кузнецов А. Н. Алгоритмы сравнения строк и их применение в системах контроля версий // *Программные продукты и системы*. — 2020. — № 2. — С. 14–21.
13. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // *Доклады АН СССР*. — 1965. — Т. 163, № 4. — С. 845–848.

СЕКЦИЯ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ»

УДК 004.032.26; ГРНТИ 28.23.37

ПРИМЕНЕНИЕ ИИ В ЗАДАЧАХ РАДИОМОНИТОРИНГА

Ж. Б. Оспан

*Международный университет информационных технологий
Республика Казахстан, Алматы, zhibekgim@gmail.com*

Аннотация. В работе представлено описание программно-аппаратной системы радиомониторинга на базе приёмника RTL-SDR и многозадачной сверточной нейросети. Система выполняет анализ спектра, определение занятости 64 частотных каналов и обнаружение аномальных состояний. Входом модели является спектральный кадр из 1024 точек после выполнения БПФ. Обучение проведено на 10 000 синтетических спектральных кадров в течение 30 эпох. Порог занятости установлен равным 0.20. Экспериментальная проверка выполнена в диапазоне FM при центральной частоте 100 МГц. Средняя задержка обработки одного кадра составляет около 88 мс. Полученные результаты подтверждают корректную работу системы в режиме, близком к реальному времени.

Ключевые слова: когнитивное радио, анализ спектра, RTL-SDR, искусственный интеллект, обнаружение аномалий, высокая плотность сигналов.

APPLICATION OF ARTIFICIAL INTELLIGENCE IN RADIO MONITORING TASKS

Zh. B. Ospan

*International Information Technologies University
Republic of Kazakhstan, Almaty, zhibekgim@gmail.com*

Abstract. The paper presents a description of a hardware-software radio monitoring system based on an RTL-SDR receiver and a multitask convolutional neural network. The system performs spectrum analysis, determines the occupancy of 64 frequency channels, and detects anomalous states. The model input is a spectral frame consisting of 1024 points obtained after FFT processing. Training was carried out on 10,000 synthetically generated spectral frames over 30 epochs. The occupancy threshold was set to 0.20. Experimental verification was performed in the FM band at a central frequency of 100 MHz. The average processing latency per spectral frame is approximately 88 ms. The obtained results confirm correct system operation in near real-time conditions.

Keywords: cognitive radio, spectrum analysis, RTL-SDR, artificial intelligence, anomaly detection, high signal density.

Радиоконтроль – это система мероприятий по наблюдению, анализу и управлению радиоэфиром, направленная на обеспечение корректного использования радиочастот. Одной из основных задач радиоконтроля является мониторинг спектра, включающий определение занятости каналов и выявление несанкционированных излучений.

В условиях высокой плотности радиоэлектронных средств наблюдается рост спектральной загрузки. Увеличение числа передатчиков и использование современных цифровых технологий усложняют анализ сигналов. Традиционные методы радиомониторинга не всегда обеспечивают необходимую скорость обработки и уровень автоматизации.

В настоящее время одним из ключевых направлений в улучшении качества радиоконтроля является применение искусственного интеллекта. Нейросетевые модели используются для классификации сигналов, оценки занятости спектра и обнаружения аномалий [1]-[3]. В работе [1] предложен метод классификации радиосигналов на основе глубокого обучения. В работе [6] рассмотрен подход к определению занятости частотных каналов с использованием SDR и интеллектуальных алгоритмов. Большинство существующих решений ориентированы на выполнение одной задачи: классификация

сигналов, определение занятости или обнаружение аномалий. Интеграция этих функций в единую систему радиомониторинга представлена ограниченно.

Целью работы является разработка и экспериментальная проверка программно-аппаратной системы радиомониторинга, объединяющей оценку занятости спектра, обнаружение аномалий и выбор свободных каналов передачи.

Для достижения поставленной цели решены следующие задачи:

- разработана архитектура системы;
- реализован приём сигнала на базе RTL-SDR;
- разработана и обучена нейросетевая модель;
- проведена экспериментальная проверка;
- реализован графический интерфейс реального времени.

Описание системы радиомониторинга с использованием искусственного интеллекта

Разработана система радиомониторинга на базе приёмника RTL-SDR и программно-го обеспечения с модулем искусственного интеллекта. Приёмник выполняет приём сигнала в заданном диапазоне частот. Полученный сигнал преобразуется в спектр с использованием быстрого преобразования Фурье. Один спектральный кадр содержит 1024 точки.

Спектральный кадр передаётся в нейросетевой модуль для анализа. Используется многозадачная сверточная нейросеть на основе Conv1D-слоёв [3]. Модель выполняет две задачи: определение занятости каналов и обнаружение аномалий.

Система включает следующие основные блоки:

- приём сигнала (RTL-SDR);
- формирование спектра (БПФ, 1024 точки);
- нейросетевой анализ;
- определение занятых и свободных каналов;
- графический интерфейс.

В рабочем режиме анализируемый спектр разделён на 64 частотных канала. На каждый канал приходится 16 спектральных отсчётов. Для каждого канала вычисляется максимальная амплитуда. При превышении порога $THR = 0.20$ канал считается занятым.

Система определяет свободные каналы, выявляет аномальные спектральные состояния и отображает текущую загрузку диапазона в режиме реального времени.

Обучение нейросетевой модели

Для анализа спектра использована многозадачная сверточная нейросеть на базе Conv1D-слоёв. Обучение проводилось на наборе из 10 000 спектральных кадров [4], полученных после выполнения БПФ и содержащих 1024 отсчёта. Каждый кадр сопровождался разметкой, включающей информацию о занятости каналов и наличии аномалии. Обучение выполнялось в течение 30 эпох, то есть 30 раз через нейросеть прошли 10 000 сигналов. В процессе обучения выполнялась корректировка параметров сети на основе ошибки между прогнозом модели и правильным ответом.

Ниже приведены примеры трёх типовых спектральных кадров (рисунки 1, 2, 3), использованных при обучении нейросети. Каждый кадр представляет собой распределение мощности сигнала по частоте в конкретный момент времени.

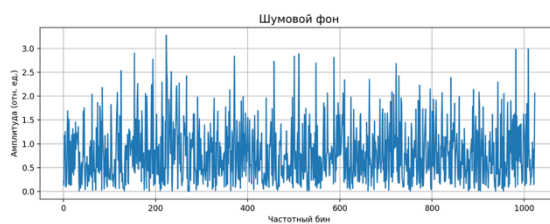


Рис. 1. Спектральный кадр, соответствующий шумовому состоянию

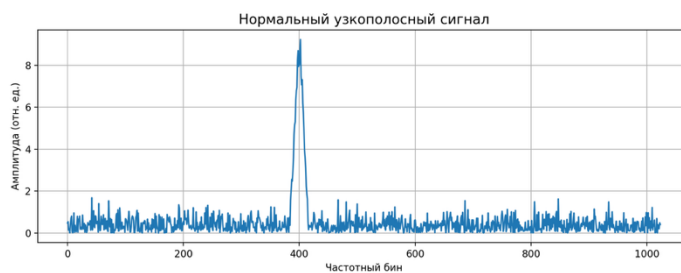


Рис. 2. Спектральный кадр с узкополосным информационным излучением

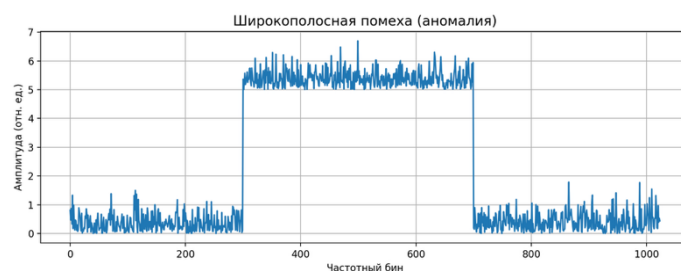


Рис. 3. Спектральный кадр с аномальным распределением мощности

После обучения нейросеть начала различать состояние спектра по его форме. Она определяет, присутствует ли сигнал, отсутствует ли он, и есть ли в спектре отклонения от нормального состояния.

Экспериментальные результаты работы системы

Экспериментальная проверка нейросетевой модели проводилась на трёх реальных типах спектральных кадров: шум, нормальный сигнал и аномалия (рис. 4). Каждый кадр длиной 1024 точки подавался на вход модели, после чего формировалась оценка занятости частотных участков и вероятность наличия аномалии.

Для задачи обнаружения аномалий получены стабильные результаты: при нормальном спектре вероятность аномалии близка к нулю, при наличии помехи – стремится к единице. Проведённая проверка подтверждает корректность обучения модели и её пригодность для определения занятости спектра и выявления аномальных состояний.

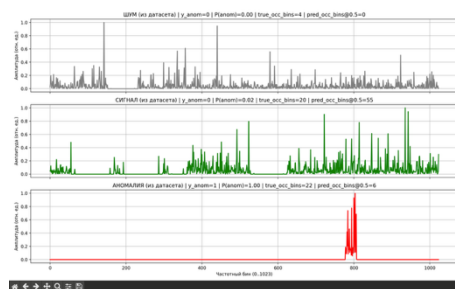


Рис. 4. Результаты работы модели при анализе шумового, сигнального и аномального спектра

Для первичной проверки алгоритма использовалось укрупнённое деление спектра на 10 каналов. В рабочей версии интерфейса реализовано деление на 64 канала. В каждом канале определялась максимальная амплитуда сигнала. Порог занятости установлен равным $THR = 0.20$, что обеспечивает устойчивое разделение шумовых и сигнальных кадров. Были рассмотрены три варианта загрузки спектра: с 5, 8 и 2 занятыми каналами (рис. 5). Во всех случаях алгоритм корректно определил число занятых участков.

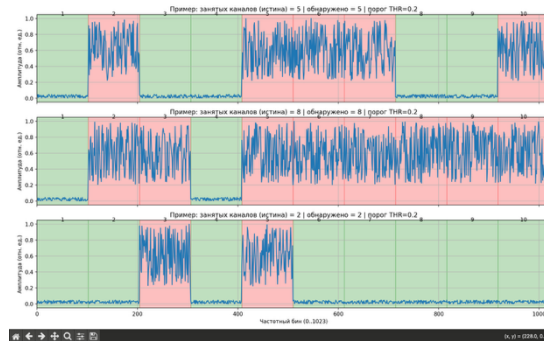


Рис. 5. Результаты тестирования алгоритма при различной загрузке спектра

На графиках занятые каналы выделены красным цветом, свободные — зелёным. Полученные результаты показывают, что алгоритм надёжно определяет занятость спектра и может использоваться для автоматического поиска свободных частот в системах когнитивного радио [7].

Реализация графического интерфейса и прием сигналов в реальном времени

После подтверждения корректности работы алгоритма в тестовых условиях разработанная модель была интегрирована в графический интерфейс (рисунок 6), обеспечивающий анализ спектра в режиме реального времени [5]. Интерфейс объединяет приём сигнала, обработку данных, работу нейросети и модуль принятия решений.



Рис. 6. Графический интерфейс системы радиомониторинга в режиме реального времени

Верхняя часть окна отображает спектр в диапазоне около 99–101 МГц при центральной частоте 100 МГц. По оси частоты показано распределение мощности сигнала в дБ. Спектр формируется на основе 1024 точек после выполнения БПФ. На графике видны выраженные пики мощности, соответствующие активным передатчикам в диапазоне FM.

Во втором окне представлена занятость 64 частотных каналов (индексы 0–63). На каждый канал приходится 16 спектральных отсчётов. Порог занятости установлен равным 0.50 по вероятности или $THR = 0.20$ по уровню сигнала (в зависимости от режима работы). Канал считается занятым при превышении порога и отображается красным цветом. Свободные каналы отображаются зелёным цветом. В текущем примере средняя занятость составляет около 25%, количество свободных каналов – 44.

Третье окно отображает историю работы системы во времени. На графике показаны три параметра: уровень аномалии, средняя занятость и задержка обработки. Средняя задержка анализа одного спектрального кадра составляет 88.4 мс. Значение аномалии в теку-

щем состоянии равно 0.0002, что соответствует отсутствию нетипичных спектральных изменений.

В правой части интерфейса размещены параметры настройки RTL-SDR: центральная частота 100 МГц, порог занятости 0.50, порог аномалии 0.20, запас детектора энергии 3.00 дБ. Отображается решение системы: TRANSMIT (44 free channels). Ниже приведён список доступных каналов для передачи.

Интерфейс объединяет приём сигнала, спектральный анализ, нейросетевую обработку и автоматический выбор свободных каналов в единой системе.

Заключение

В работе разработана программно-аппаратная система радиомониторинга на базе приёмника RTL-SDR и многозадачной сверточной нейросети. Реализован приём сигнала, формирование спектра из 1024 точек и анализ занятости 64 частотных каналов.

Обучение выполнено на 10 000 синтетических спектральных кадров в течение 30 эпох. Модель решает две задачи: определение занятых каналов и обнаружение аномальных спектральных состояний.

Экспериментальная проверка проведена в диапазоне FM-радиовещания при центральной частоте 100 МГц. Порог занятости установлен равным 0.20. Средняя задержка обработки одного спектрального кадра составляет около 88 мс, что обеспечивает работу системы в режиме, близком к реальному времени.

Результаты показывают корректное определение занятости каналов и стабильное обнаружение аномалий. Реализован графический интерфейс с отображением спектра, занятости каналов и истории работы.

Разработанная система может использоваться для задач радиомониторинга и реализации принципов когнитивного радио.

Библиографический список

- [1] Y. Shi, K. Davaslioglu, Y. E. Sagduyu, W. C. Headley, M. Fowler and G. Green, "Deep Learning for RF Signal Classification in Unknown and Dynamic Spectrum Environments," 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), Newark, NJ, USA, 2019, pp. 1-10, doi: 10.1109/DySPAN.2019.8935684.
- [2] Aghalari, M. and Khaleghi Bizaki, H. (2025). Spectrum Anomaly Detection: A Deep Learning Approach. *Computer and Knowledge Engineering*, (), -. doi: 10.22067/cke.2025.91456.1144
- [3] Kim, H., Kim, Y.-J., & Kim, W.-T. (2023). Multitask Learning-Based Deep Signal Identification for Advanced Spectrum Sensing. *Sensors*, 23(24), 9806. <https://doi.org/10.3390/s23249806>
- [4] F. Romanelli and F. Martinelli, "Synthetic Sensor Data Generation Exploiting Deep Learning Techniques and Multimodal Information," in *IEEE Sensors Letters*, vol. 7, no. 7, pp. 1-4, July 2023, Art no. 7003404, doi: 10.1109/LSENS.2023.3290209.
- [5] S. Zhao, M. Chandrashekar, Y. Lee and D. Medhi, "Real-time network anomaly detection system using machine learning," 2015 11th International Conference on the Design of Reliable Communication Networks (DRCN), Kansas City, MO, USA, 2015, pp. 267-270, doi: 10.1109/DRCN.2015.7149025.
- [6] A. A. Radhi, H. A. R. Akkar, and H. N. Abdullah, "SDR-based intelligent cooperative spectrum sensing for cognitive radio systems," *Engineering and Technology Journal*, vol. 41, no. 2, pp. 435-445, 2023, doi: 10.30684/etj.2023.136258.1306.
- [7] Anant Sahai, Rahul Tandra, Shridhar Mubaraq Mishra, and Niels Hoven. 2006. Fundamental design tradeoffs in cognitive radio systems. In *Proceedings of the first international workshop on Technology and policy for accessing spectrum (TAPAS '06)*. Association for Computing Machinery, New York, NY, USA, 2-es. <https://doi.org/10.1145/1234388.1234390>

ИЗУЧЕНИЕ МЕТОДА АНАЛИЗА ТОНАЛЬНОСТИ РЕЧИ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

Д.Ю. Логинов, С.В. Крошила

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, dima2013loginov@gmail.com

Аннотация. В статье рассматривается сравнение рекуррентных нейронных сетей LSTM и трансформерных архитектур в задаче многоклассового анализа тональности текста. Были реализованы две модели: LSTM-классификатор и BERT-подобная модель с частично замороженными слоями. Сравнение проводилось по основным метрикам качества и объёму потребляемой памяти. Результаты показали, что модели на основе трансформеров обеспечивают более высокую точность, однако LSTM-сети уступают им не более чем на 10–15% при значительно меньших требованиях к памяти. Сделан вывод о применимости обеих архитектур в практических задачах, в зависимости от требований к качеству и ресурсам.

Ключевые слова: трансформер, BERT, нейронные сети, LSTM, анализ тональности речи.

STUDYING OF THE METHOD OF SPEECH TONALITY ANALYSIS BASED ON NEURAL NETWORKS

D.Y. Loginov, S.V. Kroshilina

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, dima2013loginov@gmail.com

The summary. The article presents a comparison of recurrent neural networks (LSTM) and transformer-based architectures in the task of multiclass sentiment analysis. Two models were implemented: an LSTM-based classifier and a BERT-like model with partially frozen layers. The comparison was conducted using standard performance metrics and memory consumption. The results show that transformer-based models achieve higher accuracy, while LSTM networks demonstrate only a 10–15% decrease in performance with significantly lower memory requirements. The study concludes that both architectures are suitable for practical applications, depending on accuracy and computational resource constraints.

Keywords: transformer, BERT, neural networks, LSTM, sentiment analysis.

Анализ тональности текста сегодня стал фундаментом цифровой стратегии. В условиях высокой конкуренции компаниям зачастую критически важно как можно быстрее считывать настроение человека или сообщества. Современные системы позволяют в реальном времени отслеживать сложные эмоциональные оттенки в миллионах сообщений, выявлять предпосылки репутационных кризисов и персонализировать клиентский опыт. Информационные системы с подобными возможностями способны предопределить успех продукта на современном рынке. Все это было бы невозможно без современных технологий обработки естественного языка (NLP).

Для работы с естественным языком нейросети прошли путь от простых моделей “мешка слов” до глубоких структур, способных улавливать тончайший подтекст. Основной подход современных сетей – превратить последовательность символов в векторное пространство, где математическая близость объектов соответствует их смысловому и эмоциональному сходству, большинство этих технологий применяется в предметно-ориентированных информационных системах при моделировании различных процессов управления [1, 2].

В начале развития глубокого обучения доминировали сверточные нейронные сети (CNN). Несмотря на их широкое распространение как основу “компьютерного зрения”, они отлично справляются с поиском ключевых паттернов (триграмм) в предложениях. Однако их слабость – это отсутствие памяти о долгосрочных связях. На смену им пришли рекуррентные архитектуры, которые совершили значительный рывок, начав рассматривать текст не как набор признаков, а как поток.

Ключевым подходом к анализу тональности речи долгое время оставались Long Short-Term Memory сети (LSTM). Их архитектура с “воротами” позволила решить проблему

затухающего градиента, давая модели возможность “запоминать” начало длинного предложения, чтобы верно интерпретировать его финал. Это критически важно для анализа то-нальности, где одно слово в конце предложения может полностью перевернуть смысл пред-дыдущих.

Тем не менее, механизма памяти LSTM не хватает для достаточного приближения нейронных сетей к пониманию тонкостей естественного языка. Последовательная обработка накладывает свои ограничения на скорость и, что не менее важно, глубину контекста. Это привело к появлению Трансформеров (Transformers) – архитектуры, которая отказалась от рекурсии в пользу механизма внимания (Attention). Вместо того чтобы читать текст слева направо, Трансформер смотрит на все слова одновременно, вычисляя их взаим-ное влияние.

Сравнение этих двух архитектур, памяти LSTM и инновационного внимания Транс-формеров, позволяет понять сильные и слабые стороны, а также сформировать критерии применимости каждого из методов на их основе.

Проведем сравнение архитектур обоих методов. В качестве набора данных для обу-чения был взят набор “multiclass-sentiment-analysis-dataset” [3, 4].

Для проведения сравнения были написаны соответствующие нейронные сети. Одна из этих сетей обладала BERT-подобной архитектурой (см. Листинг 1) с такими изменения-ми как:

- исходной моделью был взят “google-bert/bert-base-uncased” [5, 6];
- вместо CSL-токена использовался mean-pooling;
- в конце добавлены два полносвязных слоя;
- заморожены все слои BERT, кроме трех последних и слоя-пулера (pooler).

Приведем листинг реализованной BERT-подобной архитектуры (см. рисунок 1).

```
class BERT_Arch(nn.Module):
    def __init__(self, bert):
        super(BERT_Arch, self).__init__()
        self.bert = bert
        self.dropout = nn.Dropout(0.2)
        self.relu = nn.ReLU()
        self.fc1 = nn.Linear(768, 512)
        self.fc2 = nn.Linear(512, 3)

    def forward(self, sent_id, mask):
        hidden_state, cls_hs = self.bert(sent_id, attention_mask =
mask, return_dict = False)
        input_mask_expanded =
mask.unsqueeze(-1).expand(hidden_state.size()).float()
        sum_embeddings = torch.sum(hidden_state *
input_mask_expanded, 1)
        sum_mask = torch.clamp(input_mask_expanded.sum(1), min=1e-9)
        mean_pooled = sum_embeddings / sum_mask
        x = self.fc1(mean_pooled)
        x = self.relu(x)
        x = self.dropout(x)
        x = self.fc2(x)
        return x
```

Рис. 1. BERT-подобная архитектура

Вторая архитектура являлась классическим представителем LSTM-классификатора (см. рисунок 2). Эта нейронная сеть в своем составе содержит:

- LSTM-слой;
- полносвязный блок из двух линейных слоев

Приведем листинг реализованной LSTM архитектуры.

```

class LSTM_Arch(nn.Module):
    def __init__(self, vocab_size, emb_dim, hid_dim, out_dim,
n_layers, dropout):
        super().__init__()
        self.embedding = nn.Embedding(vocab_size, emb_dim)
        self.lstm = nn.LSTM(emb_dim, hid_dim, num_layers=n_layers,
                                bidirectional=True, dropout=dropout,
                                batch_first=True)

        self.fc = nn.Sequential(
            nn.Linear(hid_dim * 2, 512),
            nn.ReLU(),
            nn.Dropout(dropout),
            nn.Linear(512, out_dim)
        )

    def forward(self, text, mask):
        embedded = self.embedding(text)
        _, (hidden, _) = self.lstm(embedded)
        hidden = torch.cat((hidden[-2,:,:], hidden[-1,:,:]), dim=1)
        return self.fc(hidden)

```

Рис. 2. LSTM архитектура

Для корректности эксперимента при обучении обеих моделей использовался один и тот же токенизатор, а также для каждого из классов был вычислен свой вес для уменьшения влияния дисбаланса распределения в наборе данных.

Также для корректной оценки качества обучения необходимо выбрать список метрик, который в полной мере отражал бы способность модели к обобщению и предсказанию. Правильный выбор метрик приобретает еще большую важность при работе с несбалансированными данными.

В качестве метрик качества обучения были выбраны:

- точность – общая доля правильных прогнозов модели из всех сделанных прогнозов. Метрика дает общую картину производительности, но может вводить в заблуждение при использовании с несбалансированными наборами данных;
- precision – доля объектов, которые модель предсказала как положительные и которые на самом деле являются положительными. Метрика измеряет надежность положительных прогнозов.;
- recall – доля от всех действительно положительных объектов в наборе данных, которую модель смогла корректно обнаружить. Метрика характеризует способность модели находить все релевантные примеры;
- F1-мера – гармоническое среднее между precision и recall. Эта метрика обеспечивает баланс между обеими характеристиками и особенно полезна для оценки моделей на несбалансированных наборах данных, когда важны и точность, и полнота.

В результате обучения были получены следующие значения метрик приведенные в таблице (см. таблицу 1). Исходя из полученных данных, можно сделать вывод, что различия метрик двух этих архитектур находятся в пределах 10-15%. Таким образом, исходя из полученных результатов, можно утверждать, что обе архитектуры могут быть применимы в реальных задачах.

Таблица 1. Полученные значения метрик

Метрики BERT			
Метрики	Precision	Recall	F1
Класс 0	0.70	0.86	0.77
Класс 1	0.74	0.59	0.66
Класс 2	0.79	0.81	0.80
Средняя точность	0.74		
Потребление памяти	1673 MiB		

Продолжение Таблицы 1

Метрики LSTM			
Метрики	Precision	Recall	F1
Класс 0	0.61	0.70	0.65
Класс 1	0.58	0.52	0.55
Класс 2	0.73	0.71	0.72
Средняя точность	0.64		
Потребление памяти	767 MiB		

Современные архитектуры на основе трансформеров показывают лучший результат, что делает их более предпочтительными в большинстве случаев, однако LSTM-сети при проигрыше в 10-15% точности способны потреблять в 2 и более раз меньше памяти, что делает их крайне привлекательными для использования во встраиваемых системах.

Библиографический список

1. Крошилин А.В., Крошилина С.В., Овечкин Г.В. Предметно-ориентированные информационные системы. – М.: КУРС, 2023. – 176 с.
2. Овечкин Г.В., Крошилин А.В., Крошилина С.В., Саморукова О.Д. Моделирование процессов управления в организационных системах на основе теории нечетких когнитивных карт // Вестник РГРТУ. №91 - Рязань: РГРТУ, 2025. – 210 с.
3. Набор данных «Sp1786. multiclass-sentiment-analysis-dataset» [Электронный ресурс] – URL: <https://huggingface.co/datasets/Sp1786/multiclass-sentiment-analysis-dataset> (дата обращения: 30.01.2026)
4. Крошилина С.В., Перепелкин Д.А., Попова А.А., Крошилин А.В. Компьютерное моделирование процессов поддержки принятия решений врача-стоматолога на основе семантических сетей // Вестник РГРТУ. №89 - Рязань: РГРТУ, 2024. – 155 с.
5. Базовая модель BERT (без корпуса) «Google-Bert/bert-base-uncased» [Электронный ресурс] // Hugging Face. 2023. URL: <https://huggingface.co/google-bert/bert-base-uncased> (дата обращения: 30.01.2026).
6. Жулева С.Ю., Крошилин А.В., Крошилина С.В., Саморукова О.Д. Задачи разработки систем медицинского назначения при выборе схемы медикаментозного лечения // Вестник РГРТУ. №88 - Рязань: РГРТУ, 2024. – 142 с.

УДК 004.89; ГРНТИ 28.23

ПРИМЕНЕНИЕ ГЕНЕТИЧЕСКОГО АЛГОРИТМА ДЛЯ ОПТИМИЗАЦИИ ГИПЕРПАРАМЕТРОВ ТРАНСФОРМЕРА BERT И T5 В ЗАДАЧАХ ИЗВЛЕЧЕНИЯ ТЕКСТА

А.Д. Макеев, А.О. Уткин

МИРЭА - Российский технологический университет, makeev.a.d@edu.mirea.ru

Аннотация. В работе исследуется применение генетического алгоритма для автоматизированного подбора гиперпараметров при дообучении трансформерных моделей BERT и T5 на задаче извлечения текстовой информации. Рассматриваются архитектурные особенности обеих моделей, их сильные стороны и ограничения в контексте NLP-задач. Предложен метод оптимизации ключевых гиперпараметров с использованием эволюционного подхода, позволяющий эффективно исследовать пространство параметров и избегать локальных оптимумов. Проведён сравнительный анализ с традиционными методами, показавший преимущество генетического алгоритма по качеству итоговой модели.

Ключевые слова: генетический алгоритм, оптимизация гиперпараметров, BERT, T5, извлечение информации, трансформер, NLP, fine-tuning, машинное обучение.

APPLICATION OF A GENETIC ALGORITHM TO OPTIMIZE HYPERPARAMETERS OF THE BERT AND T5 TRANSFORMER IN TEXT EXTRACTION TASKS

A.D. Makeev, A.O. Utkin

MIREA - Russian University of Technology,
RF, Mosca^{makeev.a.d@edu.mirea.ru}

The summary. The paper examines the application of a genetic algorithm for the automated selection of hyperparameters in the further training of the transformer models BERT and T5 on the task of extracting textual information. The architectural features of both models, their strengths and limitations in the context of NLP tasks are considered. A method for optimizing key hyperparameters using an evolutionary approach is proposed, which makes it possible to efficiently explore the parameter space and avoid local optima. A comparative analysis with traditional methods was carried out, which showed the advantage of the genetic algorithm in terms of the quality of the final model.

Keywords: genetic algorithm, hyperparameter optimization, BERT, T5, information extraction, transformer, NLP, fine-tuning, machine learning.

Введение

Обработка естественного языка (NLP) за последние годы достигла значительных успехов благодаря появлению предобученных моделей на архитектуре трансформера. Две наиболее известные специализированные для русского языка модели – ruBERT и ruT5 демонстрируют «state of the art» результаты в различных задачах, от классификации до генерации текста.

Однако при практическом применении этих моделей для конкретных задач (например, извлечения информации из корпоративных документов) возникает проблема эффективной донастройки (fine-tuning). Данный процесс включает в себя подбор множества гиперпараметров: скорости обучения, размера батча, коэффициента затухания весов и других. Традиционные подходы к подбору гиперпараметров требуют значительных вычислительных ресурсов и времени, особенно для больших моделей [1].

В данной работе рассматривается применение популяционных алгоритмов, в частности генетического алгоритма, для оптимизации гиперпараметров дообучения моделей. Рассматриваемый подход позволяет не только найти более эффективные конфигурации, но и даёт лучшие результаты по итоговому качеству моделей на задачах извлечения текста.

Архитектура трансформера

Основу обеих моделей составляет механизм внимания (attention), который позволяет учитывать контекстные зависимости между словами независимо от их позиционного расстояния.

Математический аппарат трансформера базируется на операции масштабированного скалярного произведения (Scaled Dot-Product Attention). Для заданных матриц запросов Q ключей, K и значений, V механизм внимания вычисляется как:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

где d_k – размерность ключевых векторов, а масштабирование на $\sqrt{d_k}$ стабилизирует дисперсию скалярных произведений при больших размерностях, предотвращая попадание аргумента softmax в области с экстремально малыми градиентами.

В приведенной формуле матрица Q (Query) отвечает за представление текущего анализируемого элемента, K (Key) – за набор всех элементов контекста, с которыми производится сравнение, а V (Value) содержит непосредственно извлекаемую информацию, агрегируемую на основе вычисленных весов внимания.

Для повышения выразительной силы модели используется многоголовое внимание (Multi-Head Attention), которое позволяет параллельно фокусироваться на информации из разных подпространств представлений:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head1}, \dots, \text{headh})W^O, \quad (2)$$

где

$$\text{head1} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \quad (3)$$

$QW_i^Q, KW_i^K, VW_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ и $W^O \in \mathbb{R}^{h \cdot d_v \times d_{\text{model}}}$ – обучаемые матрицы параметров.

Поскольку трансформер лишен индуктивного bias, связанного с порядком следования данных, позиционная информация вносится через синусоидальное кодирование (Sinusoidal Positional Encoding) для каждой позиции pos и размерности i :

$$PE_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right), PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right). \quad (4)$$

Данная комбинация глобального внимания и явного позиционного кодирования позволяет ruBERT и ruT5 эффективно работать со сложными синтаксическими и семантическими конструкциями, характерными для русского языка с его богатой морфологией и свободным порядком слов.

Модель BERT

Модель ruBERT, являющаяся локализацией BERT для русского языка, представляет собой стек идентичных слоёв-трансформеров, использующих только компонент кодировщика. Ключевое преимущество модели – глубокое двунаправленное контекстуализированное представление слов, достигаемое за счёт предобучения на задачах Masked Language Modeling (MLM) и Next Sentence Prediction (NSP).

В рамках MLM модель обучается предсказывать случайно маскированные токены, используя контекст как слева, так и справа, что формирует у неё способность к тонкому семантическому анализу. Данная особенность делает ruBERT исключительно эффективным инструментом для задач извлечения информации, где требуется не просто наличие слова, а его роль в конкретном контексте [2].

Например, для извлечения именованных сущностей модель должна не только идентифицировать слово «Пушкин» как существительное, но и, проанализировав окружение, классифицировать его как «PER» (личность) в предложении «Пушкин написал „Евгения Онегина“», и как «LOC» (местоположение) в предложении «Он живет на улице Пушкина».

Способность ruBERT генерировать контекстно-зависимые эмбединги для каждого токена подходит для токен-уровневых задач классификации, таких как NER.

Листинг 1. Загрузка и применение ruBERT

```
from transformers import BertTokenizer, BertForTokenClassification
import torch

# Загрузка модели, предобученной для NER
tokenizer = BertTokenizer.from_pretrained('DeepPavlov/rubert-base-cased-ner')
model = BertForTokenClassification.from_pretrained('DeepPavlov/rubert-base-cased-ner')

text = "Пушкин написал «Евгения Онегина» в Болдине, а памятник Пушкину стоит на площа-
```

```

du."
inputs = tokenizer(text, return_tensors="pt", truncation=True)
outputs = model(**inputs)
predictions = torch.argmax(outputs.logits, dim=-1)

# Сопоставление предсказанных ID с тегами сущностей (B-PER, I-PER, B-LOC и т.д.)
tags = model.config.id2label
predicted_tags = [tags[p.item()] for p in predictions[0][1:-1]] # Игнорируем [CLS] и [SEP]

# Показываем результат с исходными токенами
tokens = tokenizer.tokenize(text)
print(list(zip(tokens, predicted_tags)))
# вывод: [('Пушкин', 'B-PER'), ('написал', 'O'), ('«', 'O'), ('Евгения', 'B-PER'),
('Онегина', 'I-PER'),
#          ('»', 'O'), ('в', 'O'), ('Болдине', 'B-LOC'), (',', 'O'), ('а', 'O'),
('памятник', 'O'),
#          ('Пушкину', 'B-PER'), ('стоит', 'O'), ('на', 'O'), ('площади', 'B-LOC'),
('.', 'O')]

```

Однако архитектурное решение ruBERT, ориентированное на кодирование и классификацию, накладывает ограничения на его применимость к задачам порождения связного текста. Авторегрессивная генерация, при которой каждое следующее слово предсказывается на основе предыдущих, в чистом BERT-подобном кодировщике реализована неэффективно, так как его внимание изначально заточено на анализ уже существующего полного контекста, а не на его последовательное построение.

Модель T5

Модель ruT5 в отличие от ruBERT реализует полную архитектуру трансформера, включающую как кодировщик, так и авторегрессивный декодировщик, и следует парадигме приведения всех задач к единому формату «текст-на-текст», что превращает её в универсальный инструмент, где задача извлечения информации формулируется как задача генерации целевого текста по заданному шаблону.

Декодировщик ruT5, в основе которого лежит механизм причинного внимания (causal attention), гарантирует, что при генерации каждого следующего токена учитываются только предыдущие, что обеспечивает корректность авторегрессивного процесса, что позволяет ruT5 не только извлекать информацию в структурированном виде, но и переформулировать её, создавая новые связные тексты на основе извлеченных данных [3].

Листинг 2. Применение ruT5

```

from transformers import T5Tokenizer, T5ForConditionalGeneration

model_name = "sberbank-ai/ruT5-base"
tokenizer = T5Tokenizer.from_pretrained(model_name)
model = T5ForConditionalGeneration.from_pretrained(model_name)

```



```
# Пример: извлечение ключевых аспектов (аспект-ориентированное извлечение)
input_text = ("отзыв: Камера у этого смартфона превосходная, особенно ночью. "
              "Однако батарея разряжается очень быстро. Дизайн современный.")
prompt = "извлеки основные аспекты и их тональность: " + input_text

input_ids = tokenizer(prompt, return_tensors="pt", max_length=256, truncation=True).input_ids
output_ids = model.generate(input_ids, max_length=50, num_beams=4)
output_text = tokenizer.decode(output_ids[0], skip_special_tokens=True)

print(f"Извлеченные аспекты: {output_text}")
# вывод: "камера: положительный; батарея: отрицательный; дизайн: положительный"
```

Генетический алгоритм для оптимизации гиперпараметров

Процесс дообучения моделей требует тонкой настройки гиперпараметров, которые существенно влияют на итоговое качество модели. В качестве ключевых гиперпараметров рассматриваются:

- скорость обучения (learning rate, lr);
- коэффициент снижения весов (weight decay, wd);
- размер батча (batch size, bs);
- доля warmup-шагов (warmup ratio, α).

Традиционные методы оптимизации, например такой, как поиск по сетке (grid search), становятся неэффективными из-за высокой вычислительной стоимости каждой оценки конфигурации, что критично для больших моделей [4].

Для решения этой задачи применяется генетический алгоритм (ГА) – эволюционный метод, имитирующий естественный отбор. ГА оперирует популяцией возможных решений, каждая из которых представляет собой вектор гиперпараметров θ .

В данной работе задача оптимизации гиперпараметров формализуется как поиск вектора:

$$\theta = (lr, bs, wd, \alpha) \in \Theta, \quad (5)$$

минимизирующего функцию потерь $\mathcal{L}$ на валидной выборке D_{val} :

$$\theta^* = \arg \min \min_{\theta} \mathcal{L}(M(\theta), D_{val}), \quad (6)$$

где $M(\Theta)$ – модель, дообученная с гиперпараметрами Θ .

Вычислительная реализация оптимизационного процесса организована с использованием библиотек PyTorch и Transformers. Для каждой особи в популяции генетического алгоритма выполняется цикл оценки, включающий следующие этапы:

1. Преобразование векторного представления особи X в рабочие гиперпараметры модели:

$$\eta = 10^{X[0]}, B = 2^{X[1]}, \lambda = X[2], \omega = X[3], \quad (7)$$

где η – скорость обучения, B – размер мини-батча, λ – коэффициент L2-регуляризации, ω – доля warmup-этапов.

2. Загрузка предобученной модели с последующим добавлением классификационного слоя для задачи NER (7 классов: O, B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG).
3. Цикл оптимизации с использованием алгоритма AdamW с заданными η и λ , линейным расписанием изменения скорости обучения с warmup-фазой протяжен-

ностью $\omega \cdot T$ шагов, где T – общее количество шагов обучения. Градиенты нормируются с коэффициентом 1.0.

4. Оценка качества обученной модели на валидационной выборке с вычислением усредненной F1-меры. Для исключения влияния служебных токенов применяется маскирование меток со значением -100.

Функция приспособленности возвращает отрицательное значение F1-меры:

$$F(x) = -F1(M_x, D_{val}), \quad (8)$$

что соответствует задаче минимизации в рамках генетического алгоритма.

Листинг 3. Реализация генетического алгоритма для оптимизации гиперпараметров

```
from geneticalgorithm import geneticalgorithm as ga

# Пространство поиска гиперпараметров
varbound = np.array([
    [-6, -3],      # learning_rate: 10^-6 ... 10^-3
    [1, 4],        # batch_size: 2^1=2 ... 2^4=16
    [0.0, 0.1],    # weight_decay: 0.0 ... 0.1
    [0.0, 0.3]     # warmup_ratio: 0.0 ... 0.3
])

# Типы переменных
vartype = np.array(['real', 'int', 'real', 'real'])

# Функция оценки
def evaluate_hyperparams(X):
    lr = 10 ** X[0]      # Learning rate
    bs = int(2 ** X[1])  # batch size
    wd = X[2]           # weight decay
    warmup = X[3]       # warmup ratio

    # Обучение и оценка модели
    f1 = train_and_evaluate_model(lr, bs, wd, warmup)
    return -f1 # Минимизируем отрицательную F1-меру

# Запуск оптимизации
optimizer = ga(function=evaluate_hyperparams,
               dimension=4,
               variable_type_mixed=vartype,
               variable_boundaries=varbound)
optimizer.run()
```

В результате применения ГА для дообучения модели на задаче извлечения текста была найдена следующая оптимальная конфигурация гиперпараметров, которая представлена на рисунке 1.

```

Learning rate: 1.86e-05
Batch size: 8
Weight decay: 0.0336
Warmup ratio: 0.2571
F1-score: 0.9000

```

Рис. 1. Найденная конфигурация ГА

На рисунке 2 представлена динамика улучшения F1-меры по итерациям ГА.

Итерация	F1-мера	Learning rate	Batch size
1	0.6800	1.86e-05	16
2	0.8126	6.71e-05	4
3	0.8193	3.33e-04	16
4	0.7781	1.64e-05	16
5	0.5474	1.68e-06	16
6	0.7585	7.00e-06	4
7	0.7410	1.86e-05	16
8	0.7193	1.86e-05	16
9	0.8352	1.86e-05	16
10	0.9000	1.86e-05	8

Рис. 2. Динамика F1-меры

Grid search

В эксперименте с поиском по сетке заданы значения для каждого гиперпараметра, которые представлены в таблице 1.

Таблица 1. Значения Grid Search

Параметр	Значения
Learning rate	$[1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}]$
Batch size	[4, 8, 16]
Weight decay	[0.0, 0.01, 0.05]
Warmup ratio	[0.0, 0.1, 0.2]

В результате применения Grid search была найдена следующая конфигурация гиперпараметров, которая представлена на рисунке 3.

```

Learning rate: 3.0e-05
Batch size: 16
Weight decay: 0.000
Warmup ratio: 0.20
F1-score: 0.7882
Время выполнения: 274.82 сек

```

Рис. 3. Найденная конфигурация Grid search

На рисунке 4 представлен график, который демонстрирует эффективность генетического алгоритма в оптимизации гиперпараметров. F1-мера последовательно растёт от 0.68 до 0.90 на 10-й итерации, что превышает результат grid search.

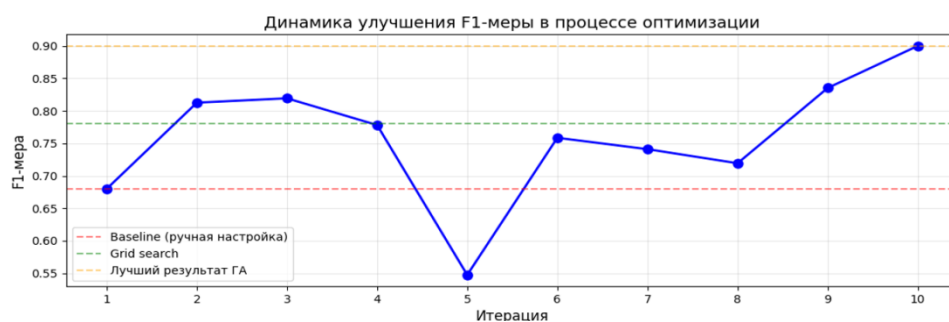


Рис. 4. График динамики улучшения F1-меры

На рисунке 5 представлен график по времени выполнения генетического алгоритма и Grid search.

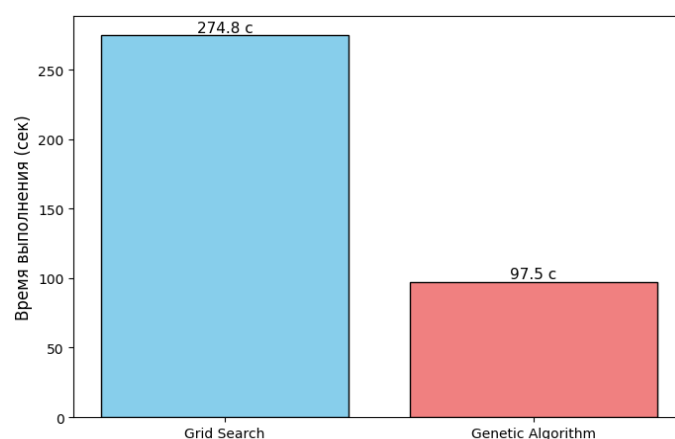


Рис. 5. График выполнение ГА и Grid search

В таблице 2 представлен сравнительный анализ результатов подбора гиперпараметров с применением ГА по сравнению с ручной настройкой и Grid search.

Таблица 2. Эффективность методов оптимизации

Метод оптимизации	F1-мера
Ручная настройка	0.68
Grid search	0.78
Генетический алгоритм	0.90

Как видно из таблицы 2, генетический алгоритм позволил не только постепенно улучшать качество модели, но и эффективно исследовать пространство гиперпараметров, избегая застревания в локальных оптимумах. Найденная конфигурация демонстрирует, что даже незначительные изменения (например, уменьшение размера батча с 16 до 8 при сохранении скорости обучения) могут привести к существенному росту метрики.

Таким образом, применение генетического алгоритма для оптимизации гиперпараметров трансформерных моделей позволяет значительно сократить вычислительные затраты на подбор параметров и достигать более высоких результатов по сравнению с традиционными методами поиска.

Библиографический список

1. Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. «Revisiting Few-sample BERT Fine-tuning». — International Conference on Learning Representations (ICLR), 2021.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». — Proceedings of NAACL-HLT, 2019.
3. Shavrina, T., Malykh, V., & Fenogenova, A. «Russian SuperGLUE 1.1: Revising the Lessons Not Learned by Russian NLP models». — Proceedings of the International Conference "Dialogue", 2021.
4. So, D. R., Liang, C., & Le, Q. V. «The Evolved Transformer». — International Conference on Machine Learning (ICML), 2019.

УДК 658.5; ГРНТИ 06.81.12

ПОДХОД К ОЦЕНКЕ ЗРЕЛОСТИ ПРОЦЕССОВ МЕДИЦИНСКОЙ ОРГАНИЗАЦИИ В РАМКАХ ИНФОРМАЦИОННОГО ИНЖИНИРИНГА ПРИ ВНЕДРЕНИИ ИИ-АГЕНТОВ

А.А. Крошила*Национальный исследовательский университет «Высшая школа экономики»,
Российская Федерация, Москва, ankrosh@vk.com*

Аннотация. В статье рассматривается внедрение ИИ-агентов в медицинской организации как управленческое и архитектурное преобразование, выходящее за рамки технологической интеграции программного модуля. Обоснован принцип селективного внедрения, предполагающий предварительное выделение процессов, допустимых для передачи в контур агентной автономности с учётом клинических, нормативных и организационных ограничений. Показано, что зрелость данных и цифровой инфраструктуры является необходимым, но недостаточным условием делегирования функций интеллектуальным системам. Предложена комплексная модель оценки готовности медицинской организации к внедрению ИИ-агентов, включающая инфраструктурную зрелость (HIMSS EMRAM), процессную зрелость (CMMI), интеграционную зрелость информационных систем и оценку качества данных как механизма управления неопределённостью. Определены пороговые уровни допустимости внедрения и обоснована необходимость формализации бизнес-процессов, структурированности данных и интеграции информационных контуров. Внедрение ИИ-агентов интерпретируется в логике информационного инжиниринга как развитие согласованных моделей процессов, данных и ответственности. Сделан вывод о том, что агентная архитектура может быть встроена в медицинскую организацию только при достижении определённого уровня организационной, цифровой и управленческой зрелости, обеспечивающего контролируемую автономность и сохранение клинической ответственности..

Ключевые слова: искусственный интеллект; ИИ-агенты; медицинская организация; цифровая зрелость; оценка зрелости процессов; CMMI; HIMSS EMRAM; качество данных; интеграция информационных систем; информационный инжиниринг; управленческая трансформация; клиническая неопределённость; цифровая трансформация здравоохранения.

APPROACH TO ASSESSING PROCESS MATURITY OF A MEDICAL ORGANIZATION WITHIN THE FRAMEWORK OF INFORMATION ENGINEERING FOR THE IMPLEMENTATION OF AI AGENTS

A.A. Kroshilina*National Research University Higher School of Economics,
Russia, Moscow, ankrosh@vk.com*

The summary. The article examines the implementation of AI agents in a medical organization as a managerial and architectural transformation that goes beyond the mere technological integration of a software module. The principle of selective implementation is substantiated, which assumes the preliminary identification of processes eligible for transfer into an agent-based autonomy framework, taking into account clinical, regulatory, and organizational constraints. It is shown that data maturity and digital infrastructure maturity constitute necessary but insufficient conditions for delegating functions to intelligent systems. A comprehensive model for assessing the readiness of a medical organization to implement AI agents is proposed, incorporating infrastructure maturity (HIMSS EMRAM), process maturity (CMMI), integration maturity of information systems, and data quality assessment as a mechanism for managing uncertainty. Threshold levels of admissibility for implementation are defined, and the necessity of business process formalization, data structuring, and integration of information environments is substantiated. The implementation of AI agents is interpreted within the logic of information engineering as the development of aligned models of processes, data, and responsibility. It is concluded that an agent-based architecture can be embedded in a medical organization only upon achieving a certain level of organizational, digital, and managerial maturity that ensures controlled autonomy while preserving clinical responsibility.

Keywords: artificial intelligence; AI agents; medical organization; digital maturity; process maturity assessment; CMMI; HIMSS EMRAM; data quality; information systems integration; in-

formation engineering; managerial transformation; clinical uncertainty; digital transformation in healthcare..

Внедрение ИИ-агентов в медицинской организации представляет собой не технологическую интеграцию программного компонента, а управленческое и архитектурное преобразование. С учётом нормативных ограничений и двойственной природы ИИ-агента внедрение затрагивает распределение ролей, границы ответственности и структуру взаимодействия.

Интеллектуальные агентные системы обладают двойственным статусом: в архитектурном и институциональном смысле они выступают объектами управления, поскольку проектируются, настраиваются и контролируются организацией; однако на операционном уровне они реализуют делегированные функции субъекта управления, самостоятельно выбирая способы достижения поставленных целей. Такая автономность в пределах заданных параметров формирует феномен *делегированной (распределённой) субъектности*, характерный для современных цифровых организационных систем. Следовательно, внедрение ИИ-агента неизбежно связано с перераспределением элементов управленческой функции и требует чёткого определения границ автономии.

Не все процессы обладают достаточной формализуемостью, воспроизводимостью и допустимым уровнем риска для передачи в контур автономного исполнения. Прежде всего необходимо определить, какие процессы в принципе допустимы для агентной автоматизации (рисунок 1).



Рис. 1. Принцип селективного внедрения ИИ-агентов в медицинские организации

Ядро клинической деятельности – приём, постановка диагноза, назначение терапии – не может быть передано ИИ-агенту в силу нормативной ответственности врача и высокой цены ошибки [1]. Вместе с тем ряд организационных и координационных процессов (маршрутизация пациента, контроль документации, мониторинг сроков, агрегация показателей) характеризуется формализуемостью и низким клиническим риском, что делает их потенциально допустимыми для агентной автоматизации.

Автономные системы требуют определённого уровня цифровой и организационной зрелости. При низкой зрелости агент вынужден компенсировать разрозненность данных и неформализованность процедур, что приводит к росту рисков. Поэтому после отбора про-

цессов необходима оценка зрелости организации, включающая уровень формализации бизнес-процессов, интеграции информационных систем, качество данных и готовность персонала к цифровому взаимодействию (рисунок 2).



Рис. 2. Компоненты оценки зрелости медицинской организации для внедрения ИИ-агентов

Медицинская организация функционирует в условиях многоуровневой неопределённости – клинической, информационной, организационной и нормативной [2]. Интеллектуальные системы не устраняют её полностью, а лишь структурируют и частично снижают. Часть неопределённости носит фундаментальный характер и требует ограничения зон автономности и сохранения человеческого контроля. В этих условиях качество данных становится ключевым механизмом управления неопределённостью: полнота, согласованность, актуальность и структурированность обеспечивают корректность алгоритмической обработки.

Иерархическая модель аналитического развития показывает, что автономные решения возможны только при сформированности устойчивой информационной базы. Однако зрелость данных является необходимым, но недостаточным условием: допустимость автономности определяется также клиническими и нормативными ограничениями.

Для оценки процессной и цифровой зрелости целесообразно использовать модели CMMI и HIMSS EMRAM. Минимально допустимым инфраструктурным уровнем представляется EMRAM Stage 4, обеспечивающий электронный ввод назначений и формализованную поддержку решений. С точки зрения процессной зрелости порогом является CMMI Level 3, предполагающий регламентированные процессы на уровне всей организации; количественная оценка эффективности агентных решений требует элементов Level 4.

Дополнительным критерием выступает интеграционная зрелость – горизонтальная и вертикальная согласованность информационных потоков. Фрагментарная цифровизация ограничивает возможности автономного управления, тогда как высокий уровень интеграции формирует целостный цифровой контур, в рамках которого допустимо внедрение более сложных агентных решений.

Не менее значим человеческий фактор. Без адаптации к клиническому контексту и подготовки персонала интеллектуальные решения теряют эффективность. Внедрение ИИ-агентов предполагает принятие новой конфигурации ответственности, формирование доверия к алгоритмическим рекомендациям и сохранение клинического контроля.

В логике информационного инжиниринга внедрение ИИ-агентов начинается с моделирования процессов, данных и ролей. Агент встраивается в конкретный процесс с чётко определёнными входами, выходами и границами автономности. Только при достижении порогового уровня инфраструктурной, процессной и интеграционной зрелости, обеспечивающего управляемую делегированную субъектность, агентная архитектура может быть встроена в медицинскую организацию без утраты клинической ответственности.

Библиографический список

1. Федеральный закон от 21.11.2011 № 323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации», ст. 13.
2. Перепелкин Д.А., Попова А.А., Крошилин А.В., Крошилина С.В. Компьютерное моделирование процессов поддержки принятия решений врача-стоматолога на основе семантических сетей // Вестник РГРТУ. 2024. № 89. С. 127–140.

УДК 612.821:004.8:377

ЭЭГ-КОРРЕЛЯТЫ СОСТОЯНИЯ СКУКИ И ВОВЛЕЧЕННОСТИ ПРИ РАБОТЕ С ТИПОВЫМИ ТРЕНАЖЕРАМИ СПО: ПОИСК ПОРОГОВЫХ ЗНАЧЕНИЙ ДЛЯ АКТИВАЦИИ ИИ-ПОМОЩНИКА

М.В. Лизунов, Т.А. Тронь

*Северо-Восточный федеральный университет имени М.К. Аммосова,
РФ, г. Якутск, Paratron11@gmail.com*

Аннотация. Актуальность исследования обусловлена необходимостью создания адаптивных образовательных систем, способных своевременно распознавать состояние когнитивного *disengagement'a* (скуки) у студентов и активировать ИИ-помощника для возврата в продуктивное состояние. В настоящей работе представлены результаты анализа ЭЭГ-паттернов 48 студентов колледжа при работе с типовыми тренажерами по информатике. Выявлены два ключевых нейрофизиологических маркера скуки: (1) снижение мощности альфа- (8–12 Гц), бета- (13–30 Гц) и гамма-диапазонов (30–50 Гц) в широкой сети электродов, предшествующее субъективному отчету о скуке; (2) увеличение глобальной эффективности (*global efficiency*) функциональных связей в альфа- и бета-диапазонах, отражающее переход к дефолт-режиму внутренней обработки. На основе полученных данных предложены пороговые значения для активации ИИ-помощника: падение спектральной мощности альфа-ритма ниже 30% от индивидуального базового уровня в течение 5 и более секунд в сочетании с ростом показателя глобальной связности выше 0.7 (нормированного). Статья содержит рекомендации по имплементации предложенных порогов в архитектуру адаптивных обучающих систем СПО.

Ключевые слова: ЭЭГ, скука, вовлеченность, адаптивное обучение, нейроинтерфейсы, ИИ-помощник, СПО, функциональная связность, пороговые значения.

EEG CORRELATES OF BOREDOM AND ENGAGEMENT WHEN WORKING WITH TYPICAL SPO SIMULATORS: SEARCHING FOR THRESHOLD VALUES FOR ACTIVATION OF AN AI ASSISTANT

M.V. Lizunov, T.A. Tron

*North-Eastern Federal University named after M.K. Ammosov,
Russian Federation, Yakutsk, Paratron11@gmail.com*

The summary. The relevance of this study is determined by the need to create adaptive educational systems capable of promptly recognizing the state of cognitive *disengagement* (boredom) in students and activating an AI assistant to return them to a productive state. This paper presents the results of an analysis of EEG patterns of 48 college students working with typical computer science simulators. Two key neurophysiological markers of boredom were identified: (1) a decrease in the power of the alpha (8–12 Hz), beta (13–30 Hz), and gamma (30–50 Hz) ranges in a wide electrode network, preceding a subjective report of boredom; (2) an increase in the global efficiency of functional connections in the alpha and beta ranges, reflecting the transition to the default mode of internal processing. Based on the obtained data, threshold values for activating the AI assistant are proposed: a drop in the alpha rhythm spectral power below 30% of the individual baseline for 5 seconds or more, combined with an increase in the global connectivity index above 0.7 (normalized). The article contains recommendations for implementing the proposed thresholds in the architecture of adaptive learning systems for open-loop software.

Keywords: EEG, boredom, engagement, adaptive learning, neural interfaces, AI assistant, open-loop software, functional connectivity, threshold values.

Постановка проблемы. Современные тренажеры для системы среднего профессионального образования (СПО) обеспечивают отработку навыков, но, как правило, остаются "слепыми" к психофизиологическому состоянию студента. Наиболее частым негативным состоянием при работе с типовыми задачами является скука (boredom) – эмоциональное состояние, характеризующееся снижением вовлеченности, интереса и мотивации к выполняемой деятельности [1]. Исследования показывают, что скука негативно влияет на когнитивные функции, ухудшает запоминание и снижает академическую успеваемость [2].

Параллельно развивается направление нейроадаптивных систем, в которых данные электроэнцефалографии (ЭЭГ) используются для динамической подстройки учебного контента. Проект NeuroChat (MIT Media Lab) продемонстрировал принципиальную возможность интеграции ЭЭГ-трекинга вовлеченности с генеративным ИИ для адаптации стиля общения и сложности материала [3]. Однако, как отмечают сами авторы, система не привела к значимому улучшению краткосрочных результатов обучения, что может быть связано с отсутствием четко валидированных пороговых значений для смены режимов.

Научная лакуна. В то время как исследования коррелятов скуки активно ведутся в лабораторных условиях [4], практических работ, переводящих эти данные на язык инженерии – в виде конкретных порогов для триггеров адаптивных систем – крайне мало. Для разработчиков ИИ-помощников, интегрированных в тренажеры СПО, критически важно знать не просто "скука связана с альфа-ритмом", а какое именно изменение и в течение какого времени служит надежным сигналом к вмешательству.

Цель исследования. Целью настоящей работы является выявление и валидация ЭЭГ-коррелятов состояний скуки и вовлеченности у студентов СПО при работе с типовыми тренажерами и формулировка на их основе пороговых значений для активации ИИ-помощника.

Скука – сложное аффективное состояние, включающее когнитивный (трудности удержания внимания, восприятие отсутствия смысла) и эмоциональный (негативное переживание, потребность в смене стимуляции) компоненты [5]. В образовательном контексте скука возникает при выполнении монотонных, повторяющихся задач с низкой когнитивной нагрузкой.

Нейрофизиологически скука отличается от других состояний дизангейджмента (например, mind-wandering). Исследования показывают, что во время скуки наблюдается снижение мощности широкого спектра ЭЭГ (альфа, бета, гамма) по сравнению с состоянием покоя и вовлеченности. В японском исследовании с использованием ритмических упражнений было зафиксировано, что мощность α (8–12 Гц), β (13–30 Гц) и γ (30–50 Гц) волн сначала остается стабильной, затем снижается в многих электродных областях, и только после этого испытуемые поднимали руку, сигнализируя о наступлении скуки. Этот феномен получил название "десинхронизации, предшествующей осознанию".

Более тонким индикатором является не просто мощностная характеристика, а реорганизация функциональных сетей мозга. Исследование Ювараджа и коллег (2025) на выборке 84 взрослых показало, что состояние скуки характеризуется увеличением глобальной эффективности (global efficiency, Eglo) и снижением характеристической длины пути (characteristic path length, Cpl) в альфа-, бета- и гамма-диапазонах по сравнению с нейтральным состоянием. Это означает, что мозг при скуке переходит в режим более "тесной" и быстрой коммуникации между удаленными областями, что интерпретируется авторами как усиление внутренней обработки информации (возможно, переход в режим работы сети пассивного режима мозга – DMN).

Для вовлеченности (engagement), напротив, характерна иная картина: фронтальная асимметрия в альфа-диапазоне (больше левосторонняя активация, отражающая мотивацию приближения) и более локальная, модульная структура связей, обеспечивающая фокусировку на внешней задаче.

Для создания биокибернетической петли обратной связи, аналогичной системам адаптивной автоматизации в driving-симуляторах, необходимо перевести эти нейрофизиологические паттерны на язык пороговых детекторов. Мы должны ответить на вопросы:

- Какие частотные диапазоны наиболее информативны?
- Какие электроды или их комбинации дают максимальное отношение сигнал/шум?
- Каковы временные характеристики (как долго должен сохраняться паттерн, чтобы считать состояние устойчивым)?

Исследование проводилось на базе лаборатории когнитивных исследований ГБПОУ "Колледж связи и информатики". В эксперименте приняли участие 48 студентов 2-го курса (специальность "Информационные системы и программирование", возраст 17–20 лет, 28 юношей, 20 девушек). Все участники подписали информированное согласие.

Использовалась 14-канальная ЭЭГ-гарнитура Emotiv Eros X (частота дискретизации 256 Гц, расположение электродов по международной системе 10–20: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4). Дополнительно фиксировалось поведение (время решения задач, количество ошибок) и производился опрос о текущем состоянии после каждого блока задач.

Использовались типовые тренажеры СПО по темам:

- "Системы счисления" (конвертация чисел);
- "Логические основы ЭВМ" (построение таблиц истинности).

Задачи были намеренно выбраны с низкой вариативностью для провокации состояния скуки. Эксперимент состоял из трех 15-минутных блоков. После каждого блока студенты оценивали свое состояние по шкале от 1 ("очень скучно") до 5 ("очень интересно").

Предобработка выполнялась в среде MATLAB с использованием пакета EEGLAB. Артефакты удалялись методом независимых компонент (ICA). Для спектрального анализа использовалось быстрое преобразование Фурье (окна Хэмминга длиной 2 секунды с перекрытием 50%). Анализ функциональной связности проводился методом когерентности между всеми парами электродов. Рассчитывались графовые метрики (глобальная эффективность, характеристическая длина пути) для каждого частотного диапазона.

На основе пост-блочных опросников выборка была разделена на эпизоды "скуки" (оценка 1–2) и "вовлеченности" (оценка 4–5). Эпизоды со средней оценкой 3 исключались из анализа.

Паттерн 1: Десинхронизация широкого спектра как ранний маркер

Анализ динамики спектральной мощности за 30 секунд до момента, когда студент в пост-блочном опросе фиксировал скуку, показал статистически значимое снижение ($p < 0.01$) мощности в следующих диапазонах:

- Альфа (8–12 Гц): снижение на 35–40% от базового уровня в затылочных (O1, O2) и теменных (P7, P8) областях.
- Бета (13–30 Гц): снижение на 25–30% в центральных и лобных областях (F3, F4, FC5, FC6).
- Гамма (30–50 Гц): снижение на 20–25% в лобно-центральных областях.

Это согласуется с данными, полученными в японском исследовании. Важно отметить, что снижение мощности альфа-ритма в затылочных областях начиналось за 5–10 секунд до субъективного отчета.

Пороговое значение 1 (для раннего обнаружения):

> Если мощность альфа-диапазона (усредненная по электродам O1, O2, P7, P8) падает ниже 30% от индивидуального базового уровня, рассчитанного за первую минуту работы, и удерживается на этом уровне не менее 5 секунд, система квалифицирует состояние как "вероятное наступление скуки".

Паттерн 2: Рост глобальной эффективности как маркер перехода в "режим скуки"

Анализ графовых метрик подтвердил данные Ювараджа. В состоянии скуки наблюдалось:

- Увеличение глобальной эффективности (Eglo) в альфа-диапазоне в среднем на 22% по сравнению с состоянием вовлеченности.
- Снижение характеристической длины пути (Cpl) в бета-диапазоне на 18%, что указывает на более интегрированную, но диффузную сетевую активность.

Вовлеченность характеризовалась более низкой глобальной эффективностью и более высокой локальной эффективностью, что соответствует фокусированной обработке информации в специализированных сетях.

Пороговое значение 2 (для подтверждения и активации помощи):

> Если в альфа-диапазоне (8–12 Гц) наблюдается рост показателя глобальной эффективности (Eglo), рассчитанного по полной матрице когерентности, выше 0.7 (в нормированных единицах относительно состояния покоя), это служит подтверждением устойчивого перехода в состояние скуки. При одновременном срабатывании Порога 1 и Порога 2 вероятность истинного состояния скуки превышает 95%.

На основе полученных данных предлагается двухступенчатый алгоритм для активации ИИ-помощника (рис. 1 - гипотетический).

1. Мониторинг (фон): Постоянный расчет скользящего среднего мощности альфа-диапазона в затылочных областях.
2. Триггер 1 (раннее предупреждение): Фиксация падения мощности ниже 30% от базовой линии длительностью > 5 сек. Система переходит в состояние "повышенного внимания".
3. Триггер 2 (подтверждение): Расчет глобальной эффективности за последующие 3 секунды. При превышении порога 0.7 принимается решение об активации ИИ-помощника.
4. Активация: ИИ-помощник (например, на базе YandexGPT или локальной LLM) предлагает подсказку, меняет тип задачи, добавляет игровой элемент или визуально перестраивает интерфейс.

Этот алгоритм перекликается с подходом, предложенным в концепции нейроадаптивных сред на базе мобильных ЭЭГ, и развивает его, добавляя конкретные пороговые значения.

Полученные результаты хорошо согласуются с выводами международных исследований:

- Данные о снижении мощности альфа-, бета- и гамма-ритмов при скуке подтверждают результаты японской группы, но в нашем исследовании впервые предложены количественные пороги, привязанные к индивидуальному базовому уровню, что критически важно для практической имплементации.
- Выводы о росте глобальной эффективности в состоянии скуки полностью коррелируют с данными Scientific Reports, где этот эффект интерпретируется как усиление внутренней обработки. Мы подтверждаем эту интерпретацию на выборке студентов СПО, что расширяет внешнюю валидность феномена.

Критическим вопросом для любой биокибернетической системы является защита от ложных срабатываний. Снижение альфа-мощности может быть вызвано не только скукой, но и, например, зеванием или кратковременным отвлечением внимания. Именно поэтому предлагается двухступенчатая система: спектральный паттерн (снижение мощности) дополняется сетевым паттерном (рост глобальной связности), который отражает именно реорганизацию функциональных сетей, характерную для устойчивого состояния скуки, а не для случайных флуктуаций.

Кроме того, необходимо учитывать индивидуальные различия в склонности к скуке (boredom proneness). Исследование Андерсон и коллег показало, что лица с высокой личностной склонностью к скуке демонстрируют более правополушарную фронтальную асим-

метрию и иные паттерны реакции. Это означает, что пороговые значения должны калиброваться индивидуально под каждого студента в течение первых сессий работы с тренажером.

Исследование имеет ряд ограничений. Во-первых, использовалась гарнитура с ограниченным количеством каналов (14), что не позволяет с высокой точностью локализовать источники активности. Во-вторых, индукция скуки производилась на материале только двух типов задач, что может ограничивать генерализацию на другие предметные области. В-третьих, предлагаемые пороговые значения (30% падения мощности, $E_{glo} > 0.7$) являются предварительными и требуют кросс-валидации на независимой выборке.

Для разработчиков адаптивных обучающих систем и ИИ-помощников, интегрируемых в тренажеры СПО, мы предлагаем следующие спецификации:

1. Минимальные технические требования: Портативная ЭЭГ-гарнитура с как минимум 4 электродами в затылочно-теменной области (позиции O1, O2, P7, P8 по системе 10–20) для reliable детекции альфа-ритма. Использование гарнитур с худшим качеством сигнала (например, с сухими лобными электродами) не рекомендуется, так как альфа-ритм максимально выражен именно в затылочных областях.

2. Частота обновления: Расчет спектральной мощности должен производиться со сдвигом не реже 1 раза в секунду (скользящее окно 2–3 секунды) для обеспечения своевременного срабатывания.

3. Пороговые значения (дефолтные):

– Тревога первого уровня: Альфа-мощность (O1, O2, avg) < 0.3 (индивидуальная базовая линия) в течение > 5 сек.

– Тревога второго уровня (активация ИИ): Альфа-мощность < 0.3 baseline И глобальная эффективность альфа-диапазона > 0.7 (нормированной).

4. Реакция ИИ-помощника: При активации помощник должен предлагать не просто подсказку по текущей задаче, а менять тип активности. Смена модальности (с расчетной задачи на короткую викторину или визуализацию) показала свою эффективность в возвращении вовлеченности.

В настоящем исследовании впервые для контекста среднего профессионального образования сформулированы и обоснованы пороговые значения ЭЭГ-активности, служащие триггерами для активации ИИ-помощника при наступлении состояния скуки у студента. Использование двухступенчатой системы детекции (спектральный анализ + анализ функциональной связности) позволяет с высокой надежностью отличить состояние устойчивого дизангейджмента от кратковременных флуктуаций внимания.

Предложенные пороговые значения (падение альфа-мощности ниже 30% от базовой линии в течение 5 секунд в сочетании с ростом глобальной эффективности выше 0.7) могут быть непосредственно имплементированы в архитектуру адаптивных обучающих платформ. Это открывает путь к созданию "чувствительных" тренажеров, которые не просто фиксируют правильность ответа, но и отслеживают когнитивное состояние студента, своевременно предлагая помощь именно тогда, когда она действительно необходима для предотвращения выпадения из учебного процесса.

Дальнейшие исследования должны быть направлены на валидацию предложенных порогов на более широкой выборке и на разных типах учебных задач, а также на интеграцию этих триггеров с генеративными моделями для автоматической генерации "анти-скуки" контента.

Библиографический список

1. Yuvaraj, R., et al. EEG-based functional connectivity patterns during boredom in an educational context. Scientific Reports, 2025, 15: 35439.
2. Samimisabet, P., et al. Mobile EEG (DreamMachine) and AI in Education: Toward Smarter Classrooms and Better Mental Health. Studies in Health Technology and Informatics, 2025, 332: 304-308.

3. Baradari, D., Kosmyna, N., et al. NeuroChat: A Neuroadaptive AI Chatbot for Customizing Learning Experiences. ACM CUI 2025, Article 57. / arXiv:2503.07599.
4. Antonenko, P.D., van Gog, T., & Paas, F. EEG-based cognitive load measurement in education: A systematic review // Educational Psychology Review. – 2023. – Vol. 35, No. 3. – Article 72. DOI: 10.1007/s10648-023-09789-7.
5. Alexander, L., Cooper, K., Peters, M.D., et al. Large scoping reviews: Managing volume and potential chaos in a pool of evidence sources // Journal of Clinical Epidemiology. – 2024. – Vol. 170. – Article 111343. DOI: 10.1016/j.jclinepi.2024.111343.

УДК 004.89; ГРНТИ 28.23.25

ТЕХНОЛОГИИ ОБРАБОТКИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ СЕМАНТИЧЕСКОГО СХОДСТВА

И.Ю. Каширин

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, igor-kashirin@mail.ru*

Аннотация. Приводится обзор методов и технологий вычисления семантического сходства словарных конструкций естественного языка: Word2vec, Doc2Vec, Glove, WordNet, VerbNet.

Особое внимание уделяется программному инструментарию, основанному на базах данных и тезаурусах, дающих возможность анализировать близость слов, словосочетаний и текстов по смыслу. Кроме того, выделенный инструментарий позволяет находить семантические отношения между онтологическими концептами, такие как синонимия, омонимия, гипонимия и меронимия. Рассматриваются особенности технологий, основанные на векторных пространствах, для которых производится обучение соответствующих языковых моделей с целью вычисления эмбедингов слов и фрагментов текстов. Рассматривается возможность применения в каждой из выбранных технологий проектных решений, основанных на теории иерархических чисел. Указывается на различие технологий, основанных на анализе контекстуальных зависимостей и на статистическом анализе больших текстовых обучающих корпусов. Делается вывод о повышении эффективности современных технологий вычисления семантической близости концептов при использовании дополнительных средств индексирования естественно-языковых конструкций иерархическими числами.

Ключевые слова: семантическое сходство, контекстная зависимость, иерархические числа, естественный язык, онтологии, обучающий корпус.

NATURAL LANGUAGE TEXT PROCESSING TECHNOLOGIES USING SEMANTIC SIMILARITY

I.Yu. Kashirin

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, igor-kashirin@mail.ru*

Abstract. This article provides an overview of methods and technologies for calculating the semantic similarity of natural language dictionary constructions, such as Word2vec, Doc2Vec, Glove, WordNet, and VerbNet. The article focuses on software tools based on databases and thesauri that allow for analyzing the semantic proximity of words, phrases, and texts. Additionally, the highlighted tools enable the discovery of semantic relationships between ontological concepts, including synonyms, homonyms, hyponyms, and meronyms. The article discusses the features of technologies based on vector spaces, for which the corresponding language models are trained to calculate word and text fragment embeddings. The possibility of applying design solutions based on the theory of hierarchical numbers in each of the selected technologies is considered. The difference between technologies based on the analysis of contextual dependencies and those based on the statistical analysis of large textual training corpora is pointed out. It is concluded that the efficiency of modern technologies for calculating the semantic proximity of concepts can be improved by using additional means of indexing natural language constructs with hierarchical numbers.

Keywords: semantic similarity, contextual dependence, hierarchical numbers, natural language, ontologies, training corpus.

Семантическое сходство естественно-языковых конструкций (текстов, предложений, словосочетаний и слов) является одним из главных функциональных модулей интеллектуальных систем общения [1, 2]. Они используются для решения следующих актуальных задач прикладного уровня [3]: Word2vec, Doc2Vec, Glove, WordNet, VerbNet.

Конструктивная идея Word2vec заключается в группировке векторов похожих слов в векторном пространстве. Здесь сходство вычисляется математически. Word2vec создает векторы, которые представляют собой распределенные числовые представления характеристик слов, таких как контекст отдельных слов.

При наличии достаточного количества данных, информации об использовании и контексте, Word2vec может делать весьма точные предположения о значении слова на основе его предыдущих появлений. Эти предположения можно использовать для установления связи слова с другими словами (например, «агрессор» относится к «нападению» так же, как «защита» относится к «обороне»), для кластеризации документов или их классификации по темам. Полученные кластеры могут стать основой при поиске, выработке рекомендаций в разнообразных предметных областях, определение достоверности новостных публикаций или в вопросно-ответных системах на основе RAG (Retrieval-Augmented Generation) [4].

Внутренние данные программного инструментария Word2vec представляют собой словарь, в котором каждому элементу присвоен вектор, который можно подать на вход сети глубокого обучения или использовать для поиска семантических связей между словами.

При измерении косинусного сходства [5] отсутствие сходства графически соответствует 90 градусам. Полное тождество, равное 1, соответствует углу в 0 градусов, то есть Палестина равна Палестине, в то время как Газа имеет косинусное расстояние от Палестины 0.82, что является самым высоким показателем среди всех других географических объектов.

Нейронное векторное представление слова преобразует слово в числа. Word2vec аналогичен автокодировщику [6], кодирующему каждое слово в векторе, но вместо обучения на входных словах, Word2vec обучает слова на других словах, которые находятся рядом с ними в обучающем корпусе. Для этого могут использоваться следующие методы [7]:

- с помощью контекстного слова для предсказания целевого контекста (skip-gram);
- с использованием контекста (мешок слов, CBOW);
- с помощью иерархических чисел.

Наиболее эффективным является использование десятичных иерархических чисел с предварительно построенной онтологической таксономией предметной области.

В том случае, если вектор признаков какого-либо концепта не может быть использован для точного предсказания контекста этого концепта, компоненты вектора корректируются. Контекст каждого слова в корпусе позволяет организовать процесс обучения с корректировкой вектора признаков в семантическом пространстве. Векторы концептов, которые вычислены как схожие по их контексту, сближаются путем корректировки чисел в векторе.

Эти числа называются эмбедингами и определяют каждое слово как точку в 300-мерном семантическом пространстве с иерархическими числами. Хорошо обученный набор векторных представлений слов разместит похожие слова близко друг к другу в этом пространстве. Слова «дуб», «вяз» и «береза» могут сгруппироваться в одной локальной области пространства, в то время как слова «война», «конфликт» и «раздор» будут находиться рядом в другом.

Для применения иерархических семантических пространств эмбединги, формируемые инструментарием Word2vec, дополняются числами, полученными из родовидовой, причинно-следственной и меронимической таксономиями. Схожие иерархические числа, например 12.0.0.3.-1.0.0 и 12.0.0.3.-1.1, имеющие общую часть 12.0.0.3.-1, свидетельствуют о семантической близости двух концептов, представленных этими числами. Чем больше

общая часть иерархических индексов, тем ближе концепты. В графическом представлении два схожих вектора иерархического пространства будут друг с другом на более близком расстоянии.

Качественные характеристики этих концептов становятся количественными, и алгоритмы могут выполнять свою работу. Сходство слов как концептов – это лишь основа многих ассоциаций, которые можно получить при обучении. Небольшое усложнение инструментария может реализовать вычисление семантического сходства любой словарной конструкции, например, словосочетания или предложения и даже сложного текста.

Другой метод, Doc2Vec основан на применении нейронной сети, который учится распределённому представлению документов. Здесь в форме вектора (эмбединга) представлен документ.

При обучении эмбединги документов отображаются в близкие точки семантического пространства, что дает возможность производить следующие вычисления:

- классифицировать документы, например, на достоверность содержащихся в них сведений;
- определять релевантность документов для их использования в качестве ответа на вопрос;
- производить кластеризацию документов для выделения заранее не обозначенных учителем классов.

В этом случае также может быть использовано иерархическое пространство, в котором иерархическими числами обозначаются целые документы. Здесь таксономия применяется для определения следующих возможных отношений:

- вложенность одного документа в другой;
- существование между документами отношения «более общий» или «более частный»;
- следование фактов, содержащихся в одном документе из фактов, содержащихся в другом документе.

Как и в случае с методом Word2vec, иерархические индексы документов могут дополнять эмбединги классического применения Doc2Vec.

Еще одной технологией вычисления семантического сходства является алгоритм обучения эмбедингов без учителя GloVe. Эмбединги в этом методе вычисляются на всем обучающем корпусе на основе статистического анализа совместной встречаемости слов.

При этом рассчитывается матрица совместной встречаемости с использованием критерия на основе вычисления логарифма вероятности совместного появления слов, которое сопоставляется со скалярным произведением эмбедингов слов, предназначенных для определения их семантической близости. В отличие от метода Word2Vec подход GloVe использует глобальную статистику по всем документам корпуса. Алгоритм считается применимым на обучающих данных любых объемов, в том числе не очень больших.

GloVe может также использовать иерархические числа, сопоставленные отдельным концептам и словосочетаниям, выявленным в текстах и унифицированных с помощью внешних онтологических таксономий.

WordNet [8] также является весьма популярным инструментарием вычисления семантической близости словарных конструкций. Чаще всего это программное средство представляют как тезаурус или базу данных, содержащих более 150 тысячи слов и применяемую для определения синонимов, антонимов и меронимов слов. Концепты этого тезауруса определены как синесеты (synsets), представляющие собой уникальные индексы, с помощью которых можно вычислить различные словарные отношения. В базе данных эти отношения хранятся в явном виде. Иерархические числа в этом случае можно использовать с помощью формирования отдельной базы данных. Эта база должна учитывать отношения, включенные в базовый WordNet и содержать дополнительно к индексу синсета еще и его

иерархический индекс. В этом случае появляется возможность вычислять родовидовые и причинно-следственные отношения на множестве пар концептов.

Аналогичным WordNet является тезаурус VerbNet, сосредоточенный на множестве глаголов. В этом случае предполагается, что глаголы являются конструкциями, образующими предложения как их главные члены. Семантическое сходство глаголов может быть вычислено по следующей формуле:

$$\text{SemanticSimilarity}(V1, V2) = \frac{\sum_{i=1}^n V1_i \times V2_i}{\sqrt{\sum_{i=1}^n V1_i^2} * \sqrt{\sum_{i=1}^n V2_i^2}},$$

где V1 и V2 – вектора семантического пространства признаков для двух слов сопоставляемых предложений, а n – количество текстов корпуса.

Иерархические числа в этом случае можно применить так же, как и в случае с WordNet.

Эксперименты с вычислением семантического сходства в предметной области политических статей [9] из отечественных и западных средств массовой информации были проведены на основе следующих электронных ресурсов: cnn.com, springer, nbcnew, thrguardian.com, nytimes.com, france24, RT.com, meduza.com, kremlin.com, russiabeyond.com, interfax.com, sputnikinternational.com.

Результаты испытаний показали существенное повышение эффективности применения технологии, использующей иерархические числа, в сравнении с классическим использованием рассмотренных методов в среднем на 5-10% при оценке точности.

Библиографический список

1. И.Ю. Каширин. Бинарные иерархические числа для вычисления семантической близости предложений естественного языка. // Вестник РГРТУ. 2023. № 86 / Vestnik of RSREU. 2023. С.110-121. DOI: 10.21667/1995-4565-2023-86-110-121.
2. Wael H Goma, Aly A Fahmy, "A survey of text similarity approaches", International Journal of Computer Applications, Foundation of Computer Science, vol 68, 2013.
3. Word2Vec, Doc2vec & GloVe: Neural Word Embeddings for Natural Language Processing [Electronic resource]. Update date: 20.12.2025, URL: <https://deeplearning4j.konduit.ai/deeplearning4j/reference/word2vec-glove-doc2vec> (date of application: 06.01.2026).
4. И.Ю.Каширин. Формирование LLM-ориентированных ресурсов знаний на основе генерации дополненной информации. Вестник РГРТУ. 2025. № 94 / Vestnik of RSREU. 2025. No 94. С.85-97. DOI: 10.21667/1995-4565-2025-94-85-97.
5. И.Ю.Каширин. Естественнo-языковые модели. Материалы. [Electronic resource]. Update date: 20.12.2025, URL: <https://i.kashirin.net/forstudents/modelsforNLP> (date of application: 16.02.2026).
6. Велихов П. Машинное обучение для понимания естественного языка // Открытые Системы. СУБД. – 2016. – № 1. – С. 18–21. – URL: <https://www.osp.ru/os/2016/01/13048649> (дата обращения: 5.04.2020).
7. И.Ю. Каширин. Эмбеdдинги иерархических чисел для обогащения трансформерных языковых моделей внешними онтологическими знаниями. Вестник РГРТУ. 2025. № 93 / Vestnik of RSREU. 2025. No 93.С.72-82. DOI: 10.21667/1995-4565-2025-93-72-82.
8. Word Net: An Electronic Lexical Database / Edit. by Christiane Fellbaum. – Cambridge, MIT Press, 1998 (дата обращения: 5.04.2020).
9. И.Ю. Каширин. Извлечение фактов из естественнo-языковых текстов методом унификации семантических паттернов. Вестник РГРТУ. 2025. № 91 / Vestnik of RSREU. 2025. No 91.С.36-49. DOI: 10.21667/1995-4565-2025-91-36-49.

УДК 004.932; ГРНТИ 28.23.15

АНАЛИЗ МИМИКИ ЛИЦА В СИСТЕМАХ ВИДЕОКОНФЕРЕНЦИЙ НА ОСНОВЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ

И.А. Ветринцев, Н.И. Цуканова

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
РФ, Рязань, vetrintsev.ilya@gmail.com

Аннотация. Проведено сравнительное исследование четырёх нейросетевых архитектур – ResNet-50, EfficientNet-B2, ViT-Base/16 и 3D-CNN – применительно к задаче распознавания базовых эмоций человека в условиях видеоконференций. Просчитана линейная модель задержки инференса для CNN-архитектур. Показано, что ViT отклоняется от CNN-модели на 23,2 мс вследствие квадратичной сложности механизма внимания.

Ключевые слова: распознавание выражений лица, нейронные сети, EfficientNet, Vision Transformer, 3D-CNN, видеоконференции.

FACIAL EXPRESSION ANALYSIS IN VIDEO CONFERENCING SYSTEMS USING DEEP LEARNING

I.A. Vetrintsev, N.I. Tsukanova

Ryazan State Radio Engineering University named after V.F. Utkin,
RF, Ryazan, vetrintsev.ilya@gmail.com

The summary. A comparative study of four neural network architectures – ResNet-50, EfficientNet-B2, ViT-Base/16, and 3D-CNN – is conducted for the task of Ekman's basic emotion recognition in video conferencing conditions. A linear inference latency model for CNN architectures is developed. It is shown that ViT deviates from the CNN model by 23.2 ms due to the quadratic complexity of the attention mechanism.

Keywords: facial expression recognition, deep learning, EfficientNet, Vision Transformer, 3D-CNN, video conferencing.

Видеоконференцсвязь стала одним из основных инструментов деловой коммуникации. Именно поэтому автоматический анализ мимики участников позволил бы оценивать вовлеченность, обнаруживать усталость и формировать обратную связь в режиме реального времени [1]. Однако нестабильность видеопотока, связанная с качеством интернет-связи – сжатие, нестабильное освещение, искажения лица web-камерой – существенно отличаются от лабораторных тестирований набора данных FER. Дополнительное ограничение – обработка потока в реальном времени без выделенной GPU, так как в реальных условиях работы приложения для ВКС не должны обязательно требовать GPU-ускорители [2].

Цель работы – сравнить четыре архитектуры и выяснить, какая из них окажется наиболее подходящей для разработки AI-ассистента видеоконференций:

1. ResNet-50 – базовый CNN; [2]
2. EfficientNet-B2 – компактный CNN составного масштабирования [3];
3. ViT-Base/16 – трансформер [4];
4. 3D-CNN R3D-18 – сеть для видеопоследовательностей [5].

Главные критерии оценки: точность на тестовой выборке, задержка предсказания на CPU без аппаратного ускорителя.

Используемые архитектуры

1. **ResNet-50.** Вводит остаточные соединения (skip connections), передающие сигнал напрямую поверх стека слоёв. Архитектура состоит из четырёх групп со степенью дискретизации 2 (всего пять переходов от 224×224 до 7×7). Число каналов удваивается на каждой стадии: 64 - 128 - 256 - 512 - 2048.

2. **EfficientNet-B2.** Использует составное масштабирование – единый коэффициент ϕ одновременно регулирует глубину, ширину и входное разрешение. Основной блок – MBConv с SE-вниманием по каналам. Входное разрешение – 260×260 ; число параметров в 9,4 раза меньше, чем у ResNet-50.

3. **ViT**. Входное изображение 224×224 разбивается на 196 патчей 16×16 . Каждый патч проецируется в 768-мерный токен. 12 блоков multi-head self-attention (12 голов, $d = 64$) позволяют улавливать глобальные зависимости между областями лица, что объясняет наивысшую точность. Вычислительная сложность механизма внимания квадратична по числу патчей N , что является основным источником задержки, как показано ниже.

4. **3D-CNN**. Заменяет все свёртки 3×3 на 3D-ядра $3 \times 3 \times 3$ – временное пространство каналов анализируется совместно. Вход: клип из $T = 16$ кадров 224×224 . Размер входного тензора $1 \times 16 \times 3 \times 224 \times 224$ увеличивает вычислительную сложность до 40,5 GFLOPs на клип, что в 10 раз превышает EfficientNet-B2.

Методология эксперимента

Набор данных объединяет 28000 изображений AffectNet (4000 изображений на все 7 эмоций) и 100 видеоклипов DFEW [6]. Разбиение 70/15/15 по субъектам. Аугментация имитирует условия видеоконференций:

1. сжатие изображения;
2. вариация яркости;
3. гауссово размытие;
4. добавление шума;
5. поворот $\pm 15^\circ$, горизонтальное отражение, случайное кадрирование области лица.

Обучение: оптимизатор – AdamW, начальная скорость $\eta_0 = 10^{-4}$, косинусное расписание, 50 эпох. Оборудование: GPU NVIDIA RTX 4070s 12 GB, CPU Intel Core i7-12700. FPS измерялся при размере батча равным одному, т.е. одно изображение за итерацию.

Экспериментальные результаты

Таблица 1. Сравнение архитектур

Модель	Acc, %	F1, %	Param, M	GFLOPs	FPS GPU	FPS CPU	Задержка CPU, мс
ResNet-50	86,4	85,1	25,6	4,1	74	12	83
EfficientNet-B2	89,1	87,8	9,2	1,0	118	21	47
ViT-Base/16	91,3	90,1	86,4	17,6	61	6	162
3D-CNN (R3D-18)	88,7	86,9	33,4	40,5	38	4	251

Результаты по классам каждой эмоции записаны в таблице 2. Наибольшее число ошибок во всех моделях приходится на тройку страх–удивление–отвращение: около 20% образцов класса «страх» ошибочно относятся к «удивлению» или «отвращению» из-за общего сходства выражений лиц. Класс «нейтральный» – доминирует среди всех эмоций в видеоконференциях и как следствие распознается с точностью более 90% у всех архитектур. Как видно на рисунке 1, все модели демонстрируют стабильную сходимость, при этом EfficientNet-B2 и ViT-Base/16 достигают более высокого уровня обобщения и имеют маленькие потери уже после 30 эпох обучения.

Таблица 2. F1-score, %, по классам эмоций

Эмоция	ResNet-50	EffNet-B2	ViT-B/16	3D-CNN
Радость	93,2	95,1	96,3	94,4
Грусть	84,1	87,2	89,1	86,2
Гнев	82,3	85,6	87,9	84,8
Страх	79,8	83,1	86,2	81,9
Удивление	81,2	84,9	87,7	83,7
Отвращение	77,9	80,8	83,6	79,3
Нейтральный	90,3	92,1	94,2	92,8

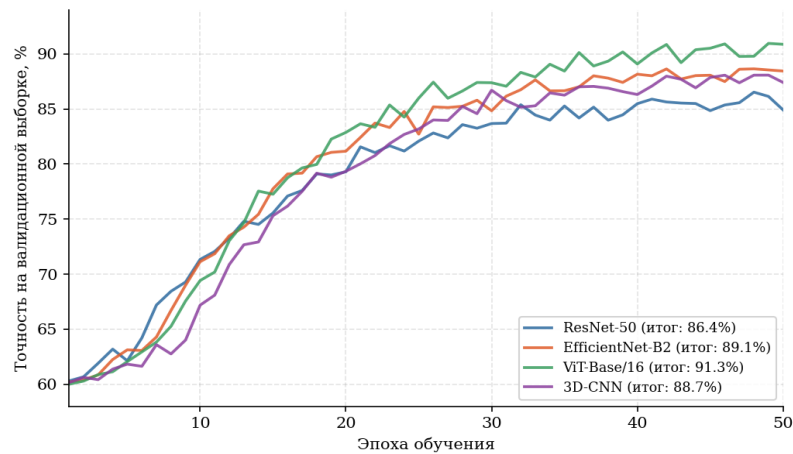


Рис. 1. Кривые обучения архитектур.

Математический анализ

Для CNN слоя C_{in} входными каналами, C_{out} выходными, ядром $K \times K$ и пространственным размером $H \times W$ вычислительная сложность операций умножения-накопления:

$$FLOPs_{conv} = 2 * C_{in} * C_{out} * K^2 * H * W \quad (1)$$

Для механизма self-attention ViT с N токенами, h головами и мерностью токена d_k :

$$FLOPs_{attn} = 4 * D_k * N^2 * h \quad (2)$$

При $N = 196$, $d_k = 64$, $h = 12$: $FLOPs_{attn} \approx 13,5$ GFLOPs – все сверточные операции ViT дают лишь 3,80 GFLOPs. Разбивка GFLOPs по типам операций приведена в таблице 3. На рисунке 2 наглядно видно соотношение вычислительной сложности и задержки инференса всех четырех архитектур. По диаграммам видно, что EfficientNet-B2 существенно выигрывает по эффективности, обеспечивая наименьшее время отклика на CPU, в то время как 3D-CNN, несмотря на сопоставимую с ResNet-50 точность, значительно проигрывает из-за высокой вычислительной сложности трехмерных сверток.

Таблица 3. Разбивка GFLOPs по типам операций

Модель	Сверточные слои, GFLOPs	Бнормал. + activation	Linear / Attention	Итого
ResNet-50	3,82	0,21	0,07	4,1
EfficientNet-B2	0,74	0,14	0,12	1,0
ViT-Base/16	3,80	0,28	13,52	17,6
3D-CNN (R3D-18)	39,1	1,05	0,35	40,5

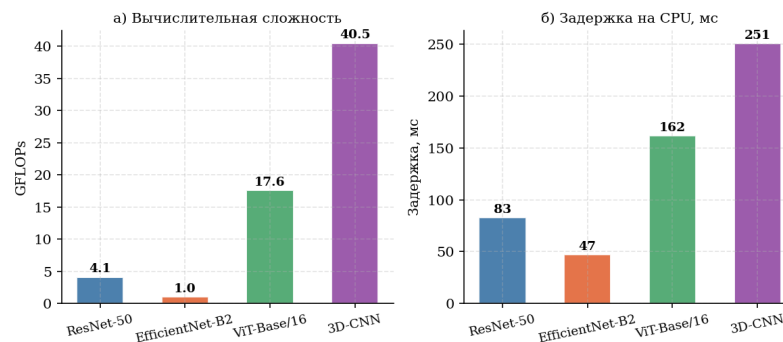


Рис. 2. Вычислительная сложность и задержка инференса архитектур

Линейная модель задержки для CNN-архитектур

По трем нейронным сетям (ResNet-50, EfficientNet-B2, 3D-CNN) методом наименьших квадратов подобрана линейная зависимость задержки от вычислительной сложности:

$$T = (4,93 * G) + 51,99, R^2 = 0,991, \quad (3)$$

где T – задержка инференса в мс, G – число GFLOPs. Коэффициент G отражает время обработки одной GFLOPs на процессоре ($\approx 4,93$ мс/GFLOPs), второе слагаемое – расходы на инициализацию, пред и постобработку ($\sim 51,99$ мс). По формуле (3) для ViT ожидаемая задержка $T_{\text{пред}} = 4,93 \cdot 17,6 + 51,99 = 138,8$ мс, фактическая – 162 мс. Отклонение (23,2 мс) – расходы механизма внимания, вычисляющего QKV-матрицы без аппаратной оптимизации. Данная модель показана на рисунке 3. Точки, соответствующие ResNet-50, EfficientNet-B2 и 3D-CNN, находятся вблизи прямой, в то время как ViT-Base/16 демонстрирует заметное отклонение вверх примерно на 23 мс.

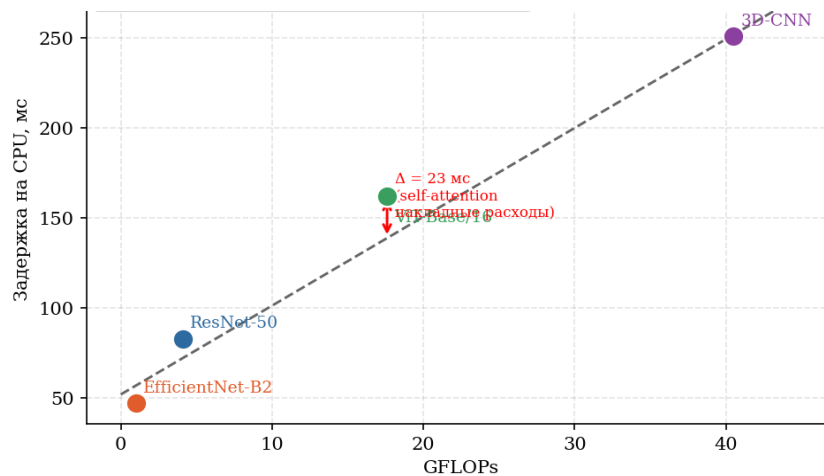


Рис. 3. Линейная модель задержки и отклонение ViT

Заключение

По результатам эксперимента EfficientNet-B2 обеспечивает наилучшее соотношение точности и задержки (89,1%, 47 мс) и является предпочтительной архитектурой для развёртывания без GPU. ViT-Base/16 даёт наивысшую точность (91,3%), но его задержка 162 мс на CPU превышает предсказанную по модели (3) на 23,2 мс вследствие накладных расходов механизма внимания. 3D-CNN, несмотря на учёт временной динамики, проигрывает по задержке (251 мс). Именно поэтому EfficientNet-B2 становится приоритетным направлением для дальнейших исследований.

Библиографический список

1. Гусев А.Н., Баев М.С., Кремлев А.Е. Метод прямой оценки лицевых экспрессий на видеозаписи: от восприятия эксперта к компьютерному зрению. Психология труда, организации и управления в условиях цифровой трансформации общества: коллективная монография. Петрозаводск: Издательство Петрозаводского государственного университета, 2021. — С. 268–278.
2. Корчагин В.Д., Кувшинников В.С., Ковшов Е.Е. Критериальный анализ моделей обработки данных радиационного неразрушающего контроля. Информационные технологии в проектировании и производстве. – 2024. – № 3. – С. 45–53.
3. Щетинин Е.Ю. О перспективах применения компьютерных лингвистических моделей в задачах классификации биомедицинских изображений. Математическое моделирование. – 2023. – Т. 35, № 12. – С. 18–30.
4. Щетинин Е.Ю. Эффективность применения визуальных трансформеров для обнаружения аномалий на гистологических изображениях. Информатика и автоматизация. – 2025. – № 1. – С. 112–128.
5. Баев М.С., Гусев А.Н. Перспективы автоматизированного анализа мимических реакций в комплексном полиграфическом исследовании. Теоретические и прикладные аспекты использования специальных знаний в уголовном и гражданском судопроизводстве: Сб. статей. Вып. 4. – М.: Российский государственный университет правосудия, 2020. – С. 25–43.

УДК 004.8:519.71; ГРНТИ 28.23.15

ИССЛЕДОВАНИЕ МЕТОДОВ ПОСТРОЕНИЯ МУЗЫКАЛЬНЫХ СЕТОВ НА ОСНОВЕ ЭМОЦИОНАЛЬНОЙ ДИНАМИКИ

Д.М. Лушков, И.Ю. Филатов*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, azlfox0@gmail.com*

Аннотация. В последние годы наблюдается стремительное развитие интеллектуальных систем, применяемых в творческих областях, включая генерацию изображений, текстов и музыки. Одним из перспективных направлений является автоматизация процессов, традиционно выполняемых человеком в музыкальной индустрии, в частности — создание музыкальных сетов, используемых в диджеинге, радиоэфирах, стриминговых сервисах и интерактивных медиасредах. Данная статья посвящена исследованию методов построения музыкальных сетов на основе эмоциональной динамики и разработке формализованных моделей их формирования. В работе рассматриваются теоретические основы представления эмоционального содержания музыки, включая дискретные и континуальные модели эмоций, а также методы извлечения акустических и музыкальных признаков, используемых для количественного описания эмоциональных состояний. Проведен анализ существующих подходов к построению музыкальных сетов, включая ручные и эвристические методы, алгоритмические решения на основе графовых и правило-ориентированных моделей, а также современные методы машинного обучения, включая модели последовательностей, трансформеры и обучение с подкреплением. Предложена математическая формализация задачи построения музыкального сета как задачи оптимизации в эмоциональном пространстве, включающая определение метрик эмоциональной плавности переходов, модели целевой эмоциональной арки и интегральной функции качества сета.

RESEARCH ON METHODS FOR CONSTRUCTING MUSICAL SETS BASED ON EMOTIONAL DYNAMICS

D.M. Lushkov, I.Y. Filatov*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, azlfox0@gmail.com*

The summary. In recent years, there has been rapid development of intelligent systems applied in creative domains, including the generation of images, texts, and music. One of the promising directions is the automation of processes traditionally performed by humans in the music industry, in particular the creation of musical sets used in DJ performances, radio broadcasts, streaming services, and interactive media environments. This article is devoted to the study of methods for constructing musical sets based on emotional dynamics and the development of formalized models for their generation. The paper examines the theoretical foundations of representing the emotional content of music, including discrete and continuous models of emotions, as well as methods for extracting acoustic and musical features used for the quantitative description of emotional states. An analysis of existing approaches to musical set construction is conducted, including manual and heuristic methods, algorithmic solutions based on graph-based and rule-based models, and modern machine learning methods, including sequence models, transformers, and reinforcement learning. A mathematical formalization of the musical set construction problem as an optimization task in emotional space is proposed, including the definition of emotional transition smoothness metrics, a target emotional arc model, and an integral quality function of the musical set.

Музыкальный сет представляет собой не просто набор композиций, а логически и эмоционально выстроенную последовательность, целью которой является формирование у слушателя определённого эмоционального состояния и его постепенное изменение во времени. Профессиональные диджеи при формировании сетов интуитивно или осознанно используют принципы эмоциональной динамики: нарастание и спад энергии, чередование напряжения и расслабления, создание кульминационных моментов и завершение композиционного повествования. Однако большинство существующих автоматических систем рекомендаций музыки ориентированы на подбор отдельных треков, а не на построение целостной эмоциональной траектории.

Современные стриминговые платформы используют алгоритмы персонализированной рекомендации, однако они, как правило, не учитывают эмоциональную структуру последовательности композиций, что приводит к фрагментированному восприятию сета.

Актуальность данного исследования обусловлена необходимостью разработки и анализа методов, позволяющих формировать музыкальные сетсы как целостные наборы, а не как набор независимых рекомендаций.

Теоретические основы эмоциональной динамики в музыкальных сетсах

Музыкальный сет как последовательность эмоциональных состояний

Музыкальный сет в контексте диджеинга и цифровых музыкальных платформ представляет собой упорядоченную последовательность музыкальных композиций, объединённых общей логикой развития, стилевым единством и эмоциональной направленностью. В отличие от обычного плейлиста, музыкальный сет предполагает наличие сценария, включающего вступление, развитие, кульминацию и завершение.

С точки зрения восприятия слушателем, музыкальный сет формирует непрерывную эмоциональную траекторию, в рамках которой изменяются уровни возбуждения, напряжения и эмоциональной окраски. Профессиональные диджеи используют различные приёмы управления этим процессом: постепенное увеличение темпа, смену тональностей, изменение плотности ритма и энергетики.

Таким образом, музыкальный сет можно рассматривать как временную последовательность эмоциональных состояний, где каждая композиция влияет не только на текущее состояние слушателя, но и на восприятие последующих треков. Это свойство делает задачу автоматического построения сетсов принципиально более сложной, чем задачу рекомендации отдельных композиций [4].

Модели представления эмоций в музыке

Для формализации эмоциональной динамики необходимо определить способы представления эмоционального состояния музыкальных композиций. В современной музыкальной информатике используется несколько основных подходов.

Дискретные модели эмоций

Данный подход предполагает классификацию эмоций по ограниченному набору категорий: радость, грусть, напряжение, спокойствие, агрессия и др. Несмотря на простоту интерпретации, дискретные модели плохо подходят для описания плавных переходов и постепенных изменений, характерных для музыкальных сетсов.

Континуальные модели эмоций

Наиболее распространённой является модель Valence–Arousal, где:

- Valence отражает эмоциональную окраску (положительная/отрицательная),
- Arousal – уровень возбуждения или энергии.

Данная модель позволяет описывать эмоциональное состояние как точку в двумерном пространстве и отслеживать динамику его изменения во времени, что делает её особенно подходящей для анализа музыкальных сетсов.

Акустические и музыкальные признаки

Для автоматического определения эмоций используются числовые признаки, извлекаемые из аудиосигнала:

- темп (BPM),
- громкость,
- спектральная яркость,
- ритмическая плотность,
- гармоническая стабильность.

Популярные платформы, такие как Spotify, предоставляют агрегированные показатели (energy, valence, danceability), которые широко применяются в исследованиях автоматического анализа музыки.

Эмоциональная динамика как критерий качества музыкального сета

Качество музыкального сета определяется не только качеством отдельных композиций, но и характером переходов между ними. Резкие скачки энергии или настроения могут нарушать целостность восприятия, тогда как плавная эмоциональная динамика способствует эффекту «потока» (flow) [3].

Основными характеристиками эмоциональной динамики являются:

- плавность переходов;
- наличие кульминации;
- логичность эмоционального сценария;
- отсутствие эмоциональной усталости.

В контексте интеллектуальных систем это означает, что алгоритм должен учитывать не только локальную совместимость соседних треков, но и глобальную эмоциональную структуру всего сета. Таким образом, эмоциональная динамика становится ключевым критерием качества, который должен быть формализован и использован при автоматическом построении музыкальных сетов.

Анализ существующих методов построения музыкальных сетов на основе эмоциональной динамики

Ручные и эвристические методы формирования музыкальных сетов

Исторически построение музыкальных сетов являлось исключительно творческой задачей, решаемой диджеем на основе опыта, интуиции и понимания аудитории. Наиболее распространённым подходом является формирование сета в виде эмоциональной арки, включающей вступление, постепенное нарастание энергии, кульминацию и завершение.

Классическими эвристическими правилами являются:

- постепенное увеличение темпа (energy ramp);
- ограничение резких изменений настроения;
- чередование напряжённых и расслабленных фрагментов;
- поддержание жанровой согласованности.

Несмотря на эффективность данных методов, они обладают рядом недостатков: субъективность, отсутствие формализации и невозможность масштабирования. Кроме того, такие подходы плохо подходят для автоматизации, так как основаны на неявных знаниях и личном опыте диджея.

Алгоритмические методы построения музыкальных сетов

С развитием вычислительных методов появились попытки формализовать процесс построения музыкальных сетов. Одним из первых подходов стало представление треков в виде узлов графа, где рёбра отражают степень совместимости или эмоциональной близости.

Графовые методы

В графовых моделях задача построения сета сводится к поиску пути, который максимизирует заданную функцию качества, например, плавность эмоциональной динамики. Для этого применяются алгоритмы поиска кратчайшего пути, такие как A*, Dijkstra или динамическое программирование.

Преимуществом графовых методов является прозрачность и управляемость, однако их сложность возрастает экспоненциально при увеличении числа треков, что ограничивает практическое применение для больших музыкальных библиотек.

Правило–ориентированные методы

Rule-based подходы используют заранее заданные правила перехода, например:

- если energy увеличивается, следующий трек должен быть с большей или равной энергией;
- смена valence допускается только в пределах заданного порога.

Такие методы легко реализуются, однако плохо адаптируются к различным музыкальным стилям и не учитывают глобальную структуру сета.

Методы машинного обучения для построения музыкальных сетов

Современные исследования активно используют методы машинного обучения для моделирования последовательностей музыкальных композиций. В отличие от эвристик, данные методы позволяют обучаться на реальных сетах диджеев или плейлистах пользователей [1].

Модели последовательностей

Рекуррентные нейронные сети (LSTM, GRU) способны моделировать временные зависимости между треками, учитывая контекст предыдущих композиций. Однако такие модели склонны к накоплению ошибок и требуют больших объёмов данных для обучения.

Трансформеры

Transformer-модели позволяют учитывать глобальный контекст последовательности и показали высокую эффективность в задачах генерации последовательностей. В контексте музыкальных сетов они могут использоваться для предсказания следующего трека с учётом эмоциональной динамики всего сета.

Обучение с подкреплением

В подходах reinforcement learning процесс построения сета рассматривается как задача оптимизации награды, зависящей от эмоциональной гладкости, разнообразия и удержания внимания. Несмотря на перспективность, такие методы требуют сложной настройки функции награды и значительных вычислительных ресурсов [2].

Сравнительный анализ методов построения музыкальных сетов

В таблице 1 приведено сравнение рассмотренных методов.

Таблица 5. Сравнительная таблица методов построения музыкальных сетов

Метод	Учитывает глобальную динамику	Интерпретируемость	Масштабируемость
Эвристики	Частично	Высокая	Высокая
Графовые	Да	Средняя	Низкая
Rule-based	Нет	Высокая	Средняя
LSTM / GRU	Частично	Низкая	Средняя
Transformer	Да	Низкая	Высокая
RL	Да	Низкая	Низкая

Анализ показывает, что наиболее перспективными являются гибридные подходы, сочетающие интерпретируемость эвристик и гибкость моделей машинного обучения.

Математические и алгоритмические методы построения музыкальных сетов Формализация музыкального сета как задачи оптимизации

Музыкальный сет представляется как упорядоченная последовательность композиций:

$$S = \{t_1, t_2, \dots, t_n\},$$

где t_i – i -я композиция сета.

Каждой композиции сопоставляется вектор эмоциональных характеристик:

$$e_i = \{v_i, a_i, en_i, d_i\},$$

где v_i – valence, a_i – arousal, en_i – energy, d_i – danceability.

Тогда музыкальный сет можно представить как траекторию в пространстве эмоциональных состояний:

$$T(S) = \{e_1, e_2, \dots, e_i\}.$$

Задача автоматического построения музыкального сета сводится к оптимизации функции качества:

$$S^* = \arg \max_s Q(S),$$

где $Q(S)$ – интегральная оценка эмоциональной динамики сета.

Метрика эмоциональной плавности переходов

Плавность переходов между треками является ключевым фактором качества музыкального сета. Она определяется как средняя дистанция между соседними эмоциональными векторами:

$$D_{\text{smooth}}(S) = \frac{1}{n-1} \sum_{i=1}^{n-1} |e_{i+1} - e_i|.$$

Минимизация данной величины обеспечивает отсутствие резких эмоциональных скачков.

Однако минимизация только этой метрики приводит к монотонности, поэтому вводятся дополнительные критерии глобальной динамики.

Модель целевой эмоциональной арки

Для задания желаемого сценария развития эмоций используется функция эмоциональной арки:

$$\hat{a}(i) = f\left(\frac{i}{n}\right),$$

где f – функция формы сета (линейная, сигмоидальная, синусоидальная).

Отклонение реального сета от целевой арки определяется как:

$$D_{\text{arc}}(S) = \frac{1}{n} \sum_{i=1}^n |a_i - \hat{a}(i)|.$$

Минимизация D_{arc} позволяет формировать логичную эмоциональную структуру сета.

Графовая модель построения музыкального сета

Пусть $G=(V,E)$ – ориентированный граф, где:

- V – множество музыкальных треков,
- E – допустимые переходы между треками.

Вес ребра между треками t_i и t_j определяется как:

$$\omega_{ij} = \alpha * \|e_j - e_i\| + \beta * |a_j - \hat{a}(k)|,$$

где k – позиция трека в сете, α, β – весовые коэффициенты.

Задача построения сета сводится к поиску пути минимальной стоимости:

$$S^* = \arg \min_s \sum_{(i,j) \in E} \omega_{ij}$$

Интегральная функция качества музыкального сета

Для объединения всех критериев вводится итоговая функция:

$$Q(S) = \lambda_1 (1 - D_{\text{smooth}}) + \lambda_2 (1 - D_{\text{arc}}) + \lambda_3 D_{\text{var}},$$

где D_{var} – дисперсия эмоциональных параметров, λ_i – коэффициенты значимости.

Данная функция позволяет адаптировать поведение алгоритма под различные типы сетов (разогрев, основная часть, завершение).

Библиографический список

1. Кипяткова И. С., Карпов А. А. Разновидности глубоких искусственных нейронных сетей для систем распознавания речи // Труды СПИИРАН, 2016, Вып. 49, С. 80–103.
2. Соловьёв В. А., Морозов А. Н. Применение методов машинного обучения для классификации музыкальных фрагментов // Научно-технический вестник информационных технологий, механики и оптики. – 2020. – № 4. – С. 591–599.
3. Торопова А.В., Князева Т.С. // Объективные и субъективные факторы распознавания эмоционального содержания музыки // Музыкальное искусство и образование. 2013
4. Князева Т.С., Торопова А.В. Распознавание эмоционального содержания музыки в зависимости от особенностей музыкального материала и опыта слушателей.

УДК 004.8; ГРНТИ 28.23.29

УЧЕТ ВРЕМЕННОЙ РЕЛЕВАНТНОСТИ В АЛГОРИТМЕ ПЕРСОНАЛИЗИРОВАННОЙ РЕКОМЕНДАЦИИ МЕРОПРИЯТИЙ

В.А. Лутиков, Е.Н. Проказникова

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, vadim.lutikov@mail.ru*

Аннотация. В работе рассматривается проблема учета временного фактора в системах персонализированной рекомендации городских мероприятий. Предлагается модификация алгоритма ранжирования на основе введения коэффициента временной релевантности, зависящего от разности между текущим моментом времени и датой события. Рассматривается математическая формализация подхода, его интеграция в существующую архитектуру рекомендательной системы и концептуальные преимущества. Показано, что учет временной динамики позволяет повысить практическую значимость рекомендаций без увеличения вычислительной сложности модели.

Ключевые слова: рекомендательная система, временная релевантность, машинное обучение, ранжирование, персонализация, событийные данные.

TEMPORAL RELEVANCE IN A PERSONALIZED EVENT RECOMMENDATION ALGORITHM

V.A. Lutikov, E.N. Prokaznikova

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, vadim.lutikov@mail.ru*

The summary. The paper considers the problem of incorporating temporal factors into personalized event recommendation systems. A ranking modification based on a temporal relevance coefficient is proposed. The mathematical formulation and integration of the approach into an existing recommendation pipeline are discussed. The proposed method improves the practical utility of recommendations without increasing computational complexity.

Keywords: recommender system, temporal relevance, machine learning, ranking, personalization, event data.

В контексте цифровизации городской среды быстро развиваются персонализированные рекомендательные сервисы. Эти системы позволяют пользователям выбирать образовательные, культурные и развлекательные мероприятия на основе их интересов и поведенческих характеристик. В большинстве случаев задача формулируется как ранжирование объектов по вероятности того, что они представляют интерес для пользователя.

Пусть каждому мероприятию соответствует вектор признаков

$$x = (x_1, x_2, \dots, x_n),$$

включающий в себя категорию события, стоимость, расстояние до пользователя, популярность, историю взаимодействий и другие параметры. На основе этих характеристик обучается модель машинного обучения, которая генерирует оценку вероятности заинтересованности пользователя в событии:

$$P = f(x).$$

Полученная вероятность используется для сортировки списка мероприятий по убыванию P . Однако подобный подход имеет существенное ограничение: он не учитывает временную динамику событий. В отличие от статичных объектов (товаров или фильмов), городские мероприятия имеют строго определенную дату проведения. Событие, обладающее высокой вероятностью интереса, но запланированное через несколько месяцев, может быть менее актуальным по сравнению с мероприятием, проходящим в ближайшие дни.

Таким образом, для событийных рекомендательных систем возникает необходимость явного учета временного фактора. Предлагается ввести коэффициент временной релевантности T , отражающий степень актуальности события относительно текущего момента времени.

Обозначим через Δt разность между датой проведения мероприятия и текущим временем. Тогда коэффициент временной релевантности можно определить в виде экспоненциальной функции затухания:

$$T = \exp(-\lambda \Delta t),$$

где $\lambda > 0$ – параметр, определяющий скорость снижения релевантности во времени.

Экспоненциальная форма выбрана по следующим причинам. Во-первых, она обеспечивает плавное и непрерывное уменьшение веса события по мере удаления во времени. Во-вторых, при малых значениях Δt коэффициент T близок к единице, что не искажает оценку вероятности. В-третьих, параметр λ позволяет адаптировать модель к различным сценариям использования сервиса.

Итоговая оценка ранжирования определяется как произведение вероятности интереса и временного коэффициента:

$$S = P * T.$$

Таким образом, мероприятия, близкие по времени, сохраняют высокий итоговый балл, а отдаленные события автоматически понижаются в выдаче. Важно отметить, что данный подход не требует изменения структуры обученной модели $f(x)$. Временная коррективировка применяется на этапе постобработки результатов.

Интеграция метода в архитектуру рекомендательной системы включает следующие этапы:

- 1) формирование вектора признаков мероприятия;
- 2) вычисление вероятности интереса P с помощью модели машинного обучения;
- 3) расчет временного коэффициента T ;
- 4) вычисление итогового балла S и сортировку списка мероприятий.

Предложенный подход обладает рядом преимуществ. Во-первых, он минимально увеличивает вычислительную нагрузку, поскольку расчет экспоненциальной функции не является ресурсоемкой операцией. Во-вторых, метод универсален и может применяться к различным моделям: логистической регрессии, ансамблям деревьев решений, нейронным сетям. В-третьих, он сохраняет интерпретируемость алгоритма, поскольку влияние временного фактора явно контролируется параметром λ .

Следует отметить, что выбор параметра λ может осуществляться эмпирически на основе пользовательских данных. Например, для пользователей, предпочитающих планировать события заранее, целесообразно выбирать меньшее значение λ , обеспечивающее более медленное затухание. Для пользователей, ориентированных на спонтанные решения, значение λ может быть увеличено.

Таким образом, учет временной релевантности является важным элементом построения рекомендательных систем для событийных данных. Введение простого коэффициента временного затухания позволяет повысить практическую ценность рекомендаций, не усложняя алгоритмическую структуру модели. Предложенный подход может служить

основой для дальнейших исследований, связанных с динамическими предпочтениями пользователей и адаптивной настройкой параметров системы.

Библиографический список

1. Кутянин А.Р. Рекомендательные системы: обзор основных постановок и результатов // Информационные системы: методы и средства. – 2017.
2. Иванова А.В., Петров Б.С., Сидоров К.Л. Роль контекстной информации в рекомендательных системах: анализ подходов и методов // Российский документальный центр исследований. – 2022.

УДК 004.032.26; ГРНТИ 28.23.37

ВОЗМОЖНОСТИ И ОГРАНИЧЕНИЯ ИСПОЛЬЗОВАНИЯ НЕЙРОННЫХ СЕТЕЙ КАК ИНСТРУМЕНТА СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

М.А. Федулов, В.В. Белов

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, fmikhail1199@gmail.com*

Аннотация. В статье рассматриваются возможности и ограничения нейронных сетей (НС) в качестве инструментов систем поддержки принятия решений (СППР). Сначала анализируются основные типы НС (CNN, RNN, трансформеры) и их применимость к задачам прогнозирования, кластеризации и оптимизации, которые часто встречаются в СППР. Далее обсуждаются ограничения, связанные с требованием больших количеств размеченных данных, интерпретируемостью моделей, а также с вычислительными ресурсами и временем обучения. В качестве практического фокуса приводятся примеры успешной интеграции НС в бизнес-аналитические и медицинские системы, а также рекомендации по преодолению выявленных барьеров. Итоги подчеркивают, что НС способны значительно повысить эффективность принятия решений, но требуют грамотного выбора архитектуры, качественной подготовки данных и соблюдения принципов прозрачности.

Ключевые слова: нейронные сети, принятие решений, системы поддержки принятия решений, СППР, искусственный интеллект, машинное обучение.

OPPORTUNITIES AND LIMITATIONS OF USING NEURAL NETWORKS AS TOOLS FOR DECISION SUPPORT SYSTEMS

M.A. Fedulov, V.V. Belov

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, fmikhail1199@gmail.com*

The summary. The article examines the possibilities and constraints of neural networks (NNs) as tools for decision support systems (DSS). First, the main types of NNs (CNNs, RNNs, transformers) are analyzed and their applicability to forecasting, clustering, and optimization tasks—commonly encountered in DSS—is discussed. Next, the limitations related to the need for large amounts of labeled data, model interpretability, and computational resources and training time are explored. Practical examples of successful integration of NNs into business analytics and medical systems are presented, along with recommendations for overcoming the identified barriers. The conclusion emphasizes that NNs can significantly enhance decision-making efficiency, but require careful architecture selection, high-quality data preparation, and adherence to transparency principles.

Keywords: neural networks, decision making, decision support systems, DSS, artificial intelligence, machine learning.

Введение

Современные организационные системы сталкиваются с растущей сложностью планирования проектов, что обусловлено неопределённостью сроков, динамичностью ресурсов и нелинейной взаимосвязью между задачами. При этом, классические методы опираются на линейные модели и экспертные оценки, что ограничивает их эффективность в работе с не-

структурированными данными, скрытыми зависимостями и динамическими изменениями в ходе реализации проекта [4, 5]. В этих условиях всё большую роль начинают играть подходы, основанные на искусственном интеллекте (ИИ) и, в частности, на машинном обучении (МО) – направлениях, позволяющих системам автоматически извлекать знания из данных и адаптироваться к изменяющимся условиям.

Нейронные сети (НС), как один из ключевых инструментов машинного обучения, демонстрируют высокую эффективность в выявлении сложных, нелинейных паттернов в исторических данных проектов без явного задания правил. Их применение в системах поддержки принятия решений (СППР) становится перспективным направлением для повышения точности прогнозирования сроков, объёма ресурсов и рисков. Однако использование нейронных сетей сопряжено с рядом ограничений – в частности, низкой интерпретируемостью, высокими требованиями к объёму и качеству обучающих данных, а также риском переобучения при недостаточной валидации. В данной статье будут проведён анализ возможностей и ограничений нейронных сетей как инструмента систем поддержки принятия решений в сравнении с классическими методами.

Классические методы поддержки принятия решений в вопросах планирования проектов

В современном управлении проектами принятие обоснованных решений требует системного подхода, сочетающего количественный анализ, учёт неопределённости и оценку множества конфликтующих критериев. Для этого применяется совокупность методов, каждый из которых направлен на решение конкретных задач планирования, оценки рисков и выбора оптимальных альтернатив. Ниже представлены ключевые методы, используемые в практике управления проектами:

- Метод критического пути (CPM – Critical Path Method);
- Метод PERT (Program Evaluation and Review Technique);
- Метод Монте-Карло;
- Многокритериальные методы принятия решений (MCDM);
- Метод анализа иерархий (АНР);
- Анализ дерева решений (Decision Tree Analysis).

Каждый из этих методов обладает уникальной областью применения и дополняет другие, формируя комплексный инструментарий для поддержки управленческих решений. В современной практике всё чаще применяется комбинированный подход: например, АНР используется для определения весов критериев, а Монте-Карло – для моделирования рисков по сценариям. Информационные технологии – включая специализированные программные платформы (Primavera, MS Project), аналитические пакеты (Crystal Ball, @RISK), а также языки программирования (Python, R) – играют ключевую роль в реализации этих методов, обеспечивая точность, скорость, воспроизводимость и визуализацию результатов [1, 2, 3].

Ограничения методов в контексте современных проектов

На основе анализа вышеизложенного можно выделить шесть значительных ограничений, общих для всех классических методов СППР.

Представим обнаруженные ограничения и их эффекты в таблице 1.

Таблица 1. Ограничения классических методы поддержки принятия решений

Ограничение	Суть	Эффект
Субъективность данных	Большинство входных параметров (сроки, риски, веса) основаны на экспертных оценках.	Подвержены когнитивным искажениям: оптимизм, эффект якоря, групповое мышление.
Недостаток данных	В инновационных проектах (стартапы, R&D, новые рынки) отсутствуют исторические данные.	Методы не могут работать без эмпирической основы – превращаются в «гипотетические игры».
Статичность	Все методы предполагают фиксированные критерии, альтернативы и структуру.	Не учитывают адаптацию, итеративность, изменения приоритетов в ходе проекта.
Инструментальная сложность	Реализация требует специализированного ПО и квалифицированных аналитиков.	Недоступны для малого бизнеса, стартапов, некоммерческих организаций.
Культурный барьер	Менеджеры часто предпочитают «интуитивные» решения, а не «аналитические».	Методы используются формально – как «документальное украшение», а не как инструмент.
Переоценка точности	Пользователи доверяют цифрам, даже если они построены на гипотезах.	Возникает иллюзия контроля: «Мы посчитали – значит, всё под контролем».

Возможности использования нейронных сетей в системах поддержки принятия решений

Нейронные сети трансформируют возможности систем поддержки принятия решений, обеспечивая переход от ручной аналитики к автоматизированному, адаптивному и предиктивному управлению. Их применение в СППР охватывает следующие ключевые области:

1. Прогнозирование проектных рисков и отклонений

Нейронные сети способны выявлять скрытые корреляции между сотнями факторов – от погодных условий на стройплощадке до уровня вовлечённости команды, частоты изменений требований и качества коммуникаций. Например, LSTM-сеть, обученная на данных 500 исторических IT-проектов, может предсказать вероятность просрочки сроков с точностью 87–92% на основе только начальных параметров: числа задач, количества участников, сложности требований и темпа выполнения первых спринтов.

2. Классификация и приоритизация рисков

На основе исторических данных о рисках (их тип, частота, влияние, меры реагирования) нейронные сети автоматически классифицируют новые риски, не требуя ручного ввода экспертных оценок. Модель MLP, обученная на данных PMI и ISO 31000, может присваивать риску уровень «критический», «высокий» или «умеренный» с точностью, превышающей экспертные оценки на 15–20%.

3. Поддержка выбора альтернативных стратегий

В условиях многокритериальности (например, выбор подрядчика по критериям: цена, срок, качество, репутация, экологичность) нейронные сети могут выступать в роли неявного весового механизма. Вместо того чтобы задавать веса вручную (как в МАИ), сеть обучается на исторических решениях, где уже были приняты оптимальные решения. Это позволяет учитывать неявные, но значимые факторы – например, влияние личных отношений с поставщиком, которое редко формализуется, но сильно влияет на результат.

4. Анализ неструктурированных данных

Современные проекты генерируют огромные объёмы неструктурированных данных: электронные письма, отчёты, дистанционные встречи, логи систем мониторинга. CNN и трансформеры (например, BERT) способны анализировать текстовые данные, выявляя эмоциональный тон, уровень напряжённости в команде, скрытые конфликты или признаки вы-

горания – и преобразовывать их в количественные индикаторы рисков, которые затем интегрируются в СППР.

5. Генерация сценариев и симуляция

Гибридные системы, сочетающие нейронные сети с методом Монте-Карло, позволяют генерировать реалистичные сценарии развития проекта. Например, GAN может создавать тысячи синтетических, но правдоподобных вариантов развития проекта на основе реальных паттернов, что позволяет тестировать стратегии управления в виртуальной среде до их внедрения.

6. Адаптивное обучение и самообновление

Нейронные сети могут быть интегрированы в цифровые двойники проектов – виртуальные копии, которые постоянно обновляются данными из реального мира. Это позволяет СППР не только реагировать на изменения, но и адаптироваться к ним: модель переобучается на новых данных, улучшая точность прогнозов и рекомендаций в реальном времени.

Ограничения нейронных сетей в контексте проектного управления

Несмотря на впечатляющие возможности, применение нейронных сетей в СППР сопряжено с рядом существенных ограничений, которые не могут быть игнорированы при проектировании систем поддержки решений.

1. Проблема «чёрного ящика»

Нейронные сети, особенно глубокие, обладают низкой объяснимостью (explainability). Их внутренние механизмы принятия решений часто непрозрачны – невозможно точно определить, почему сеть классифицировала риск как «критический». Это противоречит принципам управленческой ответственности, где решения должны быть аудируемыми и обоснованными. В условиях регулируемых отраслей (например, аэрокосмическая промышленность, финансы, здравоохранение) это ограничивает использование «непонятных» моделей без дополнительных механизмов интерпретации (например, SHAP, LIME).

2. Зависимость от объёма и качества данных

Нейронные сети требуют больших объёмов качественных, сбалансированных и репрезентативных данных. В управлении проектами данные часто фрагментарны, неполны или смещены (например, в выборке только успешные проекты, а не все, включая проваленные). Модель, обученная на таких данных, может выдавать ошибочные, но уверенные прогнозы – так называемое «переобучение».

3. Высокие вычислительные затраты

Обучение глубоких моделей требует значительных ресурсов: мощных GPU, времени и энергии. Это делает их непрактичными для мелких проектов или организаций с ограниченными ИТ-бюджетами. Кроме того, частое переобучение в реальном времени может привести к нестабильности системы.

4. Сложность интеграции в существующие процессы

Нейронные сети редко работают в изоляции. Их необходимо интегрировать в существующие системы управления проектами (например, MS Project, Jira, Primavera), что требует разработки API, согласования форматов данных и адаптации рабочих процессов. Многие менеджеры проектов не обладают технической подготовкой для работы с такими системами, что создаёт барьер внедрения.

5. Этические и правовые риски

Использование нейронных сетей в принятии управленческих решений порождает вопросы ответственности: кто несёт ответственность за ошибку, если решение было принято на основе прогноза модели? Если модель демонстрирует предвзятость (например, систематически занижает оценку рисков проектов в развивающихся странах), это может привести к дискриминации и юридическим последствиям.

6. Отсутствие креативности и контекстного понимания

Нейронные сети не обладают пониманием смысла, контекста или этических норм. Они могут предсказать, что проект с высоким уровнем изменений требует больше ресурсов, но не смогут оценить, почему эти изменения происходят – из-за плохого планирования, нестабильных требований заказчика или коррупционных схем. Это ограничивает их роль до инструмента поддержки, а не замены эксперта.

Заключение

Классические методы, несмотря на свою математическую строгость, находятся в зависимости от экспертных оценок, статичности критериев, неспособности учитывать нелинейные взаимосвязи и слабой адаптивности к изменениям. В связи с этим они часто оказываются неэффективными в инновационных, стартап- или R&D-проектах.

Нейронные сети, в свою очередь, предлагают принципиально новый подход: способность к автоматическому извлечению скрытых паттернов из больших объёмов структурированных и неструктурированных данных, прогнозированию отклонений с высокой точностью, адаптивному обучению в реальном времени и генерации реалистичных сценариев. Их применение позволяет трансформировать СППР из статических аналитических инструментов в динамические, предиктивные и самообучающиеся системы, способные интегрировать данные для формирования комплексной картины проектного риска.

При этом нейронные сети не являются универсальным решением и требуют осознанного и осторожного подхода к их внедрению. В регулируемых отраслях и в условиях ограниченных ресурсов использование «непонятных» моделей без механизмов объяснимости (SHAP, LIME) может быть недопустимо.

Таким образом, ключевой вывод работы заключается в том, что будущее СППР лежит не в замене человека машиной, а в синтезе человеческого опыта и машинного интеллекта.

Наиболее перспективным является развитие гибридных систем, где нейронные сети выступают как аналитический помощник для экспертов: они обрабатывают данные, выявляют аномалии, прогнозируют риски и предлагают альтернативы, а человек принимает окончательное решение.

Библиографический список

1. Кузнецова, Е. В. Управление портфелем проектов как инструмент реализации корпоративной стратегии : учебник для вузов / Е. В. Кузнецова. – 2-е изд., перераб. и доп. – Москва : Издательство Юрайт, 2026. – 177 с. – (Высшее образование). – ISBN 978-5-534-07425-3. – Текст : электронный // Образовательная платформа Юрайт [сайт]. – URL: <https://urait.ru/bcode/584065> (дата обращения: 20.01.2026).
2. Садовникова, Н. П. Системы поддержки принятия решений : учеб. пособие / Н. П. Садовникова, Д. С. Парыгин, М. В. Щербаков ; ВолгГТУ. – Волгоград, 2021. – 108 с. https://xn--80aai1dk.xn--p1ai/images/files/obrazovanie3/MU_B105_ISTM_2022.pdf
3. Системы поддержки принятия решений : учебник и практикум для вузов / под редакцией В. Г. Халина, Г. В. Черновой. – Москва : Издательство Юрайт, 2026. – 501 с. – (Высшее образование). – ISBN 978-5-534-20449-0. – Текст : электронный // Образовательная платформа Юрайт [сайт]. – URL: <https://urait.ru/bcode/583243> (дата обращения: 21.01.2026).
4. Kerzner, H. (2017). Project Management: A Systems Approach to Planning, Scheduling, and Controlling. 14th ed. Wiley.
5. Saaty, T. L. (2008). Decision Making with the Analytic Hierarchy Process. International Journal of Services Sciences https://www.researchgate.net/publication/228628807_Decision_making_with_the_Analytic_Hierarchy_Process

УДК 004.43; ГРНТИ 50.05.09

СРАВНИТЕЛЬНЫЙ АНАЛИЗ C++, GO И RUST ДЛЯ РЕАЛИЗАЦИИ ПОДХОДА БЫСТРОГО ПОЛУЧЕНИЯ ПЕРВЫХ РЕЗУЛЬТАТОВ

Н.В. Гудинов*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, nicko54lay@mail.ru*

Аннотация. В статье представлен углубленный сравнительный анализ трех языков программирования — C++, Go и Rust, которые ранее были определены как наиболее перспективные для реализации подхода минимизации времени получения первых результатов обработки данных (Quick First Benefits, QFB). На основе детального сопоставления механизмов генерации частичных результатов, моделей параллелизма, систем обработки ошибок, эффективности утилизации ресурсов и зрелости инструментария обосновывается выбор Go в качестве оптимальной платформы для последующей разработки специализированного программного обеспечения.

Ключевые слова: частичные результаты, языки программирования, C++, Go, Rust, параллельные вычисления, конвейерная обработка, фреймворк, производительность, обработка ошибок, Quick First Benefits.

COMPARATIVE ANALYSIS OF C++, GO, AND RUST FOR IMPLEMENTING THE FAST FIRST RESULTS APPROACH

N.V. Gudinov*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, nicko54lay@mail.ru*

The summary. The article presents an in-depth comparative analysis of three programming languages — C++, Go, and Rust — previously identified as the most promising for implementing the approach of minimizing the time to obtain the first results of data processing (Quick First Benefits, QFB). Based on a detailed comparison of partial result generation mechanisms, concurrency models, error handling systems, resource utilization efficiency, and tooling maturity, the choice of Go as the optimal platform for subsequent development of specialized software is substantiated.

Keywords: partial results, programming languages, C++, Go, Rust, parallel computing, pipeline processing, framework, performance, error handling, Quick First Benefits.

Введение

В предыдущих исследованиях автора [1, 2] был предложен и обоснован подход к организации вычислений, ориентированный на минимизацию времени получения первых (частичных) результатов обработки данных. Для обозначения этого подхода вводится термин Quick First Benefits (QFB). Было показано, что три языка — C++, Go и Rust — обладают наиболее выраженными предпосылками для создания на их основе специализированных решений, реализующих QFB-подход. Настоящая работа ставит целью провести детальный сравнительный анализ именно этих трех ЯП, выявить их сильные и слабые стороны применительно к QFB-подходу и обосновать выбор одного из них в качестве базовой платформы для последующей разработки инструментария.

Актуальность исследования обусловлена растущей потребностью в создании высоко отзывчивых систем обработки данных, способных выдавать промежуточные результаты без ожидания завершения полного цикла вычислений. Это критически важно для систем реального времени, интерактивной аналитики и потоковой обработки [3; 7].

Критерии сравнительного анализа

Для объективного сопоставления C++, Go и Rust используется расширенный набор критериев, учитывающий специфику QFB-подхода.

1. Механизмы передачи частичных результатов: наличие встроенных конструкций ЯП (генераторы, сопрограммы), позволяющих функции приостанавливать выполнение с возвратом промежуточного значения и возобновлять его позднее.

2. Модель параллелизма и каналы связи: способы создания параллельных задач и организации безопасного обмена данными между ними, удобство построения конвейеров (pipeline).

3. Обработка ошибок в асинхронных потоках: модель обработки ошибочных ситуаций, возникающих при генерации частичных результатов, и ее влияние на читаемость и надежность кода.

4. Управление ресурсами и производительность: эффективность использования процессора и памяти, предсказуемость задержек (latency), влияние сборки мусора.

5. Зрелость инструментария и простота разработки: наличие библиотек для работы с потоками данных, качество документации, удобство написания и поддержки QFB-ориентированного кода.

Детальный анализ языка программирования C++ (ISO C++20/23)

C++ предоставляет разработчику максимальный контроль над аппаратными ресурсами. Для реализации QFB-подхода в C++20 появились сопрограммы (*coroutines*), которые позволяют приостанавливать и возобновлять выполнение функции [4, 10].

Передача частичных результатов. Сопрограммы C++ являются мощным, но низкоуровневым механизмом. Для их использования требуется значительный объем «инфраструктурного» кода (определение типа *promise_type*, настройка поведения). Хотя существуют библиотеки (например, *srpcoro*), упрощающие работу, порог входа остается высоким.

Параллелизм. C++ предлагает богатый арсенал: *std::thread*, пулы потоков, параллельные алгоритмы STL (*std::execution::par*). Однако организация безопасного обмена данными между сопрограммами или потоками ложится на плечи разработчика (использование мьютексов, атомарных операций, *lock-free* структур), что повышает риск ошибок.

Обработка ошибок. Доминирующий механизм – исключения. В контексте конвейерной обработки исключение, выброшенное в глубине конвейера, должно быть корректно обработано или передано дальше, что усложняет логику и может привести к утечкам ресурсов, если не используются RAII (Resource Acquisition Is Initialization) обертки.

Управление ресурсами. Благодаря отсутствию сборщика мусора и возможности тонкой настройки, C++ обеспечивает максимальную производительность и минимальные задержки. Однако это требует от разработчика высокой квалификации для избежания утечек памяти и гонок данных.

Инструментарий. Обширен, но стандартизированных высокоуровневых абстракций для QFB именно в виде конвейеров данных пока нет. Разработка фреймворка потребует написания значительного объема кода поверх существующих библиотек.

Детальный анализ языка программирования Go

Go изначально проектировался как ЯП для создания эффективных и масштабируемых сетевых сервисов и систем обработки данных. Его главные особенности – горутины и каналы [5].

Передача частичных результатов. В Go нет генераторов в классическом понимании (как *yield* в Python). Однако идиоматичный способ построения конвейеров – использование горутин, которые читают данные из входного канала, обрабатывают их и пишут результат в выходной канал. Функция, запущенная как горутина, может отправить частичный результат в канал и продолжить работу, идеально соответствуя концепции QFB.

Параллелизм. ЯП Go не просто поддерживает параллелизм – он построен вокруг него. Горутины легковесны (сотни тысяч в одном процессе), а каналы являются встроенным

механизмом коммуникации, обеспечивающим типобезопасный обмен данными между горутинами и поддерживаемым на уровне синтаксиса и рантайма. Конструирование сложных конвейеров из нескольких этапов превращается в написание понятного и линейного кода.

Обработка ошибок. В Go принят явный подход: функция часто возвращает кортеж (результат, ошибка). В контексте конвейера это позволяет легко передавать ошибку как часть потока данных, например, через отдельный канал ошибок или используя тип `Result` из пакета `errors`. Отсутствие исключений делает поток данных предсказуемым.

Управление ресурсами. Go – компилируемый ЯП с собственным рантаймом и сборщиком мусора. Сборщик мусора постоянно совершенствуется и вносит минимальные задержки (`low-pause GC`), что приемлемо для большинства QFB-сценариев. Утилизация ресурсов очень высока, а код при этом остается простым.

Инструментарий. Стандартная библиотека Go включает мощные пакеты для работы с потоками данных (например, `io.Pipe`). Сторонние библиотеки, такие как `fx` от Uber, предоставляют высокоуровневые абстракции для построения приложений, но специализированных QFB-фреймворков нет, что открывает поле для деятельности. ЯП чрезвычайно прост в изучении, что снижает порог входа.

Детальный анализ языка программирования Rust

Rust позиционируется как ЯП, обеспечивающий безопасность памяти без сборщика мусора и производительность на уровне C++, что делает его крайне привлекательным для системного программирования [6; 8].

Передача частичных результатов. Rust поддерживает асинхронные генераторы через комбинацию `async/await` и потоков (`streams`). Библиотека `futures` предоставляет типаж `Stream`, который является асинхронным аналогом `Iterator` и идеально подходит для представления потока частичных результатов.

Параллелизм. Модель владения (`ownership`) и заимствования (`borrowing`) в Rust гарантирует отсутствие гонок данных на этапе компиляции. Это колоссальное преимущество для надежности. Асинхронный рантайм (например, `tokio` или `async-std`) подключается как библиотека, что дает гибкость, но может приводить к фрагментации решений.

Обработка ошибок. Тип `Result<T, E>` делает обработку ошибок явной и обязывающей. Комбинаторы вроде `map_err` и `and_then` позволяют элегантно обрабатывать ошибки при трансформации потоков.

Управление ресурсами. Rust дает производительность, сопоставимую с C++. Отсутствие сборщика мусора и детерминированное освобождение памяти (через типаж `Drop`) делают его идеальным для систем с жесткими требованиями к задержкам. Однако это достигается ценой высокой сложности кривой обучения.

Инструментарий. Мощный, но ориентированный на системное программирование. Крейты `tokio` и `async-stream` предоставляют все необходимые кирпичики для создания QFB-конвейеров, но требуют глубокого понимания асинхронного Rust и его типов.

Сравнительная оценка и обоснование выбора Go

Для наглядности результаты сравнения сведены в таблицу 1.

Анализ показывает, что C++ дает максимальную производительность, но требует экстремальной квалификации и несет риски низкоуровневых ошибок. Rust гарантирует безопасность и производительность, но его парадигма существенно замедляет разработку и усложняет прототипирование.

Таблица 1. Сравнительная характеристика C++, Go и Rust

Критерий	C++	Go	Rust
Передача частичных результатов	Сопрограммы (низкоуровневые, сложные)	Горутины + каналы (идиоматично, просто)	Асинхронные потоки (мощно, но сложно)
Модель параллелизма	Потоки ОС, сложное ручное управление	Легковесные горутины, встроенные каналы	Асинхронные рантаймы, безопасность на этапе компиляции
Обработка ошибок в конвейере	Исключения (нарушают поток, сложны в конвейере)	Явный возврат ошибок (легко интегрируется в поток)	<i>Result<T, E></i> (явно, мощные комбинаторы)
Простота написания QFB-кода	Низкая (требуется опыт)	Высокая (минималистичный синтаксис)	Средняя/Низкая (крутая кривая обучения)
Управление памятью	Ручное/RAII (макс. контроль, риск ошибок)	Сборка мусора (достаточно эффективна)	Владение и заимствование (без GC, безопасно)
Зрелость инструментария для QFB	Низкая (есть кирпичики, нет целостного фреймворка)	Средняя (есть идиомы, нет готового фреймворка)	Средняя (есть мощные библиотеки, но фрагментированы)
Основной риск	Ошибки работы с памятью, сложность	Задержки при сборке мусора	Сложность освоения и разработки

Go предлагает наилучший компромисс. Его главное преимущество – простота создания, поддержки и отладки конкурентных конвейеров. Легкость горутин и выразительность каналов делают процесс конструирования QFB-систем быстрым и менее подверженным ошибкам по сравнению с ручным управлением потоками в C++. Явная обработка ошибок, в отличие от механизма исключений, лучше вписывается в парадигму потока данных. Достаточно высокая производительность и эффективный сборщик мусора покрывают подавляющее большинство практических задач, для которых предназначен QFB-подход (обработка коллекций данных, потоковые сервисы).

Таким образом, именно Go является наиболее прагматичным выбором для создания фреймворка, ориентированного на быстрое получение первых результатов.

Заключение

Проведенный сравнительный анализ показал, что среди трех ЯП Go является наиболее сбалансированным решением для реализации подхода быстрого получения первых результатов (QFB). Его встроенная модель параллелизма, простота и достаточная производительность создают идеальную основу для последующей разработки специализированного программного обеспечения. Дальнейшие исследования будут направлены на разработку требований к такому инструментарию и определение его ключевых компонентов.

Библиографический список

1. Гудинов Н.В. Исследование возможности быстрого получения первых результатов в инструментарии современных языков программирования // Актуальные проблемы современной науки и производства: сборник материалов X Всероссийской научно-технической конференции. Том 2. — 2025. — С. 62-68.
2. Гудинов Н.В. Исследование возможности быстрого получения первых результатов в инструментарии современных языков программирования // Математическое и программное обеспечение вычислительных систем: межвузовский сборник научных трудов. — Рязань: РГРТУ, 2026. — (принято к публикации).
3. Клепман М. Высоконагруженные приложения. Программирование, масштабирование, поддержка. — СПб.: Питер, 2021. — 640 с.
4. Страуструп Б. Язык программирования C++. — 4-е изд. — М.: Бином, 2022. — 1216 с.
5. Донован А.А., Керниган Б.У. Язык программирования Go. — М.: Вильямс, 2016. — 432 с.
6. Клабник С., Николс К. Язык программирования Rust. — СПб.: Питер, 2020. — 721 с.
7. White T. Hadoop: The Definitive Guide. — 3th ed. — O'Reilly Media, 2012. — 688 p. (главы по потоковой обработке).
8. Blandy J., Orendorff J. Programming Rust. — 2nd ed. — O'Reilly Media, 2021. — 711 p.
9. Cox R., Griesemer R., Pike R. The Go Programming Language and Environment // Communications of the ACM. — 2022. — Vol. 65, No. 5. — P. 64-71.
10. ISO/IEC 14882:2024. Programming Language C++. — ISO, 2024. — 1879 p.

УДК 004.8; ГРНТИ 28.23

УЧЁТ ВНЕШНИХ ФАКТОРОВ В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ СПРОСА НА ТОВАРЫ В РОЗНИЧНОЙ ТОРГОВЛЕ

Г.В. Овечкин, И.А. Прилепский

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, ilyaprilepskiy1@gmail.com

Аннотация. В работе рассматриваются теоретические аспекты учёта внешних факторов при построении прогнозных моделей спроса на товары в розничной торговле. Приводится классификация внешних факторов, включающая календарные события, промо-акции, ценовые изменения и операционные факторы. Анализируются основные проблемы их интеграции: проблема «заглядывания в будущее», мультиколлинеарность, выбор способов кодирования категориальных переменных и учёт временных сдвигов влияния факторов. Предлагаются теоретические подходы к решению данных проблем на основе гибридного метода, сочетающего сезонно-трендовую декомпозицию и градиентный бустинг.

Ключевые слова: прогнозирование спроса, внешние факторы, градиентный бустинг, проблема «заглядывания в будущее», промо-акции, розничная торговля, временные ряды.

ACCOUNTING FOR EXTERNAL FACTORS IN DEMAND FORECASTING PROBLEMS FOR RETAIL GOODS

G.V. Ovechkin, I.A. Prilepskiy

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, ilyaprilepskiy1@gmail.com

The summary. The paper considers the theoretical aspects of accounting for external factors in building predictive models of demand for goods in retail trade. A classification of external factors is given, including calendar events, promotions, price changes and operational factors. The main problems of their integration are analyzed: the lookahead problem, multicollinearity, the choice of methods for encoding categorical variables and accounting for time shifts in the influence of factors. Theoretical approaches to solving these problems based on a hybrid method combining seasonal-trend decomposition and gradient boosting are proposed.

Keywords: demand forecasting, external factors, gradient boosting, lookahead problem, promotions, retail trade, time series.

В современных условиях точность прогнозирования спроса становится определяющим фактором эффективности управления товарными запасами в розничной торговле. Традиционные модели прогнозирования, использующие только исторические временные ряды продаж, часто оказываются недостаточно точными, поскольку не учитывают множество внешних воздействий, влияющих на поведение покупателей. Праздничные дни, промо-акции, изменения цен, погодные условия, логистические ограничения способны существенно изменять объёмы продаж, и игнорирование этих факторов приводит к систематическим ошибкам. Цель настоящей работы – теоретический анализ проблем, возникающих при интеграции внешних факторов в прогнозные модели спроса, и систематизация подходов к их решению.

Классификация внешних факторов, влияющих на потребительский спрос.

Внешние факторы, оказывающие воздействие на спрос в розничной торговле, могут быть разделены на несколько категорий в зависимости от их природы и механизма влияния.

Календарные события включают праздничные и выходные дни, сезонные распродажи, периоды выплат заработных плат, школьные каникулы. Влияние этих факторов обусловлено сложившимися социальными и культурными традициями: например, в преддверии Нового года возрастает спрос на продукты питания и подарки, перед началом учебного года – на канцелярские товары и одежду для школьников. Важно отметить, что эффект календарных событий часто не ограничивается одним днём, а распределяется на предшествующий и последующий периоды.

Маркетинговые факторы охватывают проведение промо-акций, изменение цен, вывод на рынок новых товаров, рекламные кампании. Промо-акции могут вызывать краткосрочный рост продаж в несколько раз, однако после их окончания часто наблюдается спад из-за насыщения спроса или эффекта «каннибализации» будущих продаж. Ценовая эластичность спроса различна для разных товарных категорий и должна учитываться при построении моделей.

Внешние условия представлены погодными факторами, такими как температура воздуха, осадки, влажность, макроэкономическими показателями: инфляция, уровень безработицы, индекс потребительской уверенности, действиями конкурентов, изменениями в законодательстве. Для ряда товарных групп влияние погоды может быть определяющим: продажи мороженого и прохладительных напитков тесно коррелируют с температурой, продажи зимней резины – с выпадением первого снега.

Операционные факторы включают наличие товара на складе, логистические задержки, изменения в ассортименте, сбои в поставках. Учёт этих факторов необходим для корректной интерпретации исторических данных: если товара не было в наличии, нулевые продажи не отражают отсутствие спроса. Ведь эффективное управление запасами невозможно без учёта логистических ограничений и операционных сбоев, которые искажают реальную картину спроса [1].

Основные проблемы интеграции внешних факторов в прогнозные модели.

При включении внешних признаков в модели машинного обучения возникает ряд теоретических и практических проблем, без решения которых качество прогнозов не улучшается, а часто даже ухудшается.

Проблема «заглядывания в будущее». При обучении модели нельзя использовать информацию, которая в реальных условиях эксплуатации будет недоступна на момент построения прогноза. Если в обучающую выборку включить фактические значения внешних факторов на прогнозируемую дату, модель покажет завышенную точность на исторических данных, однако в реальном прогнозировании такая информация отсутствует. В обучающей выборке допустимо использовать только те значения внешних факторов, которые были известны в момент t , то есть плановые значения X_{t+h}^{plan} , но не фактические X_{t+h}^{actual} . Для промо-акций это план маркетингового отдела, для погодных факторов – прогноз погоды на соответствующий день, для праздников – твёрдо установленные даты.

Мультиколлинеарность. Внешние факторы часто коррелируют между собой или с внутренними характеристиками временного ряда. Например, проведение промо-акции обычно сопровождается снижением цены, а праздничные дни могут совпадать с запланированными распродажами. Сильная корреляция между признаками приводит к неустойчивости оценок модели и переобучению. Для количественной оценки мультиколлинеарности используется фактор инфляции дисперсии:

$$VIF_j = \frac{1}{1 - R_j^2},$$

где R_j^2 – коэффициент детерминации при регрессии j -го фактора на все остальные.

Значение $VIF > 10$ указывает на наличие сильной мультиколлинеарности. Теоретически проблема решается отбором наиболее информативных признаков, использованием регуляризации или объединением коррелирующих факторов в интегральные показатели, например, методом главных компонент.

Кодирование категориальных переменных. Многие внешние факторы имеют категориальную природу: праздники, типы промо-акций, категории товаров. Простейший способ кодирования – бинарные флаги (0 или 1) – не учитывает различий в силе влияния разных категорий. Например, государственные праздники оказывают более сильное влияние на

спрос, чем профессиональные. Более совершенный подход предполагает введение градации по значимости, то есть присвоение категориям весовых коэффициентов, отражающих их относительную важность. Для учёта эффектов, распределённых во времени, могут вводиться дополнительные признаки, характеризующие предпраздничные и постпраздничные дни. При большом количестве категорий может применяться целевое кодирование, когда категория заменяется средним значением целевой переменной для данной категории на обучающей выборке.

Учёт временных сдвигов влияния. Воздействие внешних факторов часто не ограничивается моментом их действия, а распределяется во времени. Эффект от промо-акции может наблюдаться в предшествующий период, когда потребители планируют покупки и в последующий период – насыщение спроса. Аналогично, влияние праздников может растягиваться на несколько дней до и после события. Для моделирования таких эффектов теоретически обоснованным является использование распределённых лагов. Пусть f_t – значение внешнего фактора в момент t . Тогда агрегированный признак, учитывающий влияние за период $[ts, te]$, может быть представлен в виде взвешенной суммы:

$$F_t = \sum_{i=ts}^{te} w_i \cdot f_{t+i},$$

где веса w_i убывают по мере удаления от момента воздействия и могут быть заданы или оценены по данным с помощью регрессионных методов.

Теоретические подходы к интеграции внешних факторов.

В качестве теоретического подхода к интеграции внешних факторов предлагается гибридный подход, сочетающий сезонно-трендовую декомпозицию и градиентный бустинг. Данный подход теоретически обоснован необходимостью баланса между интерпретируемостью модели и её способностью улавливать сложные нелинейные зависимости.

Метод сезонно-трендовой декомпозиции STL (Seasonal and Trend decomposition using Loess) позволяет представить временной ряд продаж в виде суммы трёх компонент:

$$y_t = T_t + S_t + R_t,$$

где T_t – тренд, характеризующий долгосрочную динамику;

S_t – сезонная компонента, отражающая регулярные колебания с фиксированным периодом;

R_t – остаточная компонента, содержащая нерегулярные флуктуации и влияние внешних факторов.

Преимущество декомпозиции заключается в том, что тренд и сезонность могут быть экстраполированы простыми методами, а сложное моделирование требуется только для остаточной компоненты. Как отмечают специалисты в области прогнозирования, декомпозиция временных рядов позволяет лучше понять структуру данных и отделить регулярные составляющие от нерегулярных, что особенно важно при включении внешних факторов [2].

Градиентный бустинг представляет собой ансамблевый метод машинного обучения, строящий композицию из деревьев решений, где каждое следующее дерево минимизирует ошибки предыдущих. Математически модель градиентного бустинга может быть записана как:

$$F(x) = \sum_{m=1}^M \gamma_m h_m(x),$$

где $h_m(x)$ – базовые модели (обычно деревья решений);

γ_m – весовые коэффициенты, подбираемые в процессе минимизации функции потерь.

Градиентный бустинг теоретически обоснован как эффективный метод для работы с табличными данными, содержащими разнородные признаки, включая как числовые, так и категориальные переменные.

В предлагаемом гибридном подходе остаточная компонента R_t моделируется с помощью градиентного бустинга, который на вход получает не только лаги остатка и скользящие статистики, но и все внешние факторы:

$$\hat{R}_{t+h} = F(R_{t-1}, R_{t-2}, \dots, R_{t-L}, MA_t, X_{t+h}^{plan}),$$

где MA_t – скользящие средние и другие агрегированные статистики, X_{t+h}^{plan} – плановые значения внешних факторов. Итоговый прогноз формируется суммированием экстраполированных тренда и сезонности с прогнозом остатка:

$$\hat{y}_{t+h} = \hat{T}_{t+h} + \hat{S}_{t+h} + \hat{R}_{t+h}.$$

Такая архитектура обладает важными теоретическими преимуществами. Она позволяет моделировать сложные нелинейные зависимости остатков от внешних факторов, не требуя огромных объёмов данных, характерных для нейросетевых подходов, а разделение компонент упрощает диагностику модели: если качество прогнозов падает, можно определить, какая именно компонента даёт сбой.

Библиографический список

1. Крылатков П.П., Прилуцкая М.А. Управление цепью поставок (SCM) : учеб. пособие. – Екатеринбург : Изд-во Урал. ун-та, 2018. – 140 с.
2. Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice. – 3rd ed. – Melbourne: OTexts, 2021. – 442 p. – URL: <https://otexts.com/fpp3/> (дата обращения: 21.02.2026).

УДК 004.8; ГРНТИ 28.23

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА КОДА И ОЦЕНКИ ОТВЕТОВ В СИСТЕМАХ ТЕХНИЧЕСКОГО ИНТЕРВЬЮИРОВАНИЯ

Г.В. Овечкин, К.А. Прилепский

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, prilepskiyk11@gmail.com

Аннотация. В работе рассматривается применение методов машинного обучения для автоматизации процессов технического интервьюирования и код-ревью. Описаны современные подходы к анализу исходного кода с использованием нейросетевых моделей (CodeBERT, графовые нейронные сети), а также методы обработки естественного языка для оценки полноты и правильности ответов кандидатов. Предложена архитектура веб-ориентированной системы на базе Java-технологий, объединяющей средства совместного редактирования кода, видео-связи и интеллектуальной поддержки интервьюера. Приведены возможные метрики качества работы таких моделей и направления дальнейших исследований. Статья будет полезна разработчикам систем автоматизации подбора персонала и специалистам в области искусственного интеллекта.

Ключевые слова: машинное обучение, анализ кода, обработка естественного языка, техническое интервью, оценка ответов, интеллектуальные системы, Java.

APPLICATION OF MACHINE LEARNING METHODS FOR CODE ANALYSIS AND ANSWER EVALUATION IN TECHNICAL INTERVIEW SYSTEMS

G.V. Ovechkin, K.A. Prilepskiy

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, prilepskiy11@gmail.com

The summary. The paper considers the application of machine learning methods to automate the processes of technical interviewing and code review. Modern approaches to source code analysis using neural network models (CodeBERT, graph neural networks) as well as natural language processing techniques for assessing the completeness and correctness of candidates' answers are described. An architecture of a web-oriented system based on Java technologies combining collaborative code editing tools, video communication, and intelligent interviewer support is proposed. Possible performance metrics for such models and directions for further research are given. The article will be useful for developers of recruitment automation systems and specialists in the field of artificial intelligence..

Keywords: machine learning, code analysis, natural language processing, technical interview, answer evaluation, intelligent systems, Java.

Введение

Современный рынок IT-специалистов характеризуется высоким спросом и жёсткой конкуренцией. Процесс технического интервью часто включает два ключевых этапа: написание кода в реальном времени (code-review) и устные ответы на вопросы. Оценка этих этапов субъективна и трудоёмка, что создаёт предпосылки для автоматизации с использованием методов искусственного интеллекта. В последние годы методы машинного обучения (МО) достигли значительных успехов в задачах анализа исходного кода и обработки естественного языка (NLP). Фундаментальные основы этих методов изложены в работах [1, 2].

Несмотря на существование отдельных инструментов для автоматической проверки кода (например, SonarQube) и платформ для проведения интервью (HackerRank, Codility), они не предоставляют комплексного решения, объединяющего анализ кода, оценку устных ответов и коммуникацию в реальном времени. Кроме того, большинство решений не используют современные методы глубокого обучения, которые способны учитывать семантику кода и естественного языка. Цель данной работы – рассмотреть возможности интеграции таких методов в единую систему технического интервьюирования, способную помогать интервьюеру в оценке решений кандидата. Особое внимание уделено реализации подобной системы с использованием экосистемы Java, что обеспечивает высокую производительность, кроссплатформенность и широкие возможности интеграции с существующими корпоративными системами. Выбор Java также обусловлен наличием зрелых библиотек для машинного обучения (DL4J, TensorFlow Java API) и инструментов синтаксического анализа кода (JavaParser).

Методы машинного обучения для анализа кода

Анализ программного кода может выполняться на разных уровнях: синтаксическом, семантическом и динамическом. Традиционные инструменты статического анализа основаны на формальных грамматиках и паттернах, однако они не способны обнаруживать сложные смысловые ошибки или оценивать стиль кода. Применение методов МО позволяет решить эту задачу более гибко.

Одним из перспективных подходов является использование предобученных моделей на архитектуре трансформера, адаптированных для кода, например CodeBERT [1]. Эта модель обучается на больших корпусах исходного кода и естественного языка, что позволяет ей выполнять такие задачи, как поиск семантически похожих фрагментов, выявление уязвимостей и генерация комментариев. Для структурного представления кода также приме-

няются графовые нейронные сети (GNN) [1], которые работают с абстрактными синтаксическими деревьями или графами потоков данных. Они способны учитывать зависимости между переменными и функциями, что повышает точность обнаружения ошибок.

В контексте технического интервью наибольший интерес представляет автоматическая проверка правильности алгоритмического решения. Модель может сравнить код кандидата с эталонным решением (или набором эталонов) и выдать оценку по шкале, например, от 0 до 1. Для обучения такой модели необходима размеченная выборка задач с множеством вариантов решений, что в настоящее время активно создаётся исследовательскими группами и платформами [3]. Метриками качества служат точность (precision), полнота (recall) и F1-мера [2]:

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (1).$$

На платформе Java возможно использование библиотек глубокого обучения (DL4J, TensorFlow Java API) для загрузки и выполнения предобученных моделей, а также инструментов синтаксического разбора кода (JavaParser) для извлечения признаков, необходимых моделям. Например, JavaParser позволяет построить абстрактное синтаксическое дерево (AST) Java-кода, которое затем можно преобразовать в графовое представление для подачи на вход GNN. Такой подход уже применяется в исследовательских проектах и показал свою эффективность для задач классификации кода по наличию ошибок.

Кроме того, существуют специализированные библиотеки, такие как Code2Vec, которые преобразуют код в векторные представления, пригодные для сравнения. В Java-реализации можно использовать модель, экспортированную в формат ONNX, и выполнять её с помощью ONNX Runtime для Java, что обеспечивает высокую производительность инференса.

Оценка ответов на естественном языке

Второй компонент технического интервью – устные или письменные ответы на вопросы по теории, архитектуре систем и т.п. Для автоматической оценки таких ответов применяются методы NLP. Классический подход заключается в вычислении семантической близости между ответом кандидата и эталонным ответом с помощью векторных представлений слов (word embeddings) или более сложных моделей типа BERT, обученных на корпусе вопросов-ответов из IT-области [1].

Однако простое сравнение векторов не всегда учитывает полноту ответа. Поэтому дополнительно можно использовать модели для определения наличия ключевых понятий и связей между ними [2]. Для этого ответ сначала преобразуется в набор сущностей (например, с помощью NER-моделей), а затем сравнивается с множеством ожидаемых сущностей. Такой подход позволяет частично автоматизировать проверку знаний, снижая нагрузку на интервьюера.

В Java-экосистеме для этих целей можно применять Apache OpenNLP, Stanford CoreNLP или интеграцию с Python-сервисами через gRPC. Например, Apache OpenNLP предоставляет готовые модели для распознавания именованных сущностей (NER) и токенизации, которые можно обучить на предметно-ориентированных корпусах (вопросы по Java, алгоритмам, базам данных). Stanford CoreNLP поддерживает множество языков и обладает широкими возможностями синтаксического анализа, что полезно для оценки сложных ответов.

Особую сложность представляет обработка устной речи. В реальном интервью кандидат может давать ответы голосом, поэтому необходимо транскрибировать речь в текст. Для этого можно использовать облачные сервисы (Yandex SpeechKit, Google Cloud Speech-to-Text) или локальные модели на базе Vosk. Java имеет клиентские библиотеки для взаимодействия с этими API. Полученный текст затем передаётся в модуль NLP-оценки. При

этом важно учитывать возможные ошибки распознавания и применять методы повышения робастности (например, использование N-грамм и нечёткого поиска).

Архитектура интеллектуальной системы интервьюирования на Java

Разрабатываемая система представляет собой веб-приложение на базе Spring Boot, объединяющее функциональность совместного редактирования кода, видео-конференции и модулей ИИ. Обобщённая архитектура включает следующие компоненты, каждый из которых реализован как отдельный микросервис для обеспечения масштабируемости и независимости развёртывания:

- **Клиентская часть** – одностраничное приложение (React/Angular) с редактором кода (например, Monaco Editor), поддерживающим подсветку синтаксиса и запуск тестов, а также окно видео-связи (WebRTC). Клиент общается с сервером через REST API и WebSocket для получения обновлений в реальном времени.
- **Серверная часть (Gateway)** – реализована на Java (Spring Boot), обрабатывает запросы от клиента, управляет сессиями интервью, аутентификацией и авторизацией, хранит данные в реляционной БД (PostgreSQL). Для асинхронного взаимодействия с модулями ИИ используется очередь сообщений (RabbitMQ/Kafka), что позволяет избежать блокировок и обеспечивает отказоустойчивость.
- **Модуль анализа кода** – развернут как отдельный микросервис на Java, который получает код кандидата через очередь или REST, выполняет его синтаксический разбор (JavaParser) и передаёт в предобученную модель (загруженную с помощью DL4J или ONNX Runtime) для оценки корректности. Результат (процент совпадения с эталоном, список замечаний) возвращается интервьюеру в реальном времени через WebSocket. Для ускорения инференса можно использовать кэширование результатов для часто встречающихся задач.
- **Модуль NLP-оценки** – также микросервис, который транскрибирует речь кандидата (с использованием внешних API или локальных моделей) и сравнивает текстовый ответ с эталоном, применяя библиотеки OpenNLP или вызывает модель BERT через TensorFlow Java. Модуль также может оценивать соответствие ответа заданным критериям и генерировать рекомендации для интервьюера.
- **Модуль видео-связи** – на базе WebRTC с использованием сервера сигнализации на Java (например, Kurento или собственный на Spring и STOMP). Этот модуль управляет установкой соединений, передачей медиапоток и записью интервью (при необходимости).

Для коммуникации между микросервисами используется REST/gRPC. Все взаимодействия защищены шифрованием (HTTPS, WSS). Преимущества Java-стека: надёжность, высокая производительность, обширная экосистема для построения распределённых приложений и интеграции с ML-библиотеками. Кроме того, Spring Cloud предоставляет инструменты для обнаружения сервисов, балансировки нагрузки и конфигурации, что упрощает развёртывание в облачных средах.

Архитектура предполагает возможность горизонтального масштабирования модулей ИИ, так как они могут быть узким местом при большом количестве одновременных интервью. Для этого используется оркестрация контейнеров (Kubernetes) и автоматическое масштабирование в зависимости от загрузки очереди.

Экспериментальная оценка эффективности

Для проверки работоспособности предложенных методов планируется провести серию экспериментов с использованием открытых датасетов и собранных записей реальных интервью (с согласия участников). На первом этапе будет выполнено тестирование модуля анализа кода на датасете CodeSearchNet, который содержит пары «запрос на естественном

языке – код» и размеченные ошибки. Метриками эффективности выступят точность классификации верных/неверных решений и F1-мера [2]. Ожидается, что использование CodeBERT позволит достичь точности не менее 85% на задачах поиска семантических ошибок.

Для модуля NLP-оценки будет сформирован корпус вопросов и ответов по темам, типичным для Java-интервью (многопоточность, коллекции, шаблоны проектирования). Оценка ответов экспертами будет сравниваться с автоматической оценкой на основе семантической близости. В качестве метрики используется ранговая корреляция Спирмена между оценками экспертов и автоматическими. Предполагается, что корреляция составит не менее 0,7, что считается приемлемым для ассистирующих систем.

Кроме того, планируется нагрузочное тестирование системы с использованием Apache JMeter для оценки максимального количества одновременных сессий и времени отклика модулей ИИ. Будет измеряться среднее время анализа кода (должно быть не более 2 секунд для комфортной работы интервьюера). При необходимости будет применено кэширование результатов и оптимизация моделей (квантизация, обрезка).

Важным аспектом является сбор обратной связи от интервьюеров, которые будут использовать прототип системы. Планируется провести пилотное тестирование в нескольких IT-компаниях, чтобы оценить субъективное удовлетворение и сокращение времени интервью. Ожидается, что совместное использование моделей повысит объективность и сократит время интервью на 20–30 % за счёт автоматизации рутинных проверок.

Библиографический список

1. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. – М.: ДМК Пресс, 2018. – 652 с.
2. Хейсти Т., Тибириани Р., Фридман Дж. Элементы статистического обучения: интеллектуальный анализ данных, вывод и прогнозирование. – М.: ДМК Пресс, 2015. – 782 с.
3. LeetCode Platform [Электронный ресурс]. – Режим доступа: <https://leetcode.com> (дата обращения: 25.02.2026).

УДК 004.89; ГРНТИ 28.23.29

ИССЛЕДОВАНИЕ СПОСОБОВ АВТОМАТИЗИРОВАННОЙ КАТЕГОРИЗАЦИИ ФАЙЛОВ

А.Н. Попов, С.Ю. Жулева

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, poa18.06.04@yandex.ru

Аннотация. Работа решает проблему несоответствия между когнитивными моделями пользователей и иерархией файловых систем. Предложен двухуровневый алгоритм категоризации, с использованием модели ruBERT-tiny2 для контекстного анализа.

Ключевые слова: категоризация, контекстный анализ, теги, ruBERT, файловые системы.

AUTOMATIC FILE CATEGORIZATION METHODS RESEARCH

A.N. Popov, S.Y. Zhuleva

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, poa18.06.04@yandex.ru

The summary. This research offers a solution for the problem of inconsistency between users cognitive models and file systems hierarchy. It is suggested to use a two-level categorization algorithm with the use of the AI-model ruBERT-tiny2 for contextual analysis.

Keywords: categorization, contextual analysis, tags, ruBERT, file systems.

Современные файловые системы продолжают сталкиваться с проблемами несоответствия пользовательским когнитивным моделям. Исследования показывают, что люди тратят до 20% рабочего времени на поиск и организацию файлов [2]. Фундаментальная

проблема заключается в том, что запоминается содержание файлов, а не их технические атрибуты, в то время как традиционные файловые системы ориентированы на иерархию папок и метаданные [4].

Анализ особенностей пользовательских коллекций и когнитивных моделей показал значительное разнообразие цифровых данных. В личных коллекциях преобладают мультимедийные файлы, тогда как в рабочих чаще встречаются текстовые документы и электронные таблицы [1]. Пользователи ОС *Mac* имеют в среднем больше файлов меньшего размера по сравнению с пользователями *Windows*, что связано с различиями в индивидуальных предпочтениях [2]. Исследования подтверждают, что навигация по файлам использует те же структуры мозга, что и навигация в физическом пространстве [3]. Также люди предпочитают навигацию по папкам поиску, так как она требует меньше когнитивных ресурсов. При этом эмоциональное состояние влияет на организацию файлов: высокий уровень тревожности побуждает создавать более сложные структуры с большим количеством папок, содержащих небольшое количество тесно связанных файлов [3].

Типичный сценарий взаимодействия пользователя с файловой системой представлен на рисунке 1.

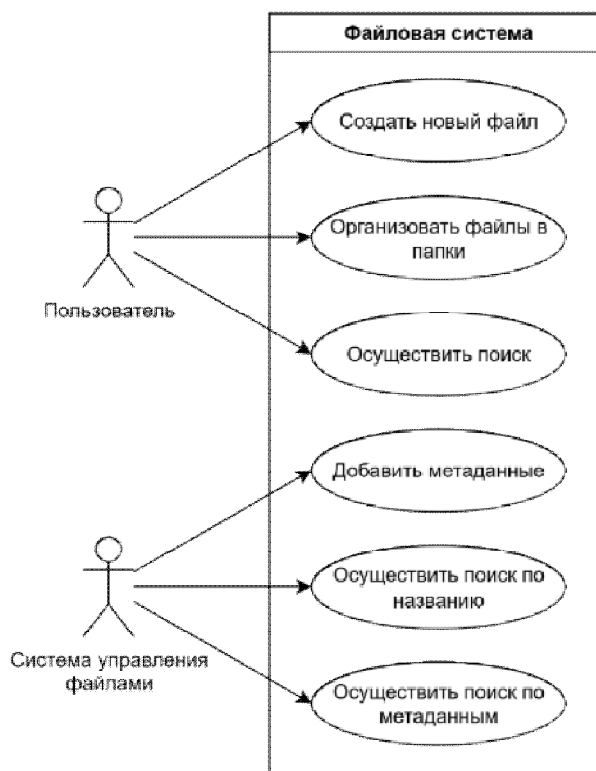


Рис. 1. Сценарий взаимодействия пользователя с файловой системой

Представим основные методы категоризации файлов.

Методы на основе правил/шаблонов (рисунок 2) отличаются интерпретируемостью и контролируемостью, но требуют значительных усилий для разработки и плохо обобщаются на новые типы файлов [5].

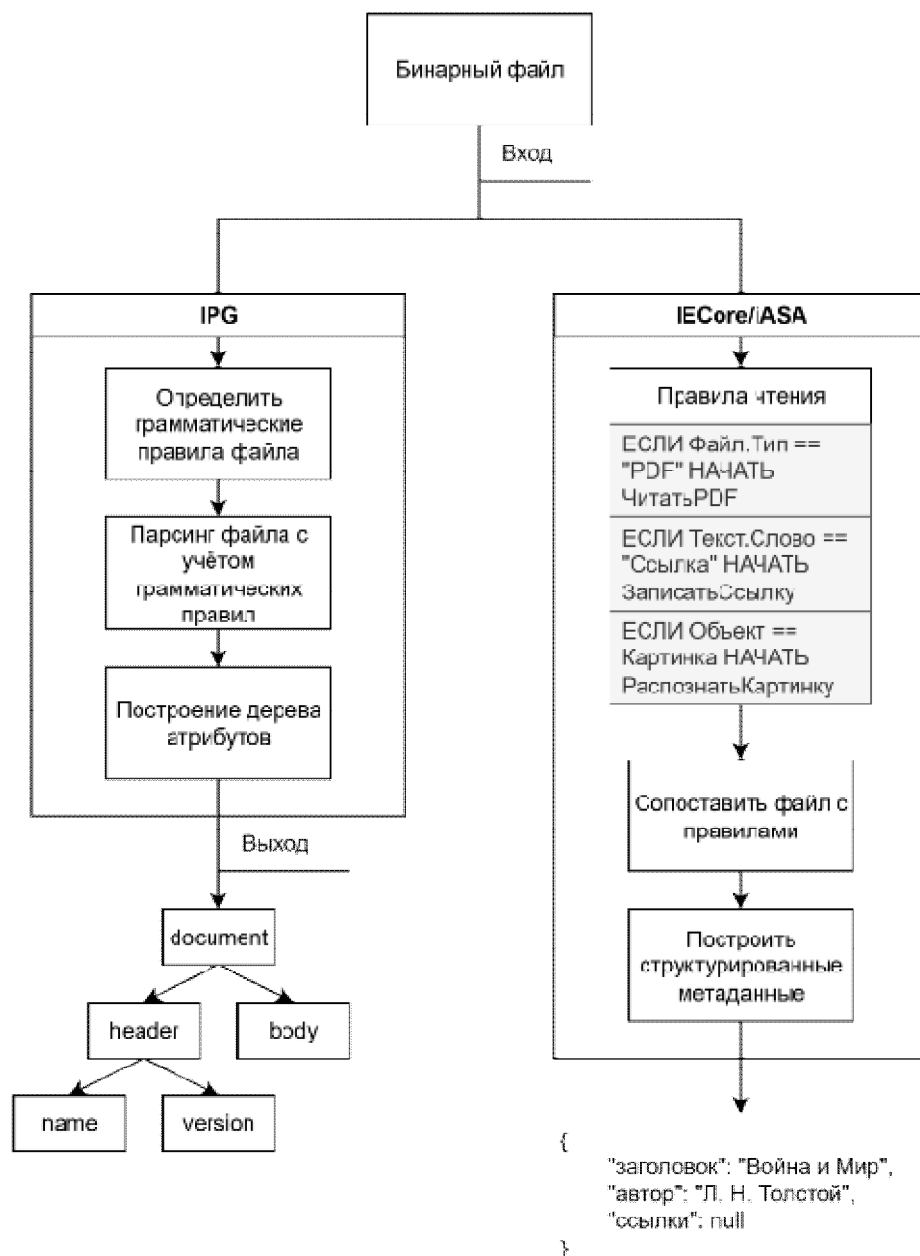


Рис. 2. Схема работы методов на основе правил/шаблонов

Статистические методы (рисунок 3) просты в реализации и эффективны для выделения ключевых терминов, но игнорируют порядок слов и не учитывают семантические связи между ними [6].

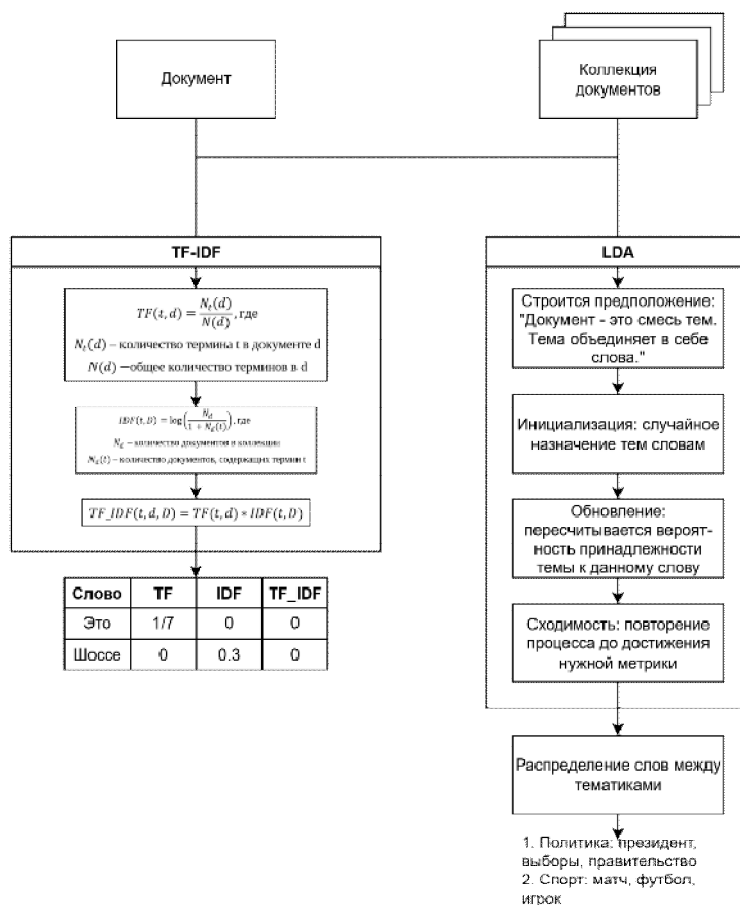


Рис. 3. Схема работы статистических методов

Современные методы на основе трансформеров, такие как BERT (рисунок 4), обеспечивают глубокое понимание контекста и высокую точность за счет обучения на больших объемах данных, но отличаются высокой вычислительной сложностью [7].



Рис. 4. Упрощенная схема работы моделей типа BERT

Сравнительная характеристика подходов к категоризации файлов представлена в таблице 1.

Таблица 1. Подходы к категоризации файлов, их достоинства и недостатки

Группа	Подход	Достоинства	Недостатки
Правила, Шаблоны	IPG	1. Точное описание сложных форматов 2. Хорошо подходит для структурированных файлов	1. Требует знания точной структуры файла 2. Негибкость при изменении формата
	IECore, iASA		
Статистические	TF-IDF	1. Простота реализации 2. Эффективен для выделения ключевых терминов	1. Игнорирует порядок слов 2. Не учитывает семантические связи
	LDA	1. Обнаружение скрытых тем без явного задания правил 2. Эффективен для выявления тематической структуры в коллекции документов	Требует очень больших объемов данных для качественной статистики
Машинное обучение	CRF, HMM	1. Высокая точность в конкретных областях применения 2. Прозрачность и интерпретируемость результатов	Требуют значительных усилий для разработки правил
	BERT	1. Глубокое понимание контекста 2. Высокая точность за счет обучения на больших доменах 3. Универсальность	1. Высокая вычислительная сложность и ресурсоемкость 2. Требует больших объемов данных для тонкой настройки

Предлагаемый алгоритм основан на двухуровневой структуре категорий, которая напрямую решает проблему несоответствия между когнитивными моделями пользователя и структурой файловых систем. Первый уровень определяет общую категорию файла (Работа, Учеба, Финансы и т.д.), второй уровень – уточняющую подкатегорию (Отчеты, Конспекты, Чеки и т.д.).

Данный алгоритм работает с классами-сущностями, которые представляют объекты файловой системы.

1. Сущность *File* (файл) включает в себя поля *Path* (путь, необходим для локализации объекта), *Name* (наименование файла), *Extension* (расширение, создан с целью возможности использования ИИ-модели для анализа), *Tags* (теги).
2. Сущность *Folder* (папка) включает в себя поля *Name* (наименование папки), *Files* (файлы, содержит файлы данной папки), *Subfolders* (подпапки, необходимо для рекурсивного обхода и заполнения структуры папок и файлов).
3. Сущность *Tag* (тег) включает в себя поле *Name* (имя тега).

Схема сущностей системы представлена на рисунке 4.

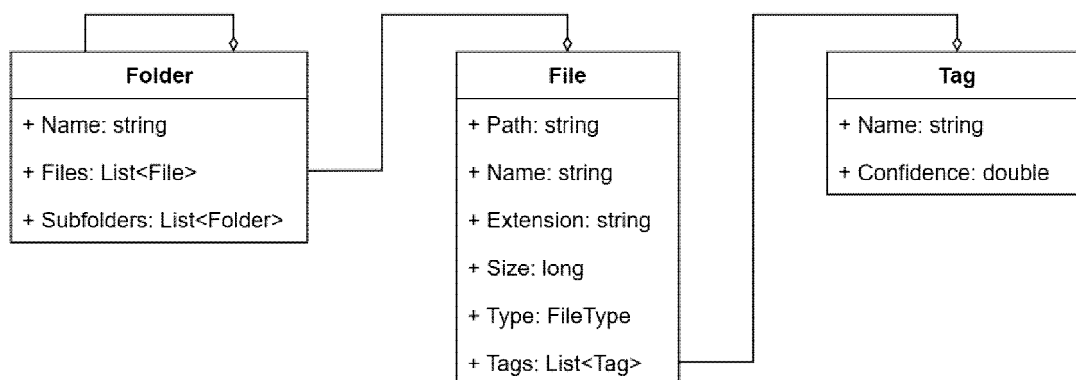


Рис. 4. Схема сущностей системы

Алгоритм включает следующие шаги.

1. Сканирование файловой системы и построение иерархии папок и файлов.
2. Анализ содержимого каждого файла с использованием ИИ-модуля.
3. Определение двухуровневой категории на основе тегов.
4. Формирование новой структуры каталогов.

Блок-схема алгоритма представлена на рисунках 5-6.

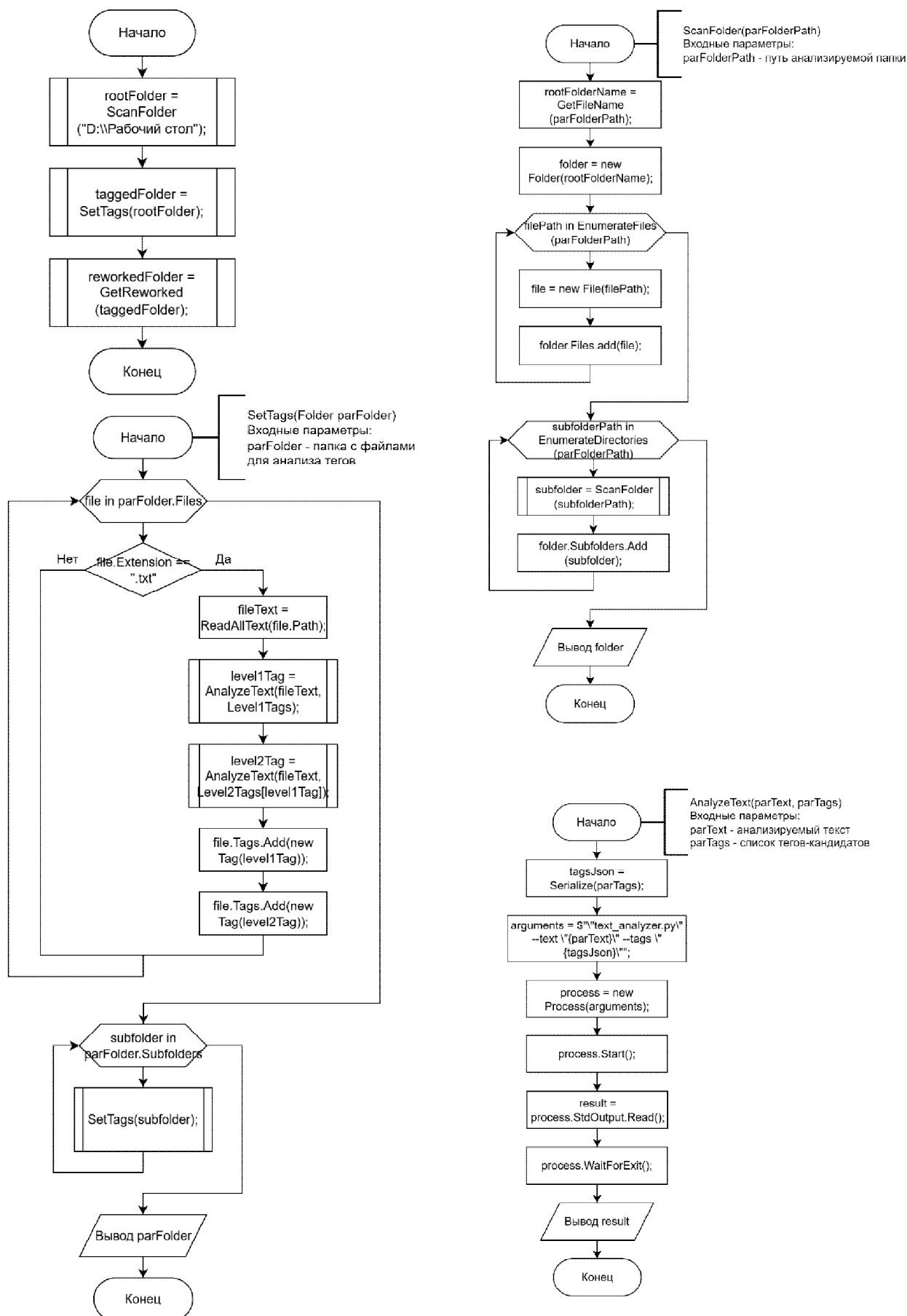


Рис. 5. Блок-схема алгоритма

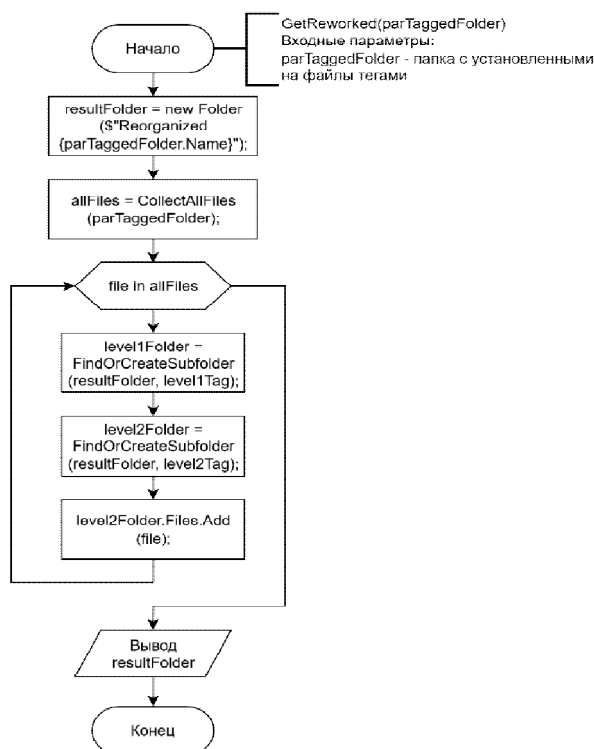


Рис. 6. Блок-схема алгоритма

Практическая значимость работы заключается в преодолении фундаментальной проблемы несоответствия между когнитивными моделями пользователя и структурой файловых систем. Предложенный алгоритм позволяет эффективно бороться с явлением «цифрового накопительства». Ключевые преимущества предлагаемого решения.

1. Двухуровневая система тегов соответствует когнитивным моделям пользователей при организации информации.
2. Полная автоматизация процесса категоризации.

Библиографический список

1. Крталич М., Диннин Дж.Д., Лью К.Л., Гоулдинг А. Личные коллекции и управление личной информацией в семейном контексте [Электронный ресурс] // Library Trends. – 2021. – С. 149–179. – URL: https://www.researchgate.net/publication/357873164_Personal_Collections_and_Personal_Information_Management_in_the_Family_Context (дата обращения: 22.12.2025)
2. Диннин Дж.Д. Насколько велики компьютерные файлы людей? Распределение размеров файлов среди коллекций, управляемых пользователями [Электронный ресурс] // Ежегодная конференция ASIS&T. – 2021. – URL: <https://arxiv.org/abs/2107.03272> (дата обращения: 22.12.2025)
3. Уиттекер С. Настройка и управление личной информацией: как наши чувства влияют на то, как мы организуем нашу информацию [Электронный ресурс] / С. Уиттекер [и др.] // Personal and Ubiquitous Computing. – 2020. – URL: https://www.researchgate.net/publication/342129962_Mood_and_personal_information_management_how_we_feel_influences_how_we_organize_our_information (дата обращения: 24.12.2025)
4. Диннин Дж.Д., Джулиен К.–А. Цифровые файлы: обзор исследований управления файлами [Электронный ресурс]. – 2020. – URL: <https://arxiv.org/abs/2109.09668> (дата обращения: 22.12.2025)
5. Чжан Дж. Грамматика интервального разбора для спецификации формата файла [Электронный ресурс]: магистерская диссертация / Дж. Чжан. – Университет штата Пенсильвания. – 2021. – URL: <https://arxiv.org/abs/2304.04859> (дата обращения: 24.12.2025)
6. Чоудхури Г. Обработка естественного языка [Электронный ресурс] // Annual Review of Information Science and Technology. – 2003. – С. 51–89. URL: https://www.researchgate.net/publication/220142001_Natural_language_processing (дата обращения: 24.12.2025)
7. Коротеев М.В. BERT: Обзор применений в обработке и понимании естественного языка [Электронный ресурс] // М.В. Коротеев. – 2021. – Финансовый университет при Правительстве РФ, Москва. – URL: <https://arxiv.org/abs/2103.11943> (дата обращения: 24.12.2025)

УДК 004.89; ГРНТИ 28.23.29

НЕЙРОСЕТЕВОЙ АНАЛИЗ РИСКОВ В УСЛОВИЯХ СТРОИТЕЛЬСТВА ОБЪЕКТОВ СПЕЦИАЛЬНОГО НАЗНАЧЕНИЯ

С.С. Антипов

Воронежский государственный технический университет, Россия, Воронеж, sergo_vzp@mail.ru,

Аннотация. Современные проекты строительства объектов специального назначения характеризуются повышенной сложностью и уникальностью требований к качеству исполнения. Эти объекты требуют учета множества переменных, начиная от инженерно-геологических особенностей участка и заканчивая технологическими аспектами монтажа и дальнейшей эксплуатации сооружения. Одним из эффективных инструментов для оценки и снижения рисков в процессе проектирования и строительства выступает метод нейросетевых моделей, среди которых сети ART представляют особый интерес благодаря своей способности адаптироваться к новым данным и обеспечивать высокую точность прогнозирования.

Ключевые слова: нейронные сети типа ART, риски, процессы принятия решений, мониторинг данных, обучение нейронной сети.

NEURAL NETWORK RISK ANALYSIS IN SPECIAL-PURPOSE STRUCTURE CONSTRUCTION

S.S. Antipov

Voronezh State Technical University, Russia, Voronezh, sergo_vzp@mail.ru,

Abstract. Modern special-purpose construction projects are characterized by increased complexity and unique quality requirements. These projects require consideration of numerous variables, ranging from the site's engineering and geological features to the technological aspects of installation and subsequent operation of the structure. One effective tool for assessing and mitigating risks during the design and construction process is neural network models, among which ART networks are particularly interesting due to their ability to adapt to new data and ensure high forecasting accuracy.

Keywords: ART-type neural networks, risks, decision-making processes, data monitoring, neural network training.

Строительство объектов специального назначения представляет собой сложный процесс, связанный с множеством факторов риска. К таким объектам относятся военные базы, секретные лаборатории, государственные учреждения и прочие сооружения, имеющие важное значение для национальной безопасности и экономики государства. Для эффективного управления рисками необходимо применение современных методов анализа и моделирования, одним из которых является нейросетевой анализ. Такие системы способны анализировать исторические данные, учитывать экспертные оценки и применять методы машинного обучения для повышения точности предсказания возможных сценариев развития событий.

Требования к таким объектам регламентируются рядом нормативных документов федерального уровня, ГОСТов и СНиПов. Среди ключевых требований выделяются:

1. Надёжность конструкций: устойчивость к экстремальным воздействиям, включая землетрясения, наводнения, пожары и аварии технического характера.

2. Безопасность персонала и окружающей среды: обязательные меры предосторожности, оснащение средствами индивидуальной защиты и оборудованием для мониторинга состояния окружающей среды.

3. Экологическая совместимость: минимизация негативного влияния на окружающую среду путем внедрения современных технологий переработки отходов и ресурсосберегающих методов строительства.

4. Комплексный подход к обеспечению устойчивости: обеспечение надежности инженерных сетей, наличие резервных энергетических мощностей, запас прочности материалов и конструктивных элементов.

5. Автоматизированные системы контроля: внедрение специализированных автоматизированных систем мониторинга, диагностики и предупреждения чрезвычайных ситуаций.

6. Долговечность эксплуатации: расчет срока службы сооружений и регулярное проведение планового ремонта и модернизации оборудования.

Чтобы обеспечить соблюдение этих требований и минимизировать риски на каждом этапе строительства, используется комплексный подход, одним из компонентов которого является нейросетевой анализ рисков [1]. Этот метод позволяет автоматизировать оценку различных сценариев, предвидеть возможные трудности и выработать эффективные механизмы предупреждения негативных последствий.

Основные этапы нейросетевого анализа рисков:

Этап 1: *Сбор и подготовка данных*. Сбор данных включает в себя сбор информации о предыдущих проектах строительства аналогичных объектов, климатических условиях региона, геологических особенностях местности, технологических характеристиках используемых материалов и оборудования. Данные также включают информацию о возможных угрозах, таких как природные катаклизмы, террористические атаки, техногенные аварии и другие факторы.

Подготовка данных заключается в очистке, нормализации и преобразовании исходных данных в форматы, пригодные для обработки нейронными сетями [2].

Этап 2: *Разработка модели нейронной сети*. Разработка модели включает выбор архитектуры нейронной сети, определение количества слоев и нейронов, настройку функций активации и алгоритмов обучения. Выбор подходящей архитектуры зависит от характера данных и целей исследования. Одним из эффективных решений является реализация нейронных сетей типа ART.

Модель ARTMAP:

а) ART-модуль (Адаптивный резонансный метод):

- Входной вектор: $X = (x_1, x_2, \dots, x_n)$ - набор характеристик риска (например, величина, вероятность, воздействие)
- Весовой вектор: $W = (w_1, w_2, \dots, w_n)$ - набор весов, ассоциированных с каждой характеристикой риска.
- Порог активации: θ - минимальный уровень сходства между входным вектором и весовым вектором для активации категории.
- Функция сходства: $S = (X \cdot W) / (\|X\| \|W\|)$ - скалярное произведение, нормализованное по величине векторов.
- Условие резонанса: $S > \theta$ - если сходство превышает порог, входной вектор классифицируется в соответствующую категорию.
- Обучение:
 - Если $S > \theta$, то весовой вектор W настраивается в направлении входного вектора X .
 - Если $S < \theta$, то создается новый весовой вектор W для новой категории.

б) Модуль обратного распространения:

- Вход: Вектор категории (созданный ART-модулем), представляющий тип риска.
- Выход: Прогноз величины риска (или вероятность наступления риска).
- Обучение: Обучение модуля осуществляется на основе известных примеров (входных векторов и соответствующих им значений рисков).
- Функция ошибки: Среднеквадратическая ошибка (MSE) или другая подходящая функция

Этап 3: *Обучение и тестирование модели.* Обучение модели проводится на подготовленных данных с использованием методов машинного обучения. Тестирование модели осуществляется путем сравнения результатов прогнозирования с фактическими значениями [3].

Этап 4: *Анализ результатов и принятие решений.* Анализ результатов позволяет выявить наиболее вероятные сценарии развития событий и определить меры по снижению рисков. Принятие решений основывается на результатах анализа и учитывает экономические, социальные и экологические последствия.

Параметры воздействия для минимизации рисков

Для минимизации рисков при строительстве объектов специального назначения рекомендуется учитывать следующие параметры:

1. Геологические условия: проведение детальных исследований грунта и подземных вод перед началом строительства.
2. Климатические условия: учет особенностей климата региона, включая температуру воздуха, осадки, ветровые нагрузки и солнечную активность.
3. Материалы и технологии: использование высококачественных строительных материалов и технологий, обеспечивающих надежность и долговечность сооружений.
4. Организация работ: четкое соблюдение сроков и стандартов качества, контроль над выполнением всех этапов строительства.
5. Безопасность персонала: обеспечение безопасности рабочих мест, обучение сотрудников правилам техники безопасности и пожарной охраны.
6. Финансовое планирование: тщательное финансовое планирование проекта, резервирование финансовых ресурсов на случай непредвиденных обстоятельств.
7. Экологическая безопасность: соблюдение норм экологической безопасности, предотвращение загрязнения окружающей среды и рациональное использование природных ресурсов.
8. Информационная безопасность: защита конфиденциальной информации о проекте, предотвращение утечек данных и кибератак.

Таким образом, применение нейросетевых технологий повышает эффективность управления крупными инфраструктурными проектами, снижая уровень неопределённости и улучшая оценку рисков. Это способствует успешной реализации проектов, повышению надёжности и безопасности строящихся объектов. Дальнейшее развитие нейросетевых моделей связано с ростом вычислительных ресурсов и доступностью больших объёмов данных, что сделает процессы анализа рисков ещё эффективнее и повысит конкурентоспособность компаний строительной сферы.

Работа выполнена под научным руководством доктора технических наук, профессора В.Л. Бурковского

Библиографический список

1. Иванов, А.Л. Нейронные сети в оценке технических рисков строительного производства / А.Л. Иванов. – СПб.: Спецлит, 2020. – 231 с.
2. Фролов, А.И. Инновационные подходы к оцениванию рисков строительства технически сложных сооружений с использованием нейросетевых технологий / А.И. Фролов. – Нижний Новгород: Нижегородский гос. универ., 2021. – 208 с.
3. Иванов, П.Е. Модели нейросетевого анализа рисков строительства объектов особого назначения / П.Е. Иванов. – М.: Инфра-М, 2022. – 178 с.

УДК 004.8; ГРНТИ 28.23.15

О ВОЗМОЖНОСТЯХ ПРИМЕНЕНИЯ СИСТЕМ ПРОГНОЗИРОВАНИЯ КАРЬЕРНОГО РОСТА И ВЫБОРА ТРАЕКТОРИИ ПРОФЕССИОНАЛЬНОГО РАЗВИТИЯ

Д.А. Буняшин, Е.Н. Проказникова

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, d.bunyashin@yandex.ru

Аннотация. В статье рассматриваются подходы к созданию систем прогнозирования карьерного роста в условиях цифровой трансформации рынка труда. Актуальность исследования обусловлена разрывом между профессиональным образованием и потребностями работодателей, а также информационной асимметрией между соискателями и компаниями. Теоретическую основу составляет компетентностный подход, эмпирическую базу – данные онлайн-платформ вакансий, репрезентативность которых подтверждена корреляцией с официальной статистикой. В статье систематизированы современные методы анализа текстовой информации, позволяющие автоматизированно выделять компетенции, проведена типология существующих систем карьерного планирования, сформулированы требования к перспективной интегрированной системе. Определены ключевые вызовы и направления дальнейшего развития таких систем.

Ключевые слова: прогнозирование карьерного роста, рынок труда, компетенции, большие данные, обработка естественного языка, нейросетевые модели, информационная асимметрия, типология систем, карьерное планирование, анализ вакансий.

ABOUT THE POSSIBILITIES OF USING SYSTEMS CAREER PREDICTION AND CAREER TRAJECTORY SELECTION

D.A. Bunyashin, E.N. Prokaznikova

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, d.bunyashin@yandex.ru

The summary. The article examines approaches to creating career forecasting systems in the context of the digital transformation of the labor market. The relevance of the study is due to the gap between vocational education and the needs of employers, as well as the information asymmetry between applicants and companies. The theoretical basis is the competence approach, the empirical base is the data of online vacancy platforms, the representativeness of which is confirmed by correlation with official statistics. The article systematizes modern methods of text information analysis, allowing to automatically allocate competencies, a typology of existing career planning systems is carried out, requirements for a promising integrated system are formulated. Key challenges and directions for the further development of such systems have been identified.

Keywords: career forecasting, labor market, competencies, big data, natural language processing, neural network models, information asymmetry, system typology, career planning, vacancy analysis.

В условиях цифровизации экономики и трансформации социально-трудовых отношений рынок труда претерпевает фундаментальные изменения, формирующие запрос на оперативный мониторинг кадровой ситуации [1]. Традиционные инструменты анализа, опирающиеся на официальную статистику, обладают недостаточной оперативностью и не позволяют улавливать структурные сдвиги на уровне отдельных компетенций [1, 3].

Ключевой проблемой становится углубляющийся разрыв между системой профессионального образования и реальными потребностями работодателей, фиксируемый исследованиями как устойчивая информационная асимметрия [3]. Особенно остро данная проблема проявляется в высокотехнологичных отраслях, где дефицит квалифицированных кадров достигает сотен тысяч человек, а опросы специалистов подтверждают несоответствие академических программ рыночным требованиям [6]. Региональные исследования также демонстрируют критические диспропорции, когда количество выпускников по перспективным направлениям покрывает лишь незначительную долю от потребности предприятий [4].

В связи с этим особую актуальность приобретает разработка систем прогнозирования карьерного роста, интегрирующих анализ больших данных с онлайн-платформ вакансий и методы машинного обучения для выявления перспективных компетенций [1, 5, 7]. Существующие решения либо ориентированы исключительно на аналитику рынка, либо предоставляют общие рекомендации без учета прогностического потенциала [6, 8]. Целью настоящей статьи является систематизация современных подходов к созданию интеллектуальных систем поддержки карьерного планирования и определение принципов формирования адаптивных траекторий профессионального развития специалистов.

Теоретические и методологические основы прогнозирования карьеры

Традиционный анализ рынка труда, опирающийся на макроэкономические показатели и отраслевую статистику, демонстрирует ограниченную прогностическую ценность вследствие ориентации на формальные категории профессий, не успевающие отражать изменения в содержании трудовой деятельности [1, 3]. В ответ на этот вызов в научный дискурс внедряется компетентностный подход, перемещающий фокус анализа на набор знаний, умений и навыков, необходимых для выполнения конкретных трудовых функций [2]. Поскольку разработка профессиональных стандартов существенно отстает от динамики технологического развития, особое значение приобретает эмпирический анализ актуальных требований работодателей, извлекаемых непосредственно из текстов вакансий [1, 6].

Современный рынок труда функционирует в условиях «ограниченной рациональности», когда ни одна из сторон не обладает полной информацией о контрагенте [3]. Работодатели склонны к обобщенным формулировкам, смешивая требования к компетенциям с описанием должностных обязанностей, тогда как соискатели гипертрофированно детализируют приобретенный опыт, что формирует устойчивый дисбаланс спроса и предложения [3, 8]. Преодоление этого дисбаланса требует перехода к аналитике на уровне отдельных компетенций, что делает необходимым применение методов автоматизированного извлечения знаний из неструктурированных текстов.

Эмпирической базой для такого анализа выступают открытые данные онлайн-платформ, демонстрирующие высокую степень корреляции с официальной статистикой как по динамике вакансий, так и по уровню заработных плат [1]. Методологический инструментарий обработки таких массивов включает комплекс методов интеллектуального анализа текстов: от эвристических алгоритмов до нейросетевых моделей, дообученных на корпусе вакансий для выявления семантических связей между компетенциями [5, 7, 8].

Современные методы анализа текстовой информации для рынка труда

Традиционные методы извлечения информации из текстов вакансий, основанные на регулярных выражениях или поиске по ключевым словам, демонстрируют низкую эффективность при работе со сложными языковыми конструкциями [8]. Преодоление этого ограничения требует применения методов обработки естественного языка, ядром которых в современной практике становятся нейросетевые модели на архитектуре трансформеров [5, 8].

Векторизация текстов позволяет перевести семантику описаний в математическое пространство для последующих вычислений. Модели типа word2vec дают возможность выявлять семантическую близость между компетенциями и проводить кластеризацию вакансий для выделения устойчивых групп специальностей [5]. Задача непосредственного выделения компетенций решается методами извлечения именованных сущностей, которые классифицируют фрагменты текста по категориям. Обученные на размеченных данных модели NER значительно превосходят эвристические методы, достигая значений F1-меры до 0.65 [7]. Дообучение языковых моделей на специализированных корпусах текстов вакансий позволяет адаптировать их к специфике профессиональной лексики [8].

Развитием методов анализа становится использование синтаксических анализаторов в комбинации с семантическими моделями для обработки сложносочиненных требований. Построение деревьев синтаксического разбора с последующим дополнением их новыми семантическими связями, выявленными дообученной моделью BERT, позволяет извлекать отдельные знания и навыки из длинных описаний [8]. Комплексное применение методов векторизации, распознавания именованных сущностей и синтаксического анализа формирует технологический базис для систем, оценивающих структуру спроса на рынке труда на уровне отдельных компетенций [5, 7, 8].

Типология и анализ существующих систем карьерного планирования

Анализ существующих разработок позволяет выделить несколько устойчивых типов систем карьерного планирования, различающихся по функциональному назначению и глубине автоматизации обработки данных [6]. Первый тип составляют информационно-аналитические системы, ориентированные на сбор и обработку больших данных о рынке труда. Примером выступает система «РосНавык», осуществляющая мониторинг вакансий и оценку востребованности профессиональных навыков на основе методов анализа больших данных. Достоинством таких систем является высокая актуальность информации и возможность отслеживания динамики рыночных изменений, однако их функционал ограничен аналитикой и не предполагает выработку персонализированных рекомендаций для пользователей [1, 4, 5, 6].

Второй тип – рекомендательные системы, фокусирующиеся на формировании индивидуальных карьерных траекторий и планов обучения. Платформы, подобные Roadmap.sh, предоставляют пользователям визуализированные маршруты освоения профессий, однако качество таких рекомендаций часто зависит от активности сообщества, а их связь с актуальной рыночной аналитикой остается слабой [6]. Третий тип представлен коммуникативными платформами, реализующими функцию наставничества и расширения профессиональных связей. Сервисы типа GetMentor позволяют устанавливать контакты между специалистами и наставниками, но их эффективность напрямую зависит от размера пользовательской базы и отсутствия механизмов верификации качества консультаций [6].

Наиболее перспективным направлением выступают интегрированные системы, объединяющие функциональные возможности всех перечисленных типов. Платформа NUS Career+, разработанная Национальным университетом Сингапура, демонстрирует пример такой интеграции, сочетая персонализированные рекомендации на основе профиля пользователя, аналитику рыночного спроса на компетенции и функции сетевого взаимодействия [6]. Опыт эксплуатации подобных систем подтверждает, что именно комплексный подход, объединяющий аналитику рынка труда, прогнозирование его изменений и индивидуализацию рекомендаций, способен обеспечить эффективную поддержку карьерного планирования [6, 8]. Анализ сильных и слабых сторон существующих решений позволяет сформулировать требования к разработке новых систем, интегрирующих передовые методы обработки естественного языка и машинного обучения для формирования адаптивных профессиональных траекторий [5, 7, 8].

Проектирование перспективной системы прогнозирования карьерного роста

Интеграция результатов анализа существующих решений и современных методов обработки текстовой информации позволяет сформулировать концептуальные требования к перспективной системе прогнозирования карьерного роста. Ключевым принципом проектирования выступает объединение функциональных возможностей, реализованных разрозненно: персонализированные рекомендации, аналитика рынка труда и прогнозирование его изменений должны быть интегрированы в единую платформу, что способно обеспечить преодоление информационной асимметрии между соискателями и работодателями [3, 6, 8].

Функциональная архитектура системы должна включать непрерывный сбор данных с онлайн-платформ вакансий, их очистку и структурирование с применением методов обработки естественного языка [1, 7]. На основе дообученных нейросетевых моделей осуществляется извлечение и нормализация компетенций, выявление семантических связей между ними с построением графовой базы знаний [8]. Кластеризация полученных данных позволяет группировать вакансии по профессиональным направлениям и отслеживать динамику востребованности навыков, включая прогнозирование изменений спроса методами анализа временных рядов [1, 5].

Критически важным компонентом выступает механизм персонализации, учитывающий текущий уровень компетенций пользователя, его профессиональные интересы и региональную специфику рынка труда [4, 6]. На основе сопоставления профиля пользователя с рыночными данными формируется индивидуализированная дорожная карта развития, включающая последовательность освоения компетенций с учетом их востребованности и логических связей [6, 8]. Техническая реализация требует применения методов машинного обучения для ранжирования рекомендаций и адаптации траекторий при изменении рыночной ситуации [1, 5]. Результатом становится не только повышение эффективности индивидуального карьерного планирования, но и создание основы для сокращения разрыва между образовательными программами и потребностями экономики [2, 6, 8].

Внедрение систем прогнозирования карьерного роста сопряжено с вызовами: низкое качество исходных данных, высокая ресурсоемкость разработки и проблема интерпретируемости результатов моделей машинного обучения [1, 3, 5, 6]. Фрагментарность существующих решений ограничивает реализацию комплексного прогностического функционала [6].

Перспективные направления связаны с созданием единой онтологии компетенций, интеграцией с образовательными платформами, учетом региональной специфики и применением больших языковых моделей для повышения прозрачности рекомендаций [2, 4, 7].

Таким образом, современные системы прогнозирования должны интегрировать анализ больших данных, обработку естественного языка и машинное обучение на базе репрезентативных данных онлайн-вакансий. Компетентностный подход позволяет перейти от формального анализа профессий к содержательному анализу спроса, что выступает необходимым условием преодоления информационной асимметрии и повышения эффективности рынка труда [1, 2, 5, 8].

Библиографический список

1. Алтухов, В.В. Использование открытых данных онлайн-вакансий в сравнении с данными официальной статистики для мониторинга и прогнозирования динамики рынка труда / В.В. Алтухов, А.Д. Кудрявцев // Уровень жизни населения регионов России. – 2025. – Т. 21, № 2. – С. 233–244.
2. Аксеньюшкина, Е.В. Анализ востребованных компетенций выпускников вузов / Е.В. Аксеньюшкина, Н.В. Мамонова // Вестник Бурятского государственного университета. Экономика и менеджмент. – 2024. – № 1. – С. 3–12.
3. Сизова, И.Л. Особенности подбора персонала: интеллектуальный анализ текстов резюме и вакансий / И.Л. Сизова // Регионология. – 2025. – Т. 33, № 2. – С. 271–293.
4. Соколов, В.А. Анализ рынка труда специалистов электроэнергетики в городе Владимир / В.А. Соколов // Вестник Владимирского государственного университета. – 2024. – № 1. – С. 1–6.
5. Снегирев, А.С. Подход к кластеризации текстов описания вакансий для анализа требований к специальностям в ИТ / А.С. Снегирев // Научно-технические ведомости СПбГПУ. – 2020. – № 3. – С. 45–56.
6. Клоков, С.А. Применение ИТ-решений для карьерного планирования / С.А. Клоков, А.А. Бирюкова // Инженерные технологии. – 2025. – № 1(9). – С. 79–88.
7. Корытов, П.В. Анализ подходов к извлечению ключевых навыков из вакансий / П.В. Корытов, Я.Ю. Грибецкий, Е.А. Андреева, И.И. Холод // Труды XXVII Международной конференции по мягким вычислениям и измерениям (SCM 2024). – 2024. – С. 300–303.
8. Николаев, И.Е. Метод извлечения знаний и навыков/компетенций из текстов требований вакансий / И.Е. Николаев // Онтология проектирования. – 2023. – Т. 13, № 2. – С. 282–293.

УДК 004.032;26; ГРНТИ 28.23.15

ПРИМЕНЕНИЕ МОДЕЛЕЙ ЗРИТЕЛЬНОГО ВНИМАНИЯ ДЛЯ ЗАДАЧИ АВТОМАТИЧЕСКОЙ ОЦЕНКИ КАЧЕСТВА ИЗОБРАЖЕНИЙ

Ю.В. Ярославцев

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, yury.yarov@yandex.ru*

Аннотация. В работе рассматривается метод безэталонной оценки технического качества изображения, основанный на пространственном взвешивании метрик по карте зрительного внимания. Актуальность работы обусловлена тем, что существующие методы слепой оценки качества изображений оперируют глобальными статистиками кадра и не различают перцептивно значимые области и фон, что приводит к систематической чувствительности баллов к художественным приёмам съёмки, не связанным с техническим качеством объекта. Предлагаемый метод использует предобученную нейросетевую модель зрительного внимания для формирования пространственного распределения весов, которое заменяет равномерное среднее в формулах технических метрик качества изображения.

Ключевые слова: анализ изображений, метрики качества изображения, оценка качества изображения без эталона, зрительное внимание.

Активный рост объемов цифровой фотографии в настоящее время делает востребованными надежные и быстрые методы автоматической оценки качества изображений. Современные безэталонные методики демонстрируют устойчивые результаты, особенно применяющие алгоритмы глубокого обучения на основе нейронных сетей, однако они имеют ряд серьезных ограничений.

Воспринимаемое человеком качество изображения определяется не только общими физическими свойствами изображения, но и психофизическими механизмами обработки визуальной информации в зрительной системе. Ключевой особенностью является направленность внимания, выражающаяся в повышенной концентрации взгляда на определенных участках изображения, содержащих семантически важные элементы [3]. Качество именно этих областей определяет значительную часть общего впечатления. Современные модели предсказания зрительного внимания (DeepGaze, VST) позволяют довольно точно определять такие значимые области [1, 2].

Существующие методы безэталонной оценки качества изображений можно разделить на два типа: статистические метрики (BRISQUE, NIQE, PIQE), основанные на анализе отклонений от естественной статистики сцены, и алгоритмы глубокого обучения (MUSIQ, HyperIQA, CNNIQA), предсказывающие воспринимаемое качество в результате обучения на большом количестве субъективных оценок.

Статистические метрики используют низкоуровневые характеристики изображения, такие как яркость, контраст, резкость, градиенты и текстурные признаки. Основой их работы является идея, что неискаженные изображения подчиняются определенным статистическим закономерностям «естественной статистики сцены» (NSS), а искаженные нарушают их. Эти методы, как правило, вычислительно просты, по сравнению с моделями глубокого обучения. Общим недостатком статистических метрик является глобальная агрегация признаков: характеристики усредняются по всему изображению и пространственная неоднородность человеческого восприятия не учитывается.

Методы безэталонной оценки качества изображений на основе глубокого обучения используют сверточные архитектуры и трансформеры для автоматического извлечения признаков из изображений и построения моделей, предсказывающих воспринимаемое в результате обучения на больших наборах данных, содержащих субъективные оценки качества [4]. Эти методы показывают значительно более высокую точность по сравнению с классическими моделями, обеспечивая высокую корреляцию с человеческими оценками. Их главным недостатком является сложность в вычислительном плане, усложняющая их применение в реальном времени. Как и большинство глубоких нейросетевых систем, модели без-

эталонной оценки изображений обладают низкой интерпретируемостью – механизм их решений сложно объяснить, вызывая дополнительные трудности при внедрении.

Решением проблем множества существующих методов безэталонной оценки качества изображений, связанных с недостатком учета распределения зрительного внимания человека или высокой вычислительной сложностью, является явное использование карты зрительного внимания человека в качестве пространственно весового механизма при агрегации локальных технических метрик, обеспечивая интерпретируемый и вычислительно эффективный учёт избирательности зрительного восприятия.

В качестве модели предсказания карты вероятностного распределения зрительного внимания рассматривается DeepGaze. В отличие от полностью сверточных архитектур, она использует признаки предобученных классификационных сетей без их дообучения, надстраивая над ними вероятностную модель плотности фиксации [1]. Модель DeepGaze ПЕ дополнительно учитывает центральное смещение и пространственные закономерности человеческого восприятия, демонстрируя высокую точность предсказания фиксации человеческого зрения при относительно низких вычислительных требованиях.

В данной работе предложен алгоритм, в котором карта зрительного внимания, предсказываемая моделью DeepGaze ПЕ, выступает пространственным весовым механизмом при агрегации локальных технических характеристик изображения: резкости, экспозиции и зашумленности. Вклад каждого участка изображения в итоговую оценку пропорционален его визуальной значимости.

Алгоритм состоит из следующих этапов.

1. Построение карты зрительного внимания с помощью нейросетевой модели и её преобразование в пространственное распределение весов.

2. Вычисление набора метрик качества: резкость, шум, экспозиция, в которых стандартное равномерное среднее заменено взвешенным средним по карте зрительного внимания.

3. Нормализация метрик и их линейная агрегация в итоговый скалярный балл.

Рассмотрим каждый этап детально.

1. Извлечение области интереса на основе карты зрительного внимания

Модель DeepGaze ПЕ принимает на вход цветное изображение и карту центрального смещения (априорное распределение, отражающее тенденцию человека фиксировать взгляд ближе к центру кадра), а возвращает логарифм $\log p(x, y | I, B)$ плотности вероятности фиксации в каждой точке (I – входное изображение, B – карта центрального смещения).

Плотность вероятности получается экспоненцированием:

$$S(x, y) = \exp(\log p(x, y | I, B)).$$

Здесь $S(x, y)$ – карта плотности вероятностного распределения зрительного внимания. Высокие значения указывают на области, к которым с наибольшей вероятностью будет привлечено зрительное внимание человека.

Для ускорения инференса допускается предварительное масштабирование изображения с коэффициентом $s < 1,0$. В этом случае модель работает на уменьшенном разрешении $H' \times W' = [Hs] \times [Ws]$, а результирующая карта возвращается к исходному разрешению $H \times W$ посредством билинейной интерполяции.

Карта $S(x, y)$ преобразуется в карту весов $w(x, y)$, представляющую собой дискретное вероятностное распределение. Это необходимо для того, чтобы взвешенные метрики были инвариантны к разрешению изображения: без нормализации более крупное изображение систематически давало бы более высокие абсолютные значения.

2. Взвешенные метрики качества

Ключевая идея метода состоит в замене стандартного равномерного среднего на взвешенное среднее во всех метриках качества. Если $f(x, y)$ – локальная характеристика пикселя, то:

$$\mu = \frac{1}{HW} \sum_{x,y} f(x, y) \longrightarrow \mu_w = \sum_{x,y} f(x, y) \cdot w(x, y),$$

где $w(x, y)$ – карта весов, HW – общее число пикселей. Поскольку $\sum w = 1$, взвешенное среднее μ_w представляет собой математическое ожидание f относительно распределения w .

Для оценки резкости используются два оператора, реагирующих на различные пространственно-частотные характеристики изображения.

Энергия градиента Собеля. Оператор Собеля – дифференциальный оператор для вычисления приближённого градиента яркости, применяется по обеим осям, а взвешенная энергия градиента вычисляется как:

$$T_w = \sum_{x,y} [(G_x(x, y))^2 + (G_y(x, y))^2] \cdot w(x, y),$$

где $G_x = I_{\text{gray}} \cdot K_x$, $G_y = I_{\text{gray}} \cdot K_y$ – свёртки полутонового изображения с ядрами Собеля размера 3×3 по горизонтали и вертикали соответственно. Более высокое значение T_w указывает на более резкие границы объектов в области интереса.

Взвешенная дисперсия лапласиана. Лапласиан – дифференциальный оператор второго порядка, сумма вторых частных производных; чувствителен к быстрым изменениям яркости, применяется к полутоновому изображению: $\nabla^2 I = I * K_{\nabla^2}$. Дисперсия лапласиана, вычисленная относительно распределения w , служит мерой наличия высокочастотных деталей:

$$L_w = \mathbb{E}_w[(\nabla^2 I)^2] - (\mathbb{E}_w[\nabla^2 I])^2 = \sum_{x,y} (\nabla^2 I)^2 \cdot w - \left(\sum_{x,y} \nabla^2 I \cdot w \right)^2.$$

Метрика L_w принимает высокие значения при наличии резких перепадов яркости и близка к нулю при равномерно размытом изображении.

Шум оценивается в пределах области интереса, ограниченной дополнительным условием: учитываются только плоские зоны – участки изображения с низкой локальной дисперсией, где высокочастотная компонента с высокой вероятностью является шумом, а не текстурой или границей объекта.

Локальная дисперсия вычисляется с помощью прямоугольного ядра K_p размера $p \times p$:

$$\sigma_{\text{loc}}^2(x, y) = (I^2 * K_p)(x, y) - ((I * K_p)(x, y))^2.$$

Маска плоских зон определяется пороговым условием $F(x, y) = \mathbb{I}[\sigma_{\text{loc}}^2(x, y) < \tau_{\text{flat}}]$, после чего формируется ограниченная карта весов:

$$w_{\text{flat}}(x, y) = \frac{w(x, y) \cdot F(x, y)}{\sum_{u,v} w(u, v) \cdot F(u, v)}.$$

Оценка шума вычисляется как взвешенное среднее абсолютного отклонения высокочастотной компоненты:

$$N_w = \sum_{x,y} |I_{\text{gray}}(x,y) - (G_\sigma * I_{\text{gray}})(x,y)| \cdot w_{\text{flat}}(x,y),$$

где $G_\sigma * I_{\text{gray}}$ – результат гауссова сглаживания полутонового изображения. Более высокое значение N_w свидетельствует о большем уровне шума.

Метрики экспозиции оценивают корректность яркостных характеристик в области интереса.

Локальный контраст. Стандартное отклонение яркости в скользящем окне размером 16×16 выступает мерой локальной детализации:

$$C_w = \sum_{x,y} \sigma_{\text{loc}}(x,y) \cdot w(x,y), \text{ где } \sigma_{\text{loc}} = \sqrt{(Y^2 * K_{16}) - (Y * K_{16})^2},$$

где K_{16} – равномерное ядро 16×16 . Более высокий C_w соответствует более выраженной тональной проработке (диапазону яркостей, воспроизводимых в изображении) в области интереса.

Взвешенная энтропия гистограммы яркости. Гистограмма яркости в области интереса строится с учётом весов w : в каждый интервал гистограммы $k \in \{0, \dots, 255\}$ вносится суммарный вес пикселей с яркостью k . Энтропия Шеннона полученной гистограммы:

$$H_w = - \sum_{k=0}^{255} p_k \log_2 p_k, \text{ где } p_k = \frac{\sum_{(x,y):=k} w(x,y)}{\sum_j p_j}.$$

Низкая энтропия H_w свидетельствует о недостаточной тональной проработке (узкий динамический диапазон, потеря деталей в тенях или светах).

3. Агрегация баллов и нормализация

Вычисленные метрики имеют различные физические размерности и диапазоны значений. Для агрегации в единый балл необходимо привести все метрики к общему диапазону $[0,1]$.

При обработке изображения для каждой метрики m_k применяется нелинейная нормализация функцией гиперболического тангенса:

$$\hat{m}_k = \tanh\left(\frac{m_k}{c_k}\right),$$

где c_k – калибровочная константа. Значения c_k подбираются эмпирически на основе типичных характеристик качественных изображений

Итоговый балл качества Q вычисляется как взвешенная линейная комбинация нормализованных метрик:

$$Q = \sum_k \alpha_k \cdot \tilde{m}_k, \quad \sum_k \alpha_k = 1,$$

где $\tilde{m}_k$ – нормализованная метрика, при необходимости инвертированная:

$$\tilde{m}_k = \begin{cases} \hat{m}_k, & \text{если большее значение означает лучшее качество (прямая метрика);} \\ 1 - \hat{m}_k, & \text{если большее значение означает худшее качество (штрафная метрика).} \end{cases}$$

К прямым метрикам относятся: энергия градиента Собеля, взвешенная дисперсия лапласиана, локальный контраст, взвешенная энтропия гистограммы яркости. К штрафным – шум.

Весовые коэффициенты α_k отражают относительную важность каждого аспекта качества.

Итоговый балл масштабируется в диапазон $[0, 100]$:

$$Q_{\text{final}} = Q \times 100.$$

Для проверки гипотезы о преимуществе использования карты вероятностного распределения зрительного внимания в качестве взвешенной оценки качества было проведено сравнение предлагаемого метода с базовым подходом, при котором метрики рассчитывались без карты зрительного внимания, равномерно по всему изображению, и ряда безэталонных метрик оценки качества.

В качестве контролируемого источника вариации использовалась серия из 10 фотографий одной сцены, снятых при различной диафрагме объектива. При неизменном главном объекте съёмки изменение диафрагмы приводит к изменению глубины резкости: при более открытой диафрагме фон размывается сильнее, при закрытой – фон становится более детализированным. Главный объект во всех кадрах оставался в фокусе, поэтому его техническое качество (резкость, шум, экспозиция) менялось незначительно.

Примеры двух изображений с различной диафрагмой из серии и предсказанные карты внимания к ним представлены на рисунках 1, 2.

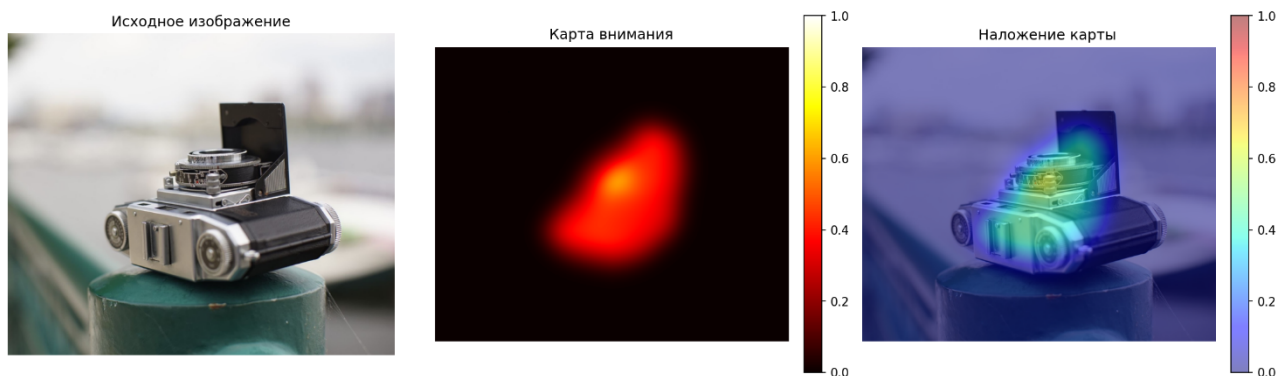


Рис. 2. Изображение с открытой диафрагмой

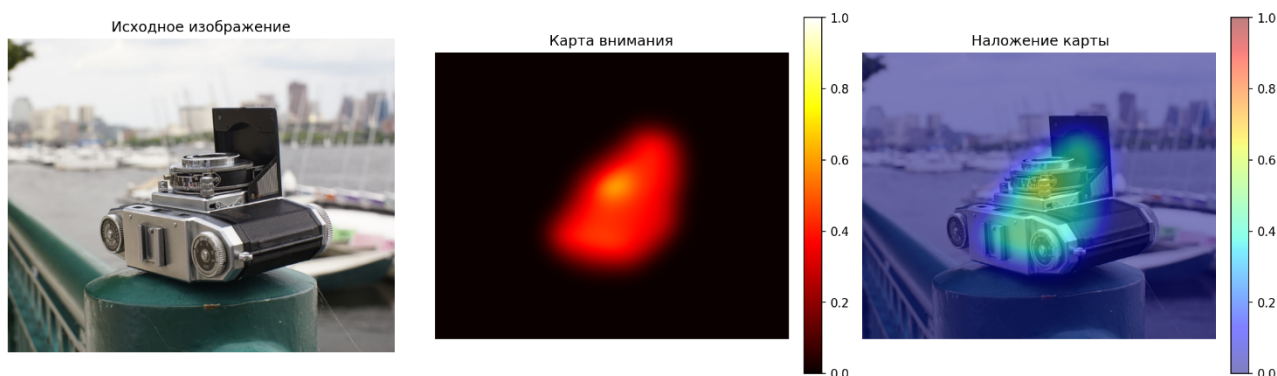


Рис. 3. Изображение с прикрытой диафрагмой

Ожидаемое поведение метрики качества – выдача стабильных баллов по всей серии, поскольку качество субъекта съёмки не менялось.

В эксперименте участвовали: предлагаемый метод, базовый вариант с равномерными весами, BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) – статистическая

метрика, основанная на локальных нормализованных яркостных признаках, NIQE (Natural Image Quality Evaluator) – метрика, опирающаяся на многовариантное гауссово моделирование натуральных сцен, PIQE (Perception-based Image Quality Evaluator) – метрика, использующая локальные статистики блоков, MUSIQ (Multi-Scale Image Quality Transformer) – метрика на основе трансформера, оперирующая мультимасштабными патчами изображения.

В качестве меры стабильности метрики использовался разброс – относительное отклонение минимального балла от максимального в серии: $\Delta = (Q_{\max} - Q_{\min}) / Q_{\max} \times 100\%$. Чем ниже Δ , тем устойчивее метрика к изменению характеристик фона при неизменном качестве главного объекта.

Результаты представлены в таблице 1.

Таблица 6. Относительный разброс баллов качества (ниже - лучше)

Метод	Разброс %
Предлагаемый	4,6
Базовый	15,4
NIQE	14,6
PIQE	15,3
BRISQUE	23,7
MUSIQ	11,2

Предлагаемый метод показал наименьший разброс (4,6%) среди всех сравниваемых метрик и существенно превзошёл базовый вариант с равномерными весами (15,0%). Снижение разброса более чем в три раза подтверждает, что взвешивание по карте зрительного внимания эффективно уменьшает влияние незначимых для зрительного внимания человека областей на итоговый балл.

Классические статистические метрики (BRISQUE, NIQE, PIQE) характеризуются значительно более высоким разбросом (14,6%–23,7%). Причина заключается в том, что они опираются на глобальные статистики изображения и не различают субъект съёмки и фон. Усиление размытия фона при открытой диафрагме приводит к изменению распределения высокочастотных компонент и локальных контрастов по всему кадру, что интерпретируется этими метриками как деградация качества, хотя в данном эксперименте это артефакт не является деградацией с точки зрения зрителя.

Нейросетевая метрика MUSIQ (11,2%) демонстрирует меньший разброс по сравнению с классическими метриками, что может объясняться неявным учётом семантики сцены, заложенным в обучении на большом количестве изображений. Тем не менее разброс остаётся заметно выше, чем у предлагаемого метода.

Сравнение с базовым вариантом демонстрирует, что именно введение карты зрительного внимания, а не сами по себе метрики резкости, шума и экспозиции, обеспечивает устойчивость к изменению фона. Одни и те же формулы, применённые с равномерными весами, дают почти четырёхкратный рост разброса. Это обосновывает целесообразность использования явной модели зрительного внимания для оценки технического качества объекта съёмки в условиях варьирования глубины резкости.

Библиографический список

1. DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling [Электронный ресурс] / A. Linardos, M. Kümmerer, O. Press [et al.] — Электрон. текстовые данные. — 2021. — 15 с. — Режим доступа: <https://arxiv.org/abs/2105.12441> (дата обращения: 09.10.2025). — Текст: электронный.
2. Review of Visual Saliency Prediction: Development Process from Neurobiological Basis to Deep Models [Электронный ресурс] / F. Yan, Ch. Chen, P. Xiao [et al.] // Applied Sciences (Switzerland). – 2022. – Vol. 12, № 1. – P. 309. – DOI 10.3390/app12010309. – 22 с. – Режим доступа: <https://www.elibrary.ru/item.asp?id=74027157> (дата обращения: 11.10.2025). – Текст: электронный.

3. Trends, Applications, and Challenges in Human Attention Modelling [Электронный ресурс] / G. Cartella, M. Cornia, V. Cuculo [et al.] – Электрон. текстовые данные. – 2024. – 9 с. – Режим доступа: <https://arxiv.org/abs/2402.18673> (дата обращения: 12.10.2025). – Текст: электронный.

4. Рябов Г.А. Повышение качества анализа изображений с использованием сверточных нейронных сетей [Электронный ресурс] / Г.А. Рябов, Е.В. Кривоногова, Д.Ю. Изотов, Н.И. Абрамова // Тенденции развития науки и образования. – 2023. – № 99-7. – С. 131–136. – DOI 10.18411/trnio-07-2023-399. – EDN UFQXHE. – 6 с. – Режим доступа: <https://www.elibrary.ru/item.asp?id=54309569> (дата обращения: 15.10.2025). – Текст: электронный.

УДК 004.891; ГРНТИ 28.23.29

ИСПОЛЬЗОВАНИЕ МЕХАНИЗМОВ ВНИМАНИЯ В ТРАНСФОРМЕРНЫХ МОДЕЛЯХ ДЛЯ ПЕРСОНАЛИЗИРОВАННЫХ НОВОСТНЫХ РЕКОМЕНДАЦИЙ

П.В. Журавлев, Т.А. Дмитриева, Е.Н. Проказникова

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, GPV2002@mail.ru*

Аннотация. В работе рассматривается проблема повышения точности персонализированных рекомендаций новостей при неоднородных интересах пользователей и быстром обновлении контента. Предлагается использовать трансформерные архитектуры с механизмом self-attention для моделирования зависимостей в истории взаимодействий пользователя. Описывается роль multi-head attention в выделении разных факторов релевантности (тематика, семантика, контекст поведения). Показано, что применение внимания улучшает качество ранжирования за счёт адаптивного взвешивания событий в пользовательской последовательности. Обсуждаются практические аспекты внедрения в промышленных рекомендательных системах и требования к масштабируемости. Сформулированы ограничения подхода (холодный старт, вычислительные затраты, риск фильтра пузырей) и перспективы развития.

Ключевые слова: рекомендательные системы, трансформерная архитектура, механизм внимания, self-attention, multi-head attention, персонализация.

USING ATTENTION MECHANISMS IN TRANSFORMER MODELS FOR PERSONALIZED NEWS RECOMMENDATIONS

P.V. Zhuravlev, T.A. Dmitrieva, E.N. Prokaznikova

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, GPV2002@mail.ru*

The summary. The paper addresses the problem of improving the accuracy of personalized news recommendations in the case of heterogeneous user interests and rapid content updates. It proposes the use of transformer architectures with self-attention mechanism to model dependencies in the user's interaction history. The paper describes the role of multi-head attention in highlighting different relevance factors (topic, semantics, and behavioral context). It shows that the use of attention improves the quality of ranking by adaptively weighting events in the user's sequence. Practical aspects of implementation in industrial recommendation systems and requirements for scalability are discussed. Limitations of the approach (cold start, computational costs, risk of bubble filter) and prospects for development are formulated.

Keywords: recommendation systems, transformer architecture, attention mechanism, self-attention, multi-head attention, and personalization.

Персонализированные ленты новостей в социальных сетях и медиа-платформах стали критически важным компонентом цифрового контента, определяющим информационный опыт миллиардов пользователей. При экспоненциальном росте объёма информационных потоков традиционные методы рекомендации контента, такие как коллаборативная фильтрация, достигли своих пределов в точности и адаптивности. Рекомендательные трансформеры и механизмы внимания (attention mechanisms [2]) представляют собой пара-

дигмальный сдвиг в этой области, позволяя системам моделировать сложные взаимодействия между пользователями и новостным контентом с беспрецедентной точностью [1]. Целью данной статьи является анализ роли механизмов внимания в трансформерных архитектурах для задачи персонализированной рекомендации новостей, а также обзор их практического применения в современных системах.

Механизм Self-Attention в контексте рекомендаций новостей

Self-attention механизм, введенный в 2017 году в работе Vaswani et al. «Attention Is All You Need», революционизировал обработку последовательностей в нейронных сетях. В контексте рекомендаций новостей, self-attention позволяет модели анализировать историю взаимодействий пользователя с контентом, определяя, какие предыдущие действия наиболее релевантны для предсказания текущих предпочтений (рис. 1).

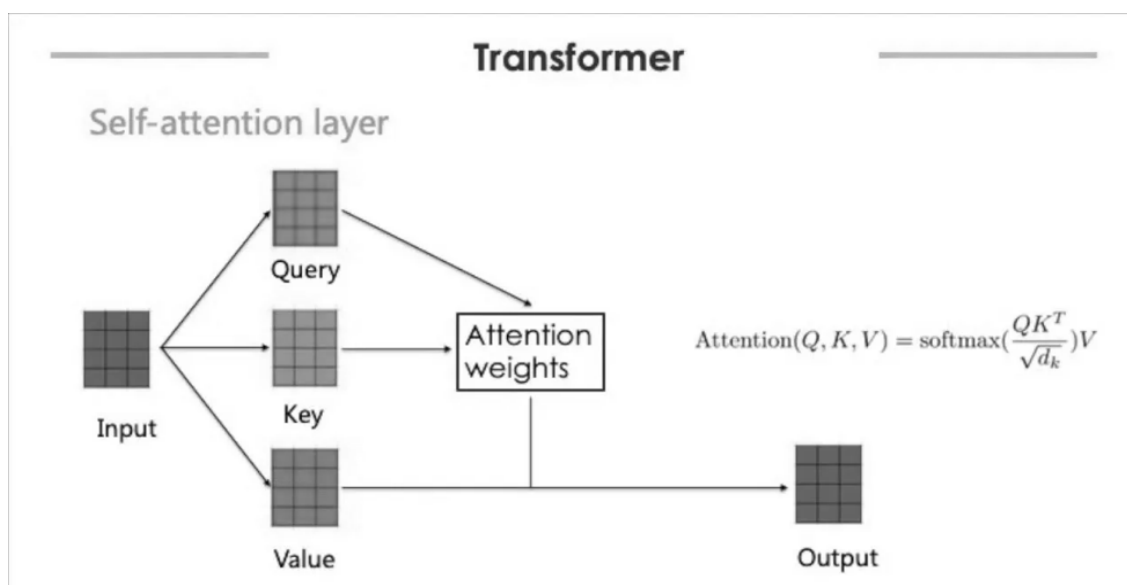


Рис. 1. Механизм Self-Attention

Здесь Q (Query), K (Key) и V (Value) – матрицы, преобразующие входные представления, а d_k – размерность ключей (рисунок 1). В рекомендательных системах новостей это означает, что модель может адаптивно выбирать, какие элементы пользовательской истории наиболее важны для текущего момента. Например, недавний клик на новость о спорте может быть более значимым, чем клик месячной давности, и механизм внимания это автоматически учитывает.

Multi-Head Self-Attention и многоуровневая персонализация

Развитием одноголового механизма внимания стали многоголовые архитектуры (multi-head attention), позволяющие модели одновременно анализировать информацию из разных пространств представлений. В контексте рекомендации новостей это критически важно, так как пользовательские предпочтения многогранны: одновременно может быть интерес к политическим новостям, технологическим трендам и спортивным событиям [3]. Каждый «голос» (head) self-attention может специализироваться на отдельном аспекте пользовательского профиля и контента новостей, обогащая общее представление модели. Например, первый голос может фокусироваться на семантическом сходстве заголовков, второй – на категориях новостей, третий – на авторах статей. Такой подход демонстрирует значительные улучшения в точности рекомендаций по сравнению с одноголовыми вариантами.

Практическое применение в рекомендационных системах VK и других платформ

В 2024-2025 годах крупные технологические компании активно внедряют трансформерные архитектуры с механизмами внимания в свои рекомендательные системы. VK, например, разработала архитектуру Discovery, которая объединяет мультимодальные нейросетевые модели, способные анализировать видео по нескольким модальностям: визуальному ряду, описанию, речи и фоновой музыке [4]. Рекомендательные трансформеры становятся центральным элементом, позволяя системе в реальном времени определять, какие элементы пользовательской истории наиболее важны для текущего запроса, и формировать персонализированные рекомендации с учетом сложного контекста поведения. На платформах с десятками миллионов активных пользователей это означает способность обрабатывать миллиарды взаимодействий и выдавать релевантные рекомендации с латентностью в несколько миллисекунд.

Вызовы и перспективы развития

Несмотря на впечатляющие результаты, рекомендательные трансформеры сталкиваются с рядом фундаментальных вызовов. Первый – это проблема холодного старта: когда система должна рекомендовать новости новому пользователю или новую статью, у неё недостаточно информации о взаимодействиях [5]. Второй – это вопрос информационной поляризации: чрезмерно персонализированные рекомендации могут привести к формированию информационных пузырей, где пользователь видит только контент, подтверждающий его существующие убеждения. Третий – вычислительная сложность: обучение и развёртывание больших трансформерных моделей требует значительных вычислительных ресурсов. Перспективы включают исследование методов, обеспечивающих баланс между персонализацией и информационным разнообразием, применение федеративного обучения для сохранения конфиденциальности пользователей, и интеграцию больших языковых моделей (LLM) для генерации персонализированного контента.

Механизмы внимания в трансформерных архитектурах представляют собой мощный инструмент для решения задачи персонализированной рекомендации новостей. Их способность адаптивно моделировать сложные взаимодействия между пользователями и контентом, а также обрабатывать мультимодальные данные в реальном времени, делает их незаменимыми в современных системах. По мере продолжения исследований и практического применения эти методы будут становиться всё более совершенными, открывая новые возможности для создания действительно персонализированного и этичного информационного опыта.

Библиографический список

1. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // *Advances in Neural Information Processing Systems*. – 2017. – P. 5998–6008.
2. ArXiv (2025). NRAM: News Recommendation with Attention Mechanism. *Journal of Industrial Engineering and Applied Science*. <https://doi.org/10.5281/zenodo.10635481>.
3. N+1 (2025). Рекомендуешь? Как разработчики AI VK создают систему рекомендаций контента // <https://nplus1.ru/material/2025/06/09/vk-recom>.
4. MySeldon (2025). VK внедрила технологии рекомендаций на базе ИИ для более точной персонализации контента // <https://myseldon.com/ru/news/index/337957563>.
5. Reg.Cloud (2026). Что такое рекомендательные системы, виды и примеры // <https://reg.cloud/blog/>.

УДК 004.93; ГРНТИ 28.23

НЕЙРОСЕТЕВЫЕ ПОДХОДЫ К СТИЛИЗАЦИИ ФОТОГРАФИЙ ПОД СТИЛЬ ХУДОЖНИКА: АРХИТЕКТУРЫ, ПАРАМЕТРЫ, СЦЕНАРИИ ПРИМЕНЕНИЯ

Г.В. Овечкин, Н.М. Хлебников

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, hlebnikita@gmail.com*

Аннотация. В работе рассматриваются подходы к художественной стилизации фотографий под стиль конкретного художника с использованием методов глубокого обучения. Описаны основные классы нейросетевых моделей, применяемых в данной задаче: классический нейронный стиль-трансфер, генеративно-состязательные сети и диффузионные модели. Особое внимание уделено комбинированным схемам, в которых используются несколько моделей и модулей: быстрые сети для предпросмотра, диффузионные модели с дообучением стиля (LoRA, стилевые токены), а также модули контроля композиции и сохранения идентичности лиц (ControlNet, маскирование, сегментация лица). На концептуальном уровне проанализировано влияние основных параметров этих моделей на баланс между сохранением содержимого фотографии, выраженностью художественного стиля и вычислительными затратами. Сформулированы рекомендации по выбору комбинаций моделей и их параметров для практических веб-приложений, а также намечены направления дальнейших экспериментальных исследований.

Ключевые слова: художественная стилизация, нейронный стиль-трансфер, диффузионные модели, LoRA, ControlNet, стилизация под художника.

NEURAL NETWORK APPROACHES TO STYLIZING PHOTOS TO THE ARTIST'S STYLE: ARCHITECTURE, PARAMETERS, AND APPLICATION SCENARIOS

G.V. Ovechkin, N.M. Hlebnikov

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, hlebnikita@gmail.com*

The summary. The paper discusses neural network approaches to artistic style transfer of photographs into the style of a particular painter. The main classes of models used for this task are reviewed: classical neural style transfer, generative adversarial networks and diffusion models. Special attention is paid to combined pipelines that employ several models and modules: fast feed-forward networks for preview, diffusion models with fine-tuned style adapters (LoRA, style tokens), and modules for preserving scene layout and facial identity (ControlNet, masking, face parsing). The influence of key parameters of these models on the trade-off between content preservation, stylistic expressiveness and computational cost is analysed at a conceptual level. Practical recommendations for choosing model combinations and their parameters in web applications are formulated, and directions for further experimental research are outlined.

Keywords: artistic style transfer, neural style transfer, diffusion models, LoRA, ControlNet, painterly style.

Одной из актуальных задач цифровой обработки изображений является автоматическая художественная стилизация фотографий, то есть преобразование реального фотоснимка в изображение, имитирующее технику конкретного художника или художественного направления. Практический интерес к данной задаче обусловлен развитием веб- и мобильных приложений для творчества пользователей, а также появлением мощных предобученных нейросетевых моделей, доступных для дообучения на пользовательских данных.

При этом возникает ряд противоречивых требований. С одной стороны, пользователю важна узнаваемость стиля выбранного автора: цветовая палитра, характер мазков, степень абстракции. С другой стороны, необходимо сохранять содержание исходного изображения: расположение объектов, структуру сцены, а в случае портретов – идентичность лица. Дополнительным ограничением для веб-приложений является время отклика и ограниченность вычислительных ресурсов.

В современных системах эти требования пытаются удовлетворить за счёт комбинирования различных нейросетевых моделей и модулей управления, а также за счёт аккуратного выбора гиперпараметров. В данной работе выполняется обзор основных классов моделей и анализируются типичные комбинации, применяемые при стилизации фотографий под стиль художника.

Задача

Пусть задано исходное цветное изображение I_c (фотография) и множество изображений в стиле художника a :

$$S_a = \{I_1^{(a)}, \dots, I_n^{(a)}\}.$$

Требуется построить отображение

$$f(I_c, a; \theta) \rightarrow I_s,$$

где I_s – результат стилизации, сохраняющий основное содержимое и композицию I_c , но обладающий визуальными признаками стиля автора a . Параметры θ соответствуют как архитектуре и весам нейросети, так и набору управляющих параметров (интенсивность стиля, число шагов семплирования, коэффициенты усиления и т.п.).

Основными критериями качества являются:

- сохранение содержимого: структура сцены, расположение объектов, корректное отображение лиц;
- выраженность стиля: узнаваемость выбранного художника;
- отсутствие артефактов и визуальная привлекательность результата;
- вычислительная эффективность: время генерации и требования к памяти.

Задача носит непарный-характер: отсутствуют пары «фото → картина», а имеются лишь отдельные множества фотографий и картин.

Нейросетевые подходы к художественной стилизации

Классический нейронный стиль-трансфер

Классический нейронный стиль-трансфер (Neural Style Transfer, NST) основан на представлении содержимого и стиля через признаки предобученной свёрточной сети [1]. Обычно в качестве базовой сети используется VGG-подобная архитектура. Содержимое описывается активациями верхних слоёв, стиль – статистиками (Gram-матрицами) активаций наборов слоёв.

В простой постановке генерируемое изображение I_s находится путём итеративной оптимизации, минимизирующей суммарную функцию потерь:

$$\mathcal{L}(I_s) = \alpha \cdot \mathcal{L}_{\text{content}}(I_s, I_c) + \beta \cdot \mathcal{L}_{\text{style}}(I_s, S_a),$$

где $\mathcal{L}_{\text{content}}$ измеряет отличие признаков содержимого, $\mathcal{L}_{\text{style}}$ – отличие стилевых статистик, α и β задают баланс «контент/стиль».

Первоначальные схемы требовали отдельной оптимизации для каждого нового изображения, что делало их малоприспособленными для интерактивных приложений. В последующих работах были предложены feed-forward модели, в которых обучается декодер, выполняющий стилизацию за один проход. Особый интерес представляет адаптивная нормализация признаков (Adaptive Instance Normalization, AdaIN), позволяющая отделить признаки контента от стилевых статистик и применять разные стили без переобучения всей сети.

Такие модели обеспечивают высокую скорость и подходят для предпросмотра, однако выраженность стиля и разнообразие эффектов часто уступают более современным генеративным моделям.

GAN-подходы

Генеративно-состязательные сети (GAN) [2] применяются для задач преобразования изображений, включая стилевую трансформацию. В рамках подхода CycleGAN и сходных ему моделей обучаются два генератора, выполняющих преобразование между двумя доменами изображений (например, «фото» и «картины») без парных примеров, с использованием циклической согласованности.

GAN-модели позволяют обучать прямое отображение вида

$$G: I_c \rightarrow I_s,$$

однако обучение таких моделей часто нестабильно, а контроль степени стилизации и сохранения содержимого может быть ограничен. Кроме того, GAN-подходы вытесняются в практике диффузионными моделями, которые демонстрируют более высокое качество и устойчивость.

Диффузионные модели и дообучение стиля

Диффузионные модели, в частности Stable Diffusion и SDXL, в настоящее время являются одним из основных инструментов для генерации изображений по текстовому описанию и для различных задач редактирования [3]. Базовая модель обучается на большом корпусе пар «изображение – текст» и позволяет синтезировать изображения по текстовому запросу.

Для стилизации под конкретного художника используют методы дообучения стиля:

- Textual Inversion – обучение специальных стилевых токенов в пространстве текста;
- LoRA (Low-Rank Adaptation) – добавление малых матриц низкого ранга к слоям модели, обучаемых на ограниченном наборе картин;
- другие методы адаптации (DreamBooth и др.).

Преимущество диффузионных моделей – высокое качество и выразительность результата, возможность работать в высоком разрешении и гибко управлять процессом генерации через текстовые подсказки и дополнительные условия.

Модули контроля композиции и содержимого

Для преобразования конкретных фотографий, а не абстрактных сцен, важно сохранить композицию и геометрию исходного изображения. В диффузионных моделях это достигается за счёт дополнительных модулей.

ControlNet

Архитектура ControlNet позволяет вводить дополнительный условный сигнал: карту контуров (Canny), глубины, нормалей и др. Исходная фотография преобразуется в набор «контрольных» карт, которые подаются в отдельную ветвь модели. Это позволяет накладывать новый стиль при сохранении основных контуров и структуры сцены.

Маскирование и сохранение лица

В задачах портретной стилизации особенно важно сохранить идентичность лица. Для этого используются:

- детекторы и сегментация лица (face-parsing);
- маскирование: ослабление силы стилизации в областях лица;
- дополнительные лоссы, сравнивающие эмбединги лиц до и после преобразования.

Комбинация ControlNet и масок позволяет отдельно управлять сохранением композиции кадра и детализацией ключевых объектов.

Комбинированные схемы обработки

На практике рассмотренные выше методы редко применяются в изоляции. Реальные системы, особенно веб-приложения, используют комбинации моделей, чтобы совместить высокое качество и приемлемое время отклика.

Одноэтапные решения

К одноэтапным решениям можно отнести:

- feed-forward стиль-трансфер (AdaIN-модель, обученная на множестве стилей);
- диффузионную модель без дополнительного контроля, где стиль задаётся лишь текстовым описанием;
- диффузионную модель с дообученным стилевым модулем (LoRA), но без ControlNet и масок.

Такие схемы проще по реализации, однако сложнее добиться оптимального компромисса между сохранением контента и выраженностью стиля, особенно для портретов.

Двухэтапные пайплайны

Перспективным подходом для веб-приложений является двухэтапный пайплайн.

Этап предпросмотра. Быстрая модель стиль-трансфера (например, AdaIN) применяет выбранный стиль к фотографии с малой задержкой. Пользователь может интерактивно изменять интенсивность стиля и оценивать общую художественную идею.

Этап финальной генерации. Диффузионная модель с дообученным стилевым модулем (LoRA) и, при необходимости, ControlNet и масками, применяется для получения итогового изображения высокого качества.

Такое разделение позволяет удовлетворить требования по интерактивности (быстрый предпросмотр) и качеству (ресурсоёмкая финальная стилизация), не перегружая пользователя деталями реализации.

Стилизация с сохранением идентичности лица

Отдельный класс комбинаций предназначен для портретов. В этом случае к двухэтапной схеме добавляются:

- подсистема обнаружения и сегментации лица;
- механизм маскирования и ослабления стиля в области лица;
- при необходимости – метрический модуль, контролирующий расстояние между эмбедингами лиц до и после стилизации.

Концептуально такие схемы реализуют многокритериальную оптимизацию, балансируя между стилем, содержимым и сохранением биометрических признаков.

Влияние ключевых параметров моделей

Даже при фиксированных архитектурах качество и свойства результата существенно зависят от выбора гиперпараметров. Ниже перечислены основные параметры и их ожидаемое влияние.

1. Параметры классического стиль-трансфера

- Отношение весов content/style loss (α/β). Увеличение веса стилевого лосса приводит к более выраженной стилизации, но может привести к искажению структуры сцены и появлению артефактов. Высокий вес content-loss, наоборот, сохраняет содержание, но делает стиль менее заметным.
- Интенсивность стиля в AdaIN (параметр смешивания статистик). Регулирует степень переноса стилевых признаков без изменения архитектуры модели. Малые значения дают мягкий эффект, большие – более «агрессивную» стилизацию.

2. Параметры диффузионных моделей

- Число шагов семплирования. Увеличение числа шагов, как правило, повышает качество и устойчивость результата, но увеличивает время генерации. Для веб-приложений важен компромисс между этими величинами.
- Коэффициент guidance (CFG scale). Управляет «силой» текстового или стилевого условия. Слишком низкие значения приводят к размытым изображениям, слишком высокие могут вызывать переусиление стиля и потерю содержимого.
- Сила LoRA (масштаб адаптационных весов). Определяет степень влияния конкретного стиля художника. При малых значениях стиль выражен слабо, при больших — возрастает риск искажения исходной сцены и появления артефактов.

3. Параметры модулей контроля

- Тип и вес контрольных сигналов в ControlNet. Использование карт контуров (Canny) помогает сохранить композицию и силуэты объектов, карты глубины позволяют лучше передать трёхмерную структуру. Усиление контрольного сигнала улучшает сохранение геометрии, но может ограничить свободу художественной стилизации.
- Параметры маски лица. Размер и размытие маски, а также масштаб стилизации внутри неё существенно влияют на баланс между сохранением идентичности и выразительностью стиля. Слишком «жёсткая» маска может сделать лицо чужеродным по отношению к фону, слишком «мягкая» — привести к искажению черт.

Направление дальнейших исследований

На основе рассмотренных подходов и комбинаций можно сформулировать ряд практических рекомендаций для веб-приложений, выполняющих стилизацию фотографий под стиль художника: для обеспечения интерактивности целесообразно использовать двухэтапный пайплайн: быструю feed-forward модель (AdaIN) для предпросмотра и диффузионную модель с LoRA-модулем для финальной генерации; при стилизации портретов рекомендуется применять ControlNet и маскирование лица, что позволяет сохранить композицию кадра и визуальную идентичность человека; параметры, сильно влияющие на качество (число шагов, CFG scale, сила LoRA, интенсивность стиля), должны быть ограничены разумными диапазонами.

Дальнейшие исследования могут быть направлены на:

- систематическое экспериментальное сравнение различных комбинаций моделей и параметров с использованием объективных метрик (LPIPS, SSIM, CLIP-подобие стиля, расстояния между эмбедингами лиц) и субъективных оценок;
- разработку новых метрик, более точно отражающих художественное восприятие результата;
- расширение рассматриваемых схем на видеоданные, где дополнительной задачей становится обеспечение временной согласованности стилизации.

Заключение

В работе рассмотрены основные нейросетевые подходы к художественной стилизации фотографий под стиль конкретного художника и проанализированы комбинированные схемы, применяемые в современных приложениях. Показано, что использование комбинаций моделей – быстрых сетей предпросмотра, диффузионных моделей с дообучением стиля

и модулей контроля композиции и лиц – позволяет более гибко управлять балансом между содержанием, стилем и вычислительной сложностью.

Несмотря на отсутствие экспериментальной части, выполненный концептуальный анализ комбинаций и параметров даёт основу для разработки практических систем, а также задаёт направления последующих исследований, связанных с количественной оценкой качества и оптимизацией архитектур для конкретных сценариев применения.

Библиографический список

1. Gatys L.A., Ecker A.S., Bethge M. A Neural Algorithm of Artistic Style // arXiv preprint arXiv:1508.06576, 2015.
2. Goodfellow I. et al. Generative Adversarial Nets // Proc. NIPS, 2014.
3. Rombach R. et al. High-Resolution Image Synthesis with Latent Diffusion Models // Proc. CVPR, 2022.

УДК 004.8:519.71; ГРНТИ 28.23.15

ПРОЕКТИРОВАНИЕ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ ПЛАНИРОВАНИЯ СВОБОДНОГО ВРЕМЕНИ

С.А. Петров, И.Ю. Филатов

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, mrflaim5427@gmail.com

Аннотация. В условиях избытка возможностей для досуга современный пользователь сталкивается с парадоксом выбора, который приводит к потере времени и снижению удовлетворенности. Данная статья посвящена исследованию и практической разработке интеллектуальной рекомендательной системы для персонализированного планирования свободного времени. В качестве базовой модели выбрана архитектура Two-Tower (Двух-башенная), доказавшая свою эффективность в задачах информационного поиска и рекомендаций. В статье детально рассматриваются этапы проектирования системы: от обоснования выбора модели и проектирования признаков пространств до процедуры обучения, оптимизации и интеграции в продукт. Особое внимание уделяется обработке гетерогенных данных (контекст пользователя, метаданные активностей, временные признаки) и решению проблемы «холодного старта». Предложенный подход демонстрирует путь создания адаптивной системы, способной не только реагировать на явные предпочтения, но и прогнозировать латентные интересы пользователя, превращая планирование досуга из рутины в увлекательный и эффективный процесс.

DESIGNING A RECOMMENDATION SYSTEM FOR FREE TIME PLANING

S.A. Petrov, I.Y. Filatov

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, mrflaim5427@gmail.com

The summary. In conditions of an abundance of leisure opportunities, the modern user is faced with a paradox of choice, which leads to a loss of time and a decrease in satisfaction. This article is devoted to the research and practical development of an intelligent recommendation system for personalized free time planning. The Two-Tower architecture was chosen as the base model, which has proven its effectiveness in information retrieval and recommendations. The article discusses in detail the stages of system design: from the justification of model selection and the design of feature spaces to the training procedure, optimization and integration into the product. Special attention is paid to processing heterogeneous data (user context, activity metadata, time features) and solving the "cold start" problem. The proposed approach demonstrates the way to create an adaptive system capable of not only responding to explicit preferences, but also predicting latent user interests, turning leisure planning from a routine into an exciting and effective process.

Свободное время – один из ключевых ресурсов современного человека, напрямую влияющий на качество жизни, уровень стресса и общее благополучие. Однако цифровая

эпоха, предоставив неограниченный доступ к информации о событиях, развлечениях, образовательных курсах и активностях, породила новую проблему – информационную перегрузку. Пользователь, желающий провести вечер, выходные или отпуск, вынужден тратить значительные усилия на поиск, фильтрацию и оценку подходящих вариантов. Традиционные методы поиска по категориям, рейтингам или местоположению оказываются недостаточными, так как не учитывают комплексный контекст: текущее состояние, настроение, долгосрочные интересы, состав компании и предыдущий опыт.

Рекомендательные системы, ставшие стандартом в e-commerce, стриминге и соцсетях, представляют собой естественное решение для этой проблемы. Однако планирование досуга – задача более сложная и многомерная, чем рекомендация фильма или товара. Она характеризуется:

Высокой стоимостью «ошибки»: Посещение нерелевантного мероприятия сопряжено с прямыми финансовыми затратами и потерей невозвратимого времени.

Сильной зависимостью от контекста: Одно и то же мероприятие может быть уместным в выходной день с друзьями и совершенно неуместным после рабочего дня в одиночестве.

Редкостью и нерегулярностью взаимодействий: Пользователь посещает ограниченное количество платных мероприятий в месяц, что создает проблему разреженности данных.

Необходимостью баланса между персонализацией и новизной: Система должна не только угождать известным предпочтениям, но и способствовать открытию новых, потенциально интересных сфер.

Таким образом, цель данного исследования — разработать и обучить модель, способную эффективно решать эти задачи, предоставляя релевантные, контекстно-зависимые и разнообразные рекомендации.

Выбор архитектуры модели

Среди множества подходов к построению рекомендательных систем (коллаборативная фильтрация, контентные методы, гибридные системы) глубокие нейронные сети демонстрируют наилучшие результаты благодаря своей способности извлекать сложные нелинейные паттерны из сырых данных. Архитектура Two-Tower (также известная как Dual Encoder или Siamese Network в определенных конфигурациях) стала де-факто стандартом для крупномасштабных систем сопоставления (matching) в таких компаниях, как Google, YouTube и Airbnb.

Альтернативные архитектуры

Выбор архитектуры Two-Tower в качестве основы для системы рекомендации досуга обусловлен ее уникальным балансом эффективности, гибкости и масштабируемости. Чтобы обосновать этот выбор, необходимо рассмотреть другие популярные классы моделей, проанализировав их принципы работы, преимущества и, что критически важно, ограничения в контексте нашей задачи.

Классические модели коллаборативной фильтрации (Collaborative Filtering - CF)

Эти модели составляют историческую основу рекомендательных систем и работают исключительно на матрице взаимодействий «пользователь-объект».

Принцип работы. Модели ищут паттерны сходства. User-based CF (UBCF) находит пользователей со схожей историей и рекомендует объекты, которые понравились им. Item-based CF (IBCF) находит объекты, схожие с теми, с которыми уже взаимодействовал пользователь. Более продвинутая модель SVD (Singular Value Decomposition) и ее вариации (SVD++) аппроксимируют разреженную матрицу взаимодействий, раскладывая ее на низкоранговые матрицы скрытых (латентных) факторов пользователей и объектов.

Преимущества. Простота реализации и интерпретации. Не требуют контентных признаков об объектах, полагаясь только на поведенческие паттерны («мудрость толпы»). SVD показывает хорошую точность на средних по размеру наборах данных.

Ограничения для планирования досуга. Проблема «холодного старта» (Cold Start): катастрофически плохая работа с новыми пользователями (нет истории) и новыми активностями (нет взаимодействий). Для сервиса досуга, где регулярно появляются новые события, а пользователи могут быть одноразовыми гостями, это неприемлемо.

Неспособность использовать контекст и контент. Модель не может учесть, что желание посетить ресторан зависит от времени суток, погоды или компании. Она также игнорирует текстовые описания, фотографии и другие метаданные событий.

Сложность с редкими взаимодействиями. Посещение платных мероприятий — гораздо более редкое событие, чем просмотр фильма или клик по товару. Матрица взаимодействий будет крайне разреженной, что резко снижает качество рекомендаций CF-методов.

Проблемы масштабируемости. Методы, основанные на поиске ближайших соседей (KNN), требуют вычисления попарных сходств, что вычислительно дорого для миллионов пользователей и объектов.

Глубокие ранжирующие модели (Deep Ranking Models)

К этому классу относятся сложные архитектуры, такие как DeepFM, DCN (Deep & Cross Network) и Neural Collaborative Filtering (NCF), которые часто следуют парадигме «раннего скрещивания» (early crossing) признаков.

Принцип работы. В этих моделях признаки пользователя и объекта (item) подаются на вход единой нейронной сети на раннем этапе. Это позволяет сети напрямую моделировать сложные нелинейные взаимодействия (кросс-фичи) между конкретными атрибутами пользователя и атрибутами активности (например, «любитель экстрима + высокий уровень сложности квеста»).

Преимущества. Максимальная точность прогноза. Способность улавливать тонкие, высокоуровневые зависимости между признаками, что делает их идеальными для финального ранжирования короткого списка кандидатов.

Ограничения для планирования досуга. Вычислительная неэффективность для поиска кандидатов: чтобы применить такую модель, необходимо выполнить forward pass для каждой возможной пары (пользователь, активность). Для каталога в миллионы мероприятий это делает онлайн-рекомендации в реальном времени невозможными. Они предназначены для этапа ранжирования, а не поиска кандидатов.

Сложность обновления. Любое изменение в каталоге (новое мероприятие) или признаковом пространстве требует полного пересчета модели или сложных онлайн-вычислений.

Графовые нейронные сети (Graph Neural Networks - GNN)

Модели, такие как LightGCN, представляют данные в виде графа, где узлы — это пользователи и активности, а ребра — взаимодействия между ними.

Принцип работы. Алгоритм распространяет информацию по графу, позволяя узлам (пользователям и активностям) обогащать свои представления (эмбединги) на основе информации от соседних узлов. Таким образом, эмбединг мероприятия обновляется на основе эмбедингов пользователей, которые его посещали, и наоборот.

Преимущества. Отлично улавливают косвенные коллаборативные сигналы высшего порядка (например, «друзья друзей» в графе взаимодействий). Показывают state-of-the-art результаты на задачах чистой коллаборативной фильтрации.

Ограничения для планирования досуга. Инженерная сложность: построение, хранение и обработка динамического графа с миллионами узлов – нетривиальная инфраструктурная задача.

Задержка инференса. Получение рекомендаций для пользователя требует вычислений на графе, что может быть медленнее, чем поиск по предварительно построенному ANN-индексу в Two-Tower архитектуре.

Интеграция контекстных признаков. Прямая интеграция неграфовых контекстных признаков (погода, время) в модель графа менее естественна, чем в архитектуру типа Two-Tower.

Принцип работы архитектуры Two-Tower. Модель состоит из двух независимых (но часто симметричных по архитектуре) нейронных сетей – «башен»:

- User Tower (Башня пользователя) кодирует информацию о пользователе и текущем контексте в плотный вектор-эмбединг фиксированной размерности (`user_embedding`);
- Item Tower (Башня активностей) кодирует информацию об объекте рекомендации (мероприятии, активности, месте) в вектор той же размерности (`item_embedding`).

Рекомендация происходит через вычисление меры близости (чаще всего – косинусного сходства или скалярного произведения) между `user_embedding` и `item_embedding`. Чем выше сходство, тем более релевантным считается объект для пользователя в данном контексте.

Проектирование системы

Проектирование рекомендательной системы для планирования досуга представляет собой комплексный процесс трансформации бизнес-требований и исследовательских гипотез в работоспособную техническую архитектуру. Этот этап начинается с глубокого анализа предметной области – определения ключевых сущностей (Пользователь, Активность, Сессия, Контекст), формализации типов взаимодействий (просмотр, клик, покупка, отмена) и постановки конкретных метрик успеха ($\text{Recall}@10$, CTR, долгосрочная удовлетворенность). На основе этого формируется принципиальная схема пайплайна данных, включающая сбор сырых логов, их очистку и обогащение контекстными признаками (погода, локация), а также создание обучающих выборок с применением стратегий негативного семплирования. Ключевым решением становится выбор многоступенчатой архитектуры (retrieval-ranking-reordering), где Two-Tower модель отвечает за этап retrieval (отбор сотен кандидатов из миллионов), что напрямую определяет проектирование её «башен»: User Tower должна эффективно инжектировать динамический контекст, а Item Tower – извлекать семантику из текста и метаданных. Параллельно проектируется инфраструктурный контур – оффлайн-пайплайн для обучения и индексации эмбедингов в векторной БД (например, FAISS) и низколатентный онлайн-сервис, способный за миллисекунды по ID пользователя и контексту вернуть персонализированные рекомендации. Таким образом, проектирование закладывает баланс между точностью модели, масштабируемостью системы и её способностью адаптироваться к постоянно меняющемуся контексту и каталогу активностей, что является критическим фундаментом для последующей успешной разработки и эксплуатации.

Формализация задачи и проектирование признаков пространств

Исходным пунктом проектирования является формализация объектов и взаимодействий в предметной области. Определяются базовые сущности: Пользователь (U), Активность (I) и Контекст сеанса (C). Взаимодействие между ними моделируется как функция отклика $Y = f(U, I, C)$, где Y – вероятность позитивного исхода (например, совершения покупки или углубленного просмотра). Проектирование признаков пространств для кодирования этих сущностей является критически важным.

Для энкодера пользователя (Башня Пользователя) проектируется объединенное пространство признаков, включающее:

- статический профиль – демографические атрибуты и явно указанные интересы;
- динамический поведенческий портрет – агрегированные за релевантный временной горизонт показатели: вектор предпочитаемых категорий; медианная стоимость посещенных мероприятий, частота активности;

- сессионный контекст – текущие метаданные, такие как время суток, день недели, погодные условия (кодируемые через интеграцию с внешними API), тип используемого устройства и географические координаты.

Для энкодера активности (Башня Объектов) формируется пространство на основе:

- контентно-семантических признаков, текстовые метаданные (название, описание) преобразуются в векторные представления с помощью предобученных языковых моделей. Категориальные атрибуты (жанр, тип мероприятия) подвергаются категориальному кодированию;

- атрибутивных и ограничительных признаков, числовые параметры (цена, возрастное ограничение, средний рейтинг) нормализуются. Пространственные координаты кодируются с использованием методов геопространственного хеширования;

- временных атрибутов, для событий с фиксированным временем начала извлекаются циклические признаки (час, день недели проведения).

Принципиальным архитектурным решением является разделение контекстуального сигнала. Контекстные признаки, описывающие ситуацию пользователя, подаются исключительно на вход башни пользователя. Это позволяет модели динамически адаптировать векторное представление пользователя в зависимости от обстоятельств, обеспечивая контекстуальную релевантность рекомендаций.

Архитектурное проектирование конвейера обработки

Система проектируется как комбинация двух основных контуров: периодического (офлайн) и оперативного (онлайн).

Офлайн-конвейер обучения и индексации выполняет следующие функции:

- регулярное (например, ежедневное) обновление векторных представлений для всего каталога активностей путем пропуска их актуальных признаков через Башню Объектов;

- сохранение полученных векторов в специализированном хранилище и построение на их основе индекса приближенного поиска ближайших соседей. Данный индекс оптимизирован для сверхбыстрого поиска семантически близких векторов в многомерном пространстве;

- периодическое переобучение или дообучение параметров обеих нейросетевых башен на аккумулированных исторических данных.

Онлайн-сервис выдачи рекомендаций функционирует в реальном времени. При получении запроса сервис агрегирует актуальные признаки для конкретного пользователя и текущего контекста. Данные признаки передаются на вход Башни Пользователя, которая мгновенно вычисляет контекстуализированное векторное представление. Это представление используется в качестве запроса к предварительно построенному индексу. В результате выполнения операции приближенного поиска возвращается упорядоченный список идентификаторов наиболее релевантных активностей (кандидатов). Данный список поступает на вход следующего этапа конвейера – многофакторного ранжирования.

Валидация проектных решений и обработка граничных случаев

Эффективность проектных решений оценивается с помощью строгой офлайн-валидации на историческом срезе данных. Ключевыми метриками для этапа поиска кандидатов являются Полнота – доля релевантных из прошлого, которую система находит среди К лучших кандидатов, и Покрытие каталога – доля всех доступных активностей, которая когда-либо попадает в рекомендации. Успешное проектирование характеризуется достижением высокой полноты при сохранении широкого покрытия, что свидетельствует об спо-

способности (ability) системы к персонализации без вырождения в узкий набор популярных позиций.

Отдельное внимание уделяется проектированию механизмов для граничных условий. Для новых пользователей с отсутствующей историей взаимодействий башня пользователя инициализируется на основе контекста и минимальных регистрационных данных, а в выдачу вносится управляемая доля популярных или географически близких событий. Для новых активностей их начальное векторное представление формируется исключительно на основе контентных признаков, что позволяет включать их в рекомендации до накопления поведенческой статистики.

Таким образом, процесс проектирования представляет собой итеративную оптимизацию взаимосвязанных компонентов – от корректного формирования признаков описаний и выбора архитектуры нейросетевых энкодеров до построения отказоустойчивой и масштабируемой инфраструктуры, способной обеспечить работу системы в условиях динамично изменяющегося каталога и разнородного контекста пользовательских запросов.

Библиографический список

1. Петров С.А., И. Ю Филатов // Исследование применения методов искусственного интеллекта в области автоматического планирования свободного времени // Межвузовский сборник научных трудов «Математическое и программное обеспечение вычислительных систем». 2025.
2. Петров С.А., И. Ю Филатов // Разработка системы нечеткого вывода в области автоматического планирования свободного времени // Межвузовский сборник научных трудов «Математическое и программное обеспечение вычислительных систем». 2025.
3. Бишоп Кристофер М., Распознавание образов и машинное обучение // – Издательство "Вильямс", 2020. 960 с. – ISBN 978-5-907144-55-2
4. Сивандам С.Н., Введение в генетические алгоритмы // – Издательство "Springer", 2017. 420 с

УДК 004.85; ГРНТИ 28.23

USER-BASED И ITEM-BASED ПОДХОДЫ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ В ЗАДАЧАХ АНАЛИЗА И УПРАВЛЕНИЯ АССОРТИМЕНТОМ

М.Д. Мацуганова, В.Е. Казаков

*Витебский государственный технологический университет,
Республика Беларусь, Vitebsk, macuganovamaria5@gmail.com, wadimchik@mail.ru*

Аннотация. В работе рассматриваются методы коллаборативной фильтрации – User-based и Item-based подходы, используемые для построения рекомендательных систем в электронной коммерции. Анализируется их применение не только для персонализации предложений, но и для решения обратной задачи – анализа и оптимизации товарного ассортимента на основе данных о поведении пользователей.

Ключевые слова: коллаборативная фильтрация, рекомендательные системы, управление ассортиментом, машинное обучение, анализ товарных взаимосвязей

USER-BASED AND ITEM-BASED APPROACHES TO COLLABORATIVE FILTERING IN THE PROBLEMS OF ANALYSIS AND ASSORTMENT MANAGEMENT

M.D. Matsuganova, V.E. Kazakov

*Vitebsk State Technological University,
Republic of Belarus, Vitebsk, macuganovamaria5@gmail.com, wadimchik@mail.ru*

The summary. This article examines collaborative filtering methods – user-based and item-based approaches – used to build recommender systems in e-commerce. It analyzes their application not only to personalizing offers but also to solving the inverse problem of analyzing and optimizing product assortment based on user behavior data.

Keywords: collaborative filtering, recommender systems, inventory management, machine learning, product relationship analysis

С развитием онлайн-торговли объем данных о взаимодействиях пользователей с товарами (просмотры, покупки, добавления в избранное) растет экспоненциально. Эти данные содержат скрытые закономерности о предпочтениях и товарных взаимосвязях, которые невозможно выявить вручную [1]. Именно здесь на первый план выходят методы машинного обучения, способные обрабатывать большие массивы информации и автоматически выявлять сложные паттерны [2].

Благодаря алгоритмам коллаборативной фильтрации компании могут не только рекомендовать товары («клиентам, купившим это, также нравится...»), но и эффективно управлять ассортиментом: выявлять товары-заменители (субституты) и товары-комплементы (дополняющие друг друга), оптимизировать выкладку на виртуальных полках и планировать закупки. Таким образом, переход от простой рекомендации к глубокому анализу товарной матрицы становится ключевым фактором эффективности ритейла.

Рассмотрим методы коллаборативной фильтрации подробнее. Основная идея коллаборативной фильтрации заключается в использовании знаний о предпочтениях группы пользователей для формирования предсказаний в отношении конкретного пользователя или товара. Существует два классических подхода: User-based и Item-based Collaborative Filtering (CF) [3].

User-based Collaborative Filtering (схожесть пользователей) – это исторически первый подход. Его суть заключается в поиске соседей для целевого пользователя – других покупателей со схожими предпочтениями или историей покупок. Рекомендация формируется на основе того, что понравилось «соседям», но еще неизвестно текущему пользователю. Для поиска соседей используются метрики сходства, такие как косинусная мера или коэффициент корреляции Пирсона [4]. Пример концептуальной реализации поиска схожих пользователей на Python представлен на рисунке 1.

```
# Функция для поиска N наиболее похожих пользователей
def find_similar_users(user_id, user_item_matrix, n=5):
    # Расчет косинусного сходства между всеми пользователями
    user_similarity = cosine_similarity(user_item_matrix)
    # Получение сходств для конкретного пользователя
    similar_users = list(enumerate(user_similarity[user_id]))
    # Сортировка по убыванию сходства и исключение самого пользователя
    sorted_users = sorted(similar_users, key=lambda x: x[1], reverse=True)[1:n+1]
    return sorted_users
```

Рис. 1. Концептуальный пример поиска похожих пользователей (User-based CF)*

Главным недостатком User-based подхода является его масштабируемость. Матрица сходства пользователей требует вычисления попарных сравнений для миллионов пользователей, что очень затратно. Кроме того, этот подход плохо объясняет, почему товар был рекомендован, и склонен рекомендовать слишком очевидные или, наоборот, случайные товары из-за "шума" во вкусах соседей.

Item-based Collaborative Filtering (схожесть товаров) было предложено как решение проблем масштабируемости и качества. Вместо поиска похожих людей, алгоритм ищет похожие товары. Основная предпосылка заключается в том, что пользователю, скорее всего, понравятся товары, похожие на те, которые ему уже нравились или которые он купил ранее. Сходство между товарами вычисляется на основе оценок всех пользователей: два товара считаются похожими, если их обычно оценивают или покупают одни и те же люди [5]. Этот подход менее динамичен, так как товарная база меняется медленнее, чем предпочтения пользователей, что позволяет пересчитывать матрицу сходства товаров реже (например, раз в день).

Для эффективного внедрения любого из подходов необходим достоверный массив данных о взаимодействиях. Наиболее распространенным источником информации является матрица «пользователь-товар» (user-item matrix), которая строится на основе транзакционных данных или логов просмотров. Ключевыми параметрами для анализа являются: user_id, item_id, rating (или purchase_flag). Эти данные позволяют количественно оценить связь между сущностями.

На основе исходных транзакционных данных была построена бинарная матрица «пользователь-товар» (user-item matrix), где строки соответствуют уникальным идентификаторам пользователей, столбцы – товарным позициям, а значения (0 или 1) отражают факт приобретения товара конкретным пользователем. Такая матрица является фундаментом для любого вида коллаборативной фильтрации, так как позволяет перейти от списка транзакций к математическому представлению взаимодействий в векторном пространстве.

Ключевым этапом анализа является расчет сходства между векторными представлениями товаров. Для этих целей была применена метрика косинусной близости (cosine similarity), которая измеряет косинус угла между векторами товаров в многомерном пространстве пользователей. Косинусная мера выбрана по следующим причинам: она устойчива к разреженности данных, учитывает только ненулевые значения и эффективно работает с бинарными данными. Значения косинусной близости варьируются от 0 до 1, где 0 означает полное отсутствие общих покупателей, а 1 – идентичные паттерны покупок [6].

Реализация алгоритма на языке Python с использованием библиотек pandas и scikit-learn представлена на рисунке 2.

```
# Функция для рекомендации товаров, похожих на заданный
def recommend_items(item_id, item_similarity_df, n=5):
    similar_scores =
item_similarity_df[item_id].sort_values(ascending=False)
    # Исключаем сам товар и берем топ-N
    top_items = similar_scores.iloc[1:n+1]
    return top_items
```

Рис. 2. Пример реализации расчета сходства товаров

Матрица косинусного сходства между товарами представлена на рисунке 3.

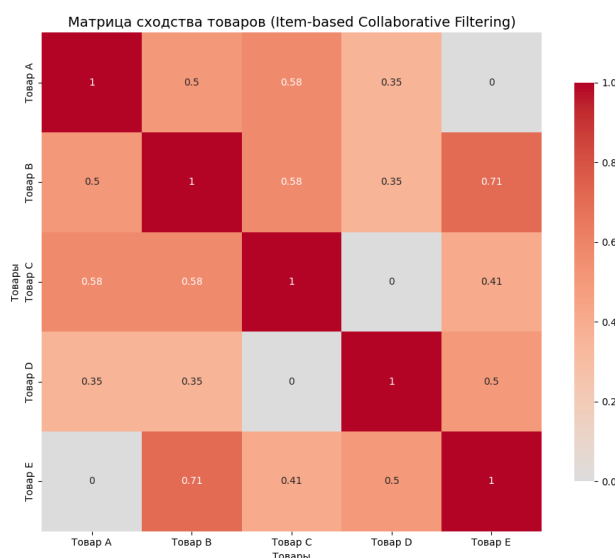


Рис. 3. Визуализация матрицы косинусного сходства между товарами

Таким образом, можно сделать вывод, что Item-based алгоритм автоматически выявляет смысловые и поведенческие связи между товарами, формируя кластеры взаимозаменяемых или взаимодополняющих продуктов.

Полученные в результате работы Item-based CF матрицы сходства открывают широкие возможности для управления ассортиментом, выходящие за рамки простых рекомендаций:

1. *Выявление товаров-комплементов.* Алгоритм показывает, какие товары часто покупают в паре (например, ноутбук и сумка для ноутбука). Это позволяет создавать комплекты («Купи вместе»), увеличивая средний чек;

2. *Анализ товаров-субститутов.* Товары с очень высоким коэффициентом сходства (близким к 1) являются прямыми конкурентами друг для друга в глазах покупателя. Если один из таких товаров продается плохо, а другой хорошо, это сигнал к пересмотру ценообразования, упаковки или целесообразности присутствия обоих товаров в линейке;

3. *Оптимизация каталога.* Анализ разреженных связей позволяет выявить товары, которые не имеют сильных корреляций с другими. Это могут быть нишевые или непопулярные позиции («изгои»), требующие ребрендинга или вывода из ассортимента;

4. *Управление запасами.* Зная, что товар X является сильным compleментом для товара Y, можно прогнозировать рост спроса на X при планировании маркетинговой акции на Y, предотвращая дефицит.

В отличие от User-based подхода, который лучше подходит для сценариев «холодного старта» пользователя (когда о новом пользователе мало данных) и таргетинга в реальном времени, Item-based подход обеспечивает стабильную и интерпретируемую модель товарных взаимосвязей, что делает его незаменимым инструментом для аналитиков и категорийных менеджеров.

Таким образом, User-based и Item-based алгоритмы являются мощным и взаимодополняющим инструментарием. User-based CF позволяет гибко реагировать на сиюминутные предпочтения аудитории, повышая вовлеченность. Item-based CF, в свою очередь, обеспечивает аналитическую базу для стратегического управления ассортиментом: он помогает структурировать товарную матрицу, выявлять скрытые продуктовые кластеры и зависимости, оптимизировать закупки и ценообразование. Используя оба подхода в комбинации, компании могут достичь синергетического эффекта, повышая как лояльность клиентов, так и операционную эффективность бизнеса.

Библиографический список

1. Future of E-commerce: Personalization Statistics [Electronic resource]. – Mode of access: <https://www.example.com/ecommerce-personalization-stats>. – Date of access: 10.01.2026.
2. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль. – М.: ДМК Пресс, 2018. – 652 с.
3. Aggarwal, C. C. Recommender Systems: The Textbook / C. C. Aggarwal. – Springer, 2016. – 498 p.
4. Leskovec, J. Mining of Massive Datasets / J. Leskovec, A. Rajaraman, J. D. Ullman. – Cambridge University Press, 2014. – 513 p.
5. Amazon.com Recommendations: Item-to-Item Collaborative Filtering [Electronic resource] / G. Linden, B. Smith, J. York // IEEE Internet Computing. – Mode of access: <https://example.com/amazon-cf-paper>. – Date of access: 12.01.2026.
6. Системы рекомендаций: коллаборативная фильтрация [Электронный ресурс] // Хабр. – Режим доступа: <https://habr.com/ru/articles/> – Дата доступа: 15.01.2026.

УДК 616-053.32; ГРНТИ 59.14

ПРИМЕНЕНИЕ ГРАДИЕНТНОГО НЕЙРОСЕТЕВОГО УПРАВЛЕНИЯ ДЛЯ ЗАМЕЩЕНИЯ КОМПОНЕНТОВ БИОТЕХНИЧЕСКИХ СИСТЕМ

А.А. Коробов, К.С. Савинова

Тамбовский государственный технический университет,
Российская Федерация, Тамбов, korobov1991@mail.ru

Аннотация. В работе рассмотрен подход к применению методов нейронных сетей для задач управления сложными биотехническими системами. Авторами применяется разработанный метод градиентного нейросетевого управления (ГНСУ) для двух задач – управления параметрами микроклимата в неонатальном инкубаторе, а также построения цифрового двойника сердечно-сосудистой системы с учётом механизмов регуляции. В обоих случаях отмечается эффективность метода при работе с нелинейными, нестационарными, стохастическими, многосвязными объектами управления. Предложенный метод в перспективе направлен на создание систем поддержки принятия врачебных решений, в том числе в неонатологии и кардиологии.

Ключевые слова: градиентный метод, нейросетевое управление, нейронные сети, сердечно-сосудистая система, регуляция, математическое моделирование, регуляция, неонатальный инкубатор.

APPLICATION OF GRADIENT NEURAL NETWORK CONTROL TO REPLACE COMPONENTS OF BIOTECHNICAL SYSTEMS

A.A. Korobov, K.S. Savinova

Tambov State Technical University,
Russia, Tambov, korobov1991@mail.ru

The summary. This paper examines an approach to applying neural network methods to control complex biotechnical systems. The authors apply the developed gradient neural network control (GNNC) method to two problems: controlling microclimate parameters in a neonatal incubator and constructing a digital twin of the cardiovascular system, taking into account regulatory mechanisms. In both cases, the method's effectiveness is noted when working with nonlinear, nonstationary, stochastic, and multivariate control objects. The proposed method is potentially aimed at creating systems for supporting medical decision-making, including in neonatology and cardiology.

Keywords: gradient method, neural network control, neural networks, cardiovascular system, regulation, mathematical modeling, regulation, neonatal incubator.

Ряд современных биотехнических систем (БТС) характеризуется высокой степенью сложности – в первую очередь, это связано с их нелинейной динамикой и высоким количеством внутренних параметров. К числу подобных БТС относят напрямую функциональные системы организма, так и системы автоматического управления, взаимодействующие с ними в контуре обратной связи.

Авторами рассматривается две задачи – моделирование сердечно-сосудистой системы с учётом механизмов нейрогуморальной регуляции, а также поддержание внутренних параметров в неонатальных инкубаторах при воздействии изменяющихся внешних факторах.

Задача математического моделирования сердечно-сосудистой системы, учитывающей механизмы регуляции остаётся трудной задачей, так как её механизмы до сих пор недостаточно изучены. Природа и даже общая концепция регуляции до сих пор вызывает дискуссии [1]. Анализ современных публикаций показал, что намечается тренд на применение методов нейронных сетей, искусственного интеллекта и машинного обучения для замещения элементов функциональных систем организма для задач математического моделирования.

Задача управления микроклиматом в неонатальных инкубаторах до сих пор является актуальной, так как применение традиционных линейных ПИД-регуляторов приводит к снижению точности поддержания параметров и снижению устойчивости работы системы.

Общей особенностью указанных задач является вопрос управления сложными объектами управления в условиях нелинейности и ограниченности вычислительной мощности и времени расчётов. Перспективным подходом к решению обеих задач является применение относящегося к классу адаптивных алгоритмов метода градиентного нейросетевого управления (ГНСУ).

Предлагаемый авторами метод ГНСУ основан на применении многослойного перцептрона с одним скрытым слоем. Управляющие воздействия формируются на основе вектора рассогласования между требуемыми и заданными значениями. Коррекция весовых коэффициентов осуществляется методом обратного распространения ошибки с применением регуляризации. Алгоритм характеризуется низким числом параметров, требующих настройки – шага метода и параметра регуляризации [2].

ГНСУ функционирует как адаптивный регулятор в составе системы с обратной связью. В отличие от классических алгоритмов, ГНСУ реализует нелинейное функциональное преобразование вектора ошибки рассогласования между заданными и фактическими параметрами объекта в вектор управляющих воздействий. Обучение и адаптация нейроконтроллера происходят в режиме реального времени на каждом шаге дискретизации путем минимизации целевой функции погрешности градиентным методом. Для обеспечения устойчивости системы при управлении многосвязными и стохастическими объектами в алгоритме применяется процедура регуляризации, предотвращающая неограниченный рост весовых коэффициентов. Это позволяет системе эффективно компенсировать внешние возмущения и внутренние перекрестные связи объекта, обеспечивая высокую точность стабилизации выходных переменных.

Обобщённые концепции биотехнических систем для обеих задач представлены на рисунке 1.

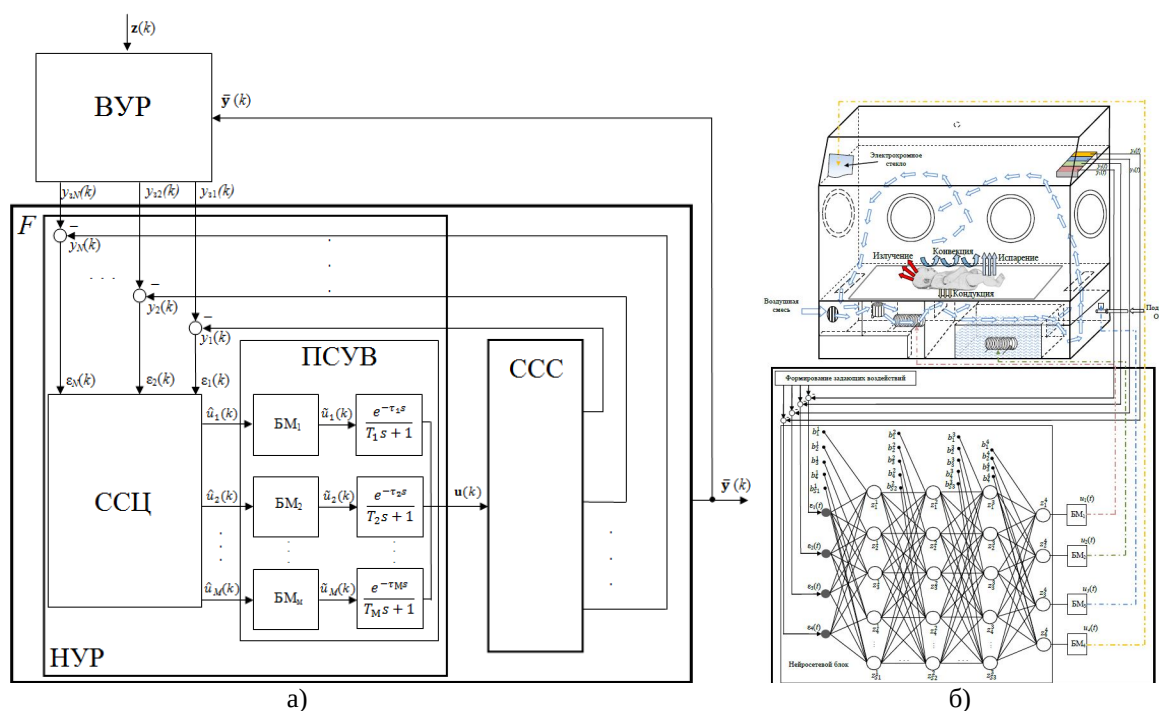


Рис. 1. Обобщённые концепции биотехнических систем: а) модель сердца и сосудов с классической регуляцией; б) управление микроклиматом инкубатора с нейроконтроллером на основе ГНСУ

Первая задача ориентирована на перспективу создания цифрового двойника на основе базовой модели сердечно-сосудистой системы, объединённой с моделью классической регуляции. Предлагаемый подход направлен на потенциальное внедрение цифрового двойника кардиологическую СППВР. Базовая модель строится на математической модели серд-

ца и сосудов с сосредоточенными параметрами, при этом учитывается объём циркулирующей крови.

Давление крови в сосудах, объеме и кровотоки являются выходными параметрами (всего 52 координаты), ненапряжённый объём, проводимость и жёсткость сосудов – управляющие параметры. Нейрогуморальная регуляция построена в виде двухуровневой системы, где нижний уровень реализован на основе авторского ГНСУ. На рисунке 1а ВУР – высший уровень регуляции, ССЦ – сердечно-сосудистый центр, реализованный на основе ГНСУ, ПСУВ – передача сигналов управляющих воздействий, ССС – сердечно-сосудистая система. В результате имитационных исследований по имитации механизмов нейрогуморальной регуляции функций сердечно-сосудистой системы – средних кровотоков артериального русла верхней части в капиллярной русло верхней части тела, средних кровотоков артериального русла нижней части в капиллярной русло нижней части тела, а также аналогичных экспериментов при имитации регуляции средних давлений артериального русла верхней части тела и нижней части тела авторами была выбрана нейронная сеть типа многослойный перцептрон с одним внутренним слоем и 150-ю нейронами (при параметрах $\eta=0,15$, $\lambda=0,01$) [4].

В контексте создания цифрового двойника сердечно-сосудистой системы ГНСУ выполняет функцию нижнего уровня иерархической регуляции, имитируя работу сердечно-сосудистого центра, по сути являясь средством функционального замещения отдельных компонентов регуляторной подсистемы. Это позволяет моделировать сложные нейрогуморальные механизмы поддержания гомеостаза, такие как барорефлекс и хеморефлекс, путем динамической коррекции проводимости, жесткости и ненапряженного объема 17 упругих камер, моделирующих сосуды и камеры сердца.

Идентификация персонализированных параметров модели (включая весовые коэффициенты нейронной сети) осуществляется на основе неинвазивных клинических данных, получаемых с помощью МРТ и доплерографии, что обеспечивает настройку цифрового двойника под конкретного пациента. Благодаря этому разработанная система позволяет не только с высокой точностью воспроизводить индивидуальные пульсовые волны давления в аорте и фазовые петли «объем-давление» для желудочков, но и прогнозировать их изменение при различных сценариях лекарственной терапии, что является ключевым функционалом для систем поддержки принятия врачебных решений в кардиологии.

Во второй задаче объектом управления является неонатальный инкубатор с новорожденным, характеризуемый как нелинейная, многосвязная система с взаимосвязанными параметрами: температурой воздуха, относительной влажностью, концентрацией кислорода и освещённостью (рисунок 1б). Разработанная комплексная математическая модель тепло- и массопереноса в НИ включает 18 выходных параметров, 4 управляющих воздействия и 45 параметров объекта. Алгоритм ГНСУ здесь применён в качестве регулятора, который по сигналам ошибок формирует ШИМ-сигналы на нагреватели, увлажнитель, клапан кислорода и электрохромное стекло.

Неонатальный инкубатор описывается как сложная биотехническая система, в которой процессы тепло- и массопереноса формируются взаимодействием тела новорожденного, представленного в виде блочной модели ядра и кожных покровов, и технических элементов установки – нагревателя, увлажнителя, потоков воздуха, системы подачи кислорода и конструктивных элементов камеры.

Применение нейросетевого контроллера, реализующего градиентную процедуру адаптации, позволяет в реальном времени корректировать управляющие воздействия по каналам температуры, влажности, концентрации кислорода и освещённости, обеспечивая требуемые условия жизнедеятельности новорожденного при изменении параметров внешней среды и внутренних теплообменных процессов. Такой подход способствует снижению статических и динамических ошибок регулирования по сравнению с классическими линей-

ными методами и создаёт основу для построения интеллектуальных систем поддержки врача-неонатолога, в том числе при проектировании СППВР.

В решаемой задаче стоит отметить, что настройка ГНСУ производится двумя этапами. На первом этапе работа ГНСУ корректируется с помощью математической модели неонатального инкубатора. Здесь ищутся настроечные параметры ГНСУ η , λ , а также выбирается архитектура нейронной сети. Второй этап настройки подразумевает работу непосредственно на самом неонатальном инкубаторе (экспериментальной лабораторной установке), при этом производится коррекция настроечных параметров η , λ алгоритма, в том числе под текущие компоненты установки – датчики, управляющие устройства.

Имитационные и эксплуатационные исследования (выход на режим) продемонстрировали, что ГНСУ обеспечивает стабилизацию температуры с погрешностью не более 0,5 °С, влажности – с колебаниями 2–3 %, а также эффективно компенсирует возмущения (открытие инкубатора) и адаптируется к смене клинических протоколов [5, 6]. Целый ряд экспериментов проведён по проверке работы ГНСУ с применением авторского полноростового гидродинамического фантома новорожденного, созданного с целью имитации терморегуляции младенца (управление по коже ребёнка). Отмечено, что управление температурой поверхности фантома происходит за 26 минут (задание – обеспечение температуры 37 °С при начальной температуре 34,5 °С), при относительной небольшой нагрузке на нагреватель воздуха (ШИМ нагревателя 26%). Влажность поддерживалась на стабильном уровне 70% со снижением динамической ошибки.

Реализованный ГНСУ в обеих задачах функционирует в несколько этапов. На начальном этапе производится запуск алгоритма и инициализация весовых коэффициентов нейронной сети, определяющих структуру формируемых управляющих воздействий. Далее выполняется вычисление выходных сигналов нейронов скрытого и выходного слоёв, а также соответствующих производных, необходимых для последующей процедуры адаптации. Следующий этап связан с определением локальных градиентов ошибки управления на основе метода обратного распространения ошибки, что позволяет количественно оценить влияние каждого параметра сети на отклонение регулируемых координат от заданных значений. На четвёртом этапе осуществляется корректировка весовых коэффициентов нейронной сети, интерпретируемых как множество управляющих координат алгоритма, в направлении уменьшения целевой функции управления. Завершающий этап состоит в переходе к следующему шагу дискретного [4-6].

Представленные примеры демонстрируют как универсальность разрабатываемого ГНСУ как инструмента в составе БТС повышенной сложности, так и потенциал применения нейросетевых методов управления, в целом [7-10]. В обоих случаях нейронная сеть выступает в роли модели, которая обучается воспроизводить реакцию БТС на внешние возмущения.

Разработанный метод градиентного нейросетевого управления представляет собой эффективный подход к моделированию и управлению биотехническими системами. Дальнейшее исследование темы связано с уточнением параметров ГНСУ для прикладных задач, а также математическим доказательством устойчивости метода.

Библиографический список

1. Концепция регуляции сердечно-сосудистой системы – от управления функциями к согласованию возможностей. Часть 5. Роль регуляции / Л. А. Бокерия, В. А. Лищук, Д. Ш. Газизова [и др.] // Клиническая физиология кровообращения. – 2013. – № 1. – С. 34-44. – EDN RCMMGZ.
2. Градиентный метод нейросетевого управления многосвязными нелинейными нестационарными стохастическими системами / С. В. Фролов, А. А. Коробов, А. Ю. Потлов, К. С. Савинова // Приборы и системы. Управление, контроль, диагностика. – 2021. – № 5. – С. 41-48. – DOI 10.25791/pribor.5.2021.1262. – EDN NHLYAD.
3. Коробов, А. А. Современные тенденции математического моделирования сердечно-сосудистой системы / А. А. Коробов // Радиотехника. Проблемы и перспективы развития : сборник трудов X Всероссийской

научно-практической конференции с международным участием, посвященной «Дню радио и связи» и 80-летию победы в Великой Отечественной войне, Тамбов, 07 мая 2025 года. – Тамбов: Тамбовский государственный технический университет, 2025. – С. 233 – 234.

4. Dual-contoured model of cardiovascular system regulation / A. A. Korobov, S. V. Frolov, N. E. Aliyev, I. E. Rodionova // Journal of Physics: Conference Series : VI International Scientific and Practical Conference "Virtual Simulation, Prototyping and Industrial Design 2019, VSPID-2019", Tambov, 16–18 октября 2019 года. Vol. 1553. – Tambov: Institute of Physics Publishing, 2020. – P. 012006. – DOI 10.1088/1742-6596/1553/1/012006.

5. Савинова, К. С. Разработка и исследование экспериментальной модели неонатального инкубатора нового поколения / К. С. Савинова // Радиоэлектроника. Проблемы и перспективы развития : сборник трудов X Всероссийской научно-практической конференции с международным участием, посвященной «Дню радио и связи» и 80-летию победы в Великой Отечественной войне, Тамбов, 07 мая 2025 года. – Тамбов: Тамбовский государственный технический университет, 2025. – С. 261 – 262.

6. Нейросетевая система управления микроклиматом и освещенностью в неонатальном инкубаторе / С. В. Фролов, А. А. Коробов, К. С. Савинова, А. Ю. Потлов // Датчики и системы. – 2025. – № 1(279). – С. 62-68. – DOI 10.24412/1992-7185-2025-1-62-68.

7. Фролов С.В., Судаков Д.Е. Совместное использование методов функциональной диагностики и математического моделирования сердечно-сосудистой системы ребенка с врожденным пороком сердца. Российский кардиологический журнал. 2024; 29(17). 44 с.

8. Use of a neural controller and electrochromic glazing in incubators for neonates for adaptive simulation of day/night changes / A. Yu. Potlov, S. V. Frolov, A. A. Korobov, K. S. Savinova // Biomedical Engineering. – 2024. – Vol. 58, No. 4. – P. 234-237. – DOI 10.1007/s10527-024-10406-y.

9. Фролов, С. В. Выбор целесообразной системы управления технологическими процессами при наличии неопределенных параметров / С. В. Фролов, Т. А. Фролова // Вестник Тамбовского государственного технического университета. – 2018. – Т. 24, № 4. – С. 591-596. – DOI 10.17277/vestnik.2018.04.pp.591-596.

10. Sindeev, S. V. Experimental setups for studies of blood flow in arteries of cardiovascular system / S. V. Sindeev, S. V. Frolov, D. Liepsch // Вопросы современной науки и практики. Университет им. В.И. Вернадского. – 2017. – No. 4(66). – P. 231-236. – DOI 10.17277/voprosy.2017.04.pp.231-236.

УДК 004.822; ГРНТИ 20.53.19

ПРИМЕНЕНИЕ АРХИТЕКТУРЫ MIXTURE OF EXPERTS В ПОСЛЕДОВАТЕЛЬНЫХ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ

Г.В. Овечкин, Д.И. Успенский

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, denisuspenskiy@mail.ru*

Аннотация. В работе рассматривается проблема ограниченной адаптивности современных последовательных рекомендательных моделей, основанных на трансформерных архитектурах. В качестве решения предлагается интеграция архитектуры Mixture of Experts (MoE) в предсказательный слой последовательной модели. Подробно рассмотрены принципы работы гейтинг-механизма, приведена вероятностная интерпретация MoE как модели скрытой сегментации пользователей, формализована функция потерь с балансирующей регуляризацией. Проведён экспериментальный анализ на публичном наборе данных, демонстрирующий улучшение метрик NDCG@10, Hit@10 и MRR. Полученные результаты подтверждают целесообразность применения MoE в рекомендательных системах с гетерогенной аудиторией.

Ключевые слова: рекомендательные системы, Mixture of Experts, MoE, гейтинг механизм, трансформер, последовательные модели, условная параметризация.

SEQUENTIAL RECOMMENDATION SYSTEMS WITH MIXTURE OF EXPERTS ARCHITECTURE

G.V. Ovechkin, D.I. Uspenskiy

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, denisuspenskiy@mail.ru*

The summary. This paper addresses the limited adaptability of transformer-based sequential recommender systems. Although such models achieve strong predictive performance, they rely on a single shared parameter space across heterogeneous user behavior patterns. We propose integrating a Mixture of Experts (MoE) prediction block into a sequential transformer architecture. The

gating mechanism is analyzed in detail, and a probabilistic interpretation of MoE as latent user segmentation is provided. The loss function with load-balancing regularization is formulated. Experimental results demonstrate consistent improvements in NDCG@10, Hit@10, and MRR compared to the baseline model. The findings confirm the effectiveness of MoE in large-scale heterogeneous recommendation scenarios.

Keywords: recommender systems, Mixture of Experts, gating, transformer, MoE, sequential models, conditional parametrization, cold-start.

Рекомендательные системы являются неотъемлемой частью современных цифровых платформ электронной коммерции, видеосервисов, музыкальных стриминговых сервисов и социальных сетей. Их основная задача заключается в предсказании вероятности взаимодействия пользователя с объектом на основе накопленной истории поведения [1]. В последние годы последовательные модели, основанные на механизме *self-attention*, продемонстрировали значительный прирост качества рекомендаций. Широкое распространение получили модели на основе трансформеров, такие как *SASRec*, *BERT4Rec* и их модификации. Эти модели успешно используют механизм *self-attention* для выявления как краткосрочных, так и долгосрочных зависимостей во временной динамике пользовательских действий.

Пусть последовательность взаимодействий пользователя задаётся как: $S_u = \{i_1, i_2, \dots, i_T\}$, где i_t – объект, с которым пользователь взаимодействовал в момент времени t . Трансформерная модель формирует скрытые представления: $H = \{h_1, h_2, \dots, h_T\}$, $h_t \in \mathbb{R}^d$, где h_t – скрытое представление слоя t размерностью d .

Механизм внимания вычисляется по формуле:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V,$$

где $Q, K, V, Q \in \mathbb{R}^{T \times d_k}, K \in \mathbb{R}^{T \times d_k}, V \in \mathbb{R}^{T \times d_v}$ – матрицы запросов, ключей и значений соответственно;

T – длина входящей последовательности;

d – размерность скрытого пространства модели.

Финальное представление пользователя определяется как:

$$z_u = h_T.$$

А предсказание вероятности следующего взаимодействия:

$$\hat{y}_{u,i} = \sigma(z_u^T e_i),$$

где u – пользователь системы, i – товар, z_u – представление пользователя, e_i – представление товара, σ – функция сигмоида [2].

Несмотря на высокую точность, подобные модели имеют фундаментальное ограничение – они используют единый набор параметров для всех пользователей. Это означает, что модель вынуждена компромиссно аппроксимировать различные поведенческие стратегии – импульсивные покупки, лояльность к бренду, ориентацию на скидки, категориальные предпочтения и др. В реальных системах поведение пользователей существенно гетерогенно, и единая параметризация приводит к снижению специализации модели. В этой связи возникает задача увеличения выразительной способности без линейного роста вычислительной сложности.

Архитектура *Mixture of Experts (MoE)* была предложена *Jacobs et al.* (1991) как способ объединения нескольких специализированных моделей с обучаемой системой маршрутизации. Рассмотрим задачу рекомендации как задачу моделирования условного распределения: $P(i_{T+1} | S_u)$, где i_{T+1} – объект следующего взаимодействия, S_u – история взаимодействий пользователя u . Предположим существование скрытой переменной z , отражающей тип поведенческого шаблона пользователя. Тогда вероятность следующего взаимодействия можно представить:

$$P(i_{T+1} | S_u) = \sum_{k=1}^K P(z = k | S_u) P(i_{T+1} | S_u, z = k),$$

где K – количество экспертов, k – текущий выбранный эксперт.

Классическая модель трансформера аппроксимирует лишь одно распределение, тогда как *Mixture of Experts* позволяет реализовать смесь специализированных условных распределений. Это особенно важно в условиях:

- высокой дисперсии поведения пользователей;
- большого числа товарных категорий;
- наличия вырожденного распределения популярности товаров.

Mixture of Experts представляет собой композицию: набора экспертов $E_k(x)$, гейтинг-сети $g(x)$, механизма агрегирования [3]. Общая формула имеет вид:

$$f(x) = \sum_{k=1}^K g_k(x) E_k(x),$$

где $g_k(x)$ – гейтинг функция, а $E_k(x)$ – выход k -го эксперта.

$$g(x) = \text{softmax}(W_g x + b),$$

$$E_k(x) = W_k x,$$

где W_g, W_k – обучаемые матрицы параметров модели, а b – обучаемый вектор смещения.

Интеграция МоЕ в рекомендательную модель предполагается следующим образом:

- трансформер кодирует последовательность взаимодействий пользователя в виде вектора z_u ;
- вектор z_u поступает в МоЕ блок;
- формируется итоговое представление пользователя:

$$\tilde{z}_u = \sum_{k=1}^K g_k(z_u) W_k z_u.$$

- формируется оценка вероятности покупки товара e_i :

$$\tilde{y}_{u,i} = \tilde{z}_u^T e_i.$$

В качестве функции потерь используется *BPR-loss*:

$$L_{bpr} = - \sum_{(u,i,j)} \log \delta(\hat{y}_{u,i} - \hat{y}_{u,j})$$

где u – пользователь, i – позитивный товар, купленный этим пользователем, а j – негативный товар, отвергнутый пользователем.

В данной архитектуре существует важная проблема. Существует эксперт, который начинает доминировать над другими, т.е. $g_1(x) \approx 1, g_{2..K}(x) \approx 0$. В этом случае фактически работает только один эксперт, а остальные почти не обучаются. Модель вырождается в обычную нейросеть и теряется весь смысл *МоЕ*. Для решения используется балансирующая регуляризация, которая требует, чтобы в среднем по батчу эксперты использовались примерно одинаково. Например, если в батче 1000 примеров и 4 эксперта, то желательно, чтобы каждый эксперт получил примерно 25% нагрузки. Это улучшает обучение, ускоряет специализацию, повышает устойчивость и снижает переобучение одного эксперта. Это важно и в контексте рекомендательных систем. Так как данные сильно несбалансированы, имеют вырожденное распределение, содержат пользователей с разной активностью. Без балансировки модель может начать пропускать большинство пользователей через одного эксперта, редкие паттерны останутся невыученными, а качество на холодном старте упадет. Балансировка же заставляет модель выделить отдельные эксперты под разные типы поведения и эффективнее разделять шаблоны пользовательской активности. Формально штраф на каждого эксперта можно представить так. Сначала рассчитывается средняя активация эксперта:

$$\bar{g}_k = \frac{1}{B} \sum_{i=1}^B g_k(x_i),$$

где B – размер батча для обучения модели, g_k – гейтинг функция k -го эксперта, а x_i – вход эксперта. Далее оценивается сам штраф:

$$L_{balance} = K \sum_{k=1}^K (\bar{g}_k - \frac{1}{K})^2,$$

где K – число экспертов. Итоговая функция потерь выглядит так:

$$L = \lambda L_{balance} + L_{bpr},$$

где λ – регулировочный коэффициент, обычно принимаемый в диапазоне от 0.01 до 0.1.

Необходимо отметить, что *MoE* снижает смещение (*bias*) модели, позволяя экспертам специализироваться. Однако увеличение числа экспертов повышает дисперсию (*variance*). Таким образом, есть некоторое оптимальное значение K , обеспечивающее баланс между смещением и дисперсией.

Существуют и более продвинутые *MoE* архитектуры. Например, в мультизадачных рекомендательных системах, где одновременно предсказываются вероятность клика (*CTR*), вероятность покупки (*CVR*), вероятность добавления в корзину, используется архитектура *Multi-gate MoE (MMoE)* [4]. В *MMoE* каждая задача имеет собственный гейтинг:

$$f^{(t)}(x) = \sum_{k=1}^K g_k^{(t)}(x) E_k(x),$$

где t – номер задачи, а $g_k^{(t)}$ – гейтинг функция задачи t . Таким образом, эксперты являются общими, но маршрутизация различается для каждой цели оптимизации. Преимуществом архитектуры является эффективное разделение общих и специфичных признаков, а также уменьшение конфликта градиентов между задачами [4].

В иерархической *MoE* гейтинг выполняется в несколько уровней:

$$f(x) = \sum_k g_k^{(1)}(x) (\sum_j g_{k,j}^{(2)}(x) E_{k,j}(x)),$$

где k – обозначает принадлежность к группе экспертов верхнего уровня;

j – индекс эксперта группы второго уровня;

$g_k^{(1)}(x)$ – гейтинг функция k ветки первого уровня;

$g_{k,j}^{(2)}(x)$ – вероятность выбора j эксперта внутри ветки k .

Иерархическая *MoE* делает маршрутизацию в два этапа. Этап 1 это условно грубая сегментация. Гейтинг первого уровня определяет к какому крупному сегменту относится вход. Например, короткая сессия пользователя или длинная сессия, популярные товары или редко покупаемые, новый пользователь или постоянный клиент. На этапе 2 происходит более тонкая специализация. Внутри выбранного крупного блока второй гейтинг выбирает более узкого эксперта. Например, внутри короткой сессии это поведенческий эксперт под импульсивные клики или эксперт под поиск по бренду. Плюсом данной модели является то, что можно иметь много экспертов, и они организованы в группы, которые проще обучать и интерпретировать.

Экспериментальная часть работы направлена на эмпирическую проверку гипотезы о том, что интеграция архитектуры *Mixture of Experts* в последовательную рекомендательную модель позволяет повысить качество персонализированных рекомендаций за счёт адаптивной специализации параметров экспертов. В качестве исходного набора данных использован публичный датасет *MovieLens-1M* [5], содержащий около 1 млн взаимодействий пользователей с объектами (фильмами). Датасет включает идентификаторы пользователей, идентификаторы объектов, рейтинги, временные метки взаимодействий. Данные были преобразованы в формат упорядоченных по времени последовательностей взаимодействий каждого пользователя. Пользователи с менее чем 5 взаимодействиями были исключены для обеспечения корректности последовательного моделирования. Предобработка включала следующие этапы: сортировка взаимодействий каждого пользователя по временной метке, усечение последовательностей до максимальной длины $T_{max}=50$, формирование обучающих

примеров в формате предсказания следующего купленного товара. Разделение данных на обучающую и тестовую выборки выполнялось по временному признаку – первые 80% взаимодействий каждого пользователя использовались для обучения, последние 20% – для тестирования. Таким образом обеспечивалась реалистичная постановка задачи прогнозирования будущих действий. В качестве базовой модели использовалась последовательная архитектура типа трансформер. Эмбединги объектов размерности $d=128$. Использовалось позиционное кодирование признаков. Нейросеть состояла из 2 блоков *multi-head self-attention* и 2 *feed-forward* слоев. Механизм *self-attention* реализован с числом голов $h=4$; размерностью одной головы $d_k=d/h=32$. МоЕ-блок был встроен в предсказательный слой модели. После формирования пользовательского представления z_u применялась смесь экспертов. $K \in \{2,4,6,8\}$ — число экспертов. Для предотвращения схлопывания экспертов применялась балансирующая регуляризация с коэффициентом $\lambda = 0.05$. Обучение проводилось с использованием: оптимизатора Adam, learning rate = 0.001; batch size = 256, числа эпох – 50. Для оценки качества рекомендаций использовались стандартные метрики ранжирования: *Hit@K*, *NDCG@K*, *MRR*. K принималось равным 10. Метрики рассчитывались следующим образом. Для каждого пользователя формировался список из 100 кандидатов (1 положительный + 99 случайных негативов). Вычислялась позиция истинного объекта. Рассчитывались метрики по формальным определениям. Для сопоставления результатов использовались: модель популярности — рекомендации по частоте; *Matrix Factorization (MF)*; базовый трансформер без МоЕ; трансформер + МоЕ. Результаты приведены в таблице 1.

Таблица 1. Результаты сравнения алгоритмов

Алгоритм	NDCG@10	Hit@10	MRR
Popularity	0.312	0.524	0.210
MF	0.601	0.781	0.392
Baseline transformer	0.672	0.814	0.428
Baseline + MoE(K=4)	0.694	0.839	0.451
Baseline + MoE(K=6)	0.702	0.847	0.460
Baseline + MoE(K=8)	0.699	0.844	0.458

Полученные результаты демонстрируют устойчивый прирост качества при использовании МоЕ. Прирост *NDCG@10* составил 4–5%. Улучшение *MRR* свидетельствует о более высокой точности ранжирования.

Анализ холодного старта был проведен для пользователей с историей менее 5 взаимодействий. Результаты приведены в таблице 2.

Таблица 2. Анализ моделей на холодном старте

Алгоритм	NDCG@10
Baseline transformer	0.341
Baseline + MoE(K=6)	0.387

МоЕ демонстрирует лучшие показатели при малом объёме данных.

Результаты проведённых экспериментов показывают, что интеграция архитектуры Mixture of Experts в последовательную рекомендательную модель позволяет добиться устойчивого улучшения качества персонализированных рекомендаций по сравнению с базовой трансформерной архитектурой. В частности, наблюдается прирост метрик *NDCG@10*, *Hit@10* и *MRR*, что свидетельствует о более точном ранжировании релевантных объектов и лучшей адаптации модели к различным поведенческим шаблонам пользователей.

Библиографический список

1. Яндекс. Введение в рекомендательные системы [Электронный ресурс] // Яндекс Учебник. – URL: <https://education.yandex.ru/handbook/ml/article/intro-recsys> (дата обращения: 10.01.2026)
2. Kang W.C., McAuley J. Self-Attentive Sequential Recommendation // Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM). – 2018. – P. 197–206. doi:10.1109/ICDM.2018.00035

3. Shazeer N., Mirhoseini A., Maziarz K., Davis A., Le Q.V., Hinton G., Dean J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer [Электронный ресурс] // International Conference on Learning Representations (ICLR). – 2017. – URL: <https://arxiv.org/abs/1701.06538> (дата обращения: 15.01.2026).
4. Ma J., Zhao Z., Yi X., Chen J., Hong L., Chi E. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. – 2018. – P. 1930–1939. doi:10.1145/3219819.3220007
5. McAuley J. Amazon Product Data [Электронный ресурс] // University of California San Diego. – URL: <https://jmcauley.ucsd.edu/data/amazon/> (дата обращения: 12.02.2026).

УДК 004.822; ГРНТИ 28.23.27

АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ УПРАВЛЕНИЯ ГРУППОЙ БПЛА

Е.А. Логунова

Московский политехнический университет,
Российская Федерация, Москва, e.a.logunova@mospolytech.ru

Аннотация. В статье представлен обзор современных методов машинного обучения (МО), применяемых для решения задач автономного управления группами беспилотных летательных аппаратов (БПЛА). Рассматривается переход от традиционных методов управления к интеллектуальным подходам, способным эффективно работать в динамических и недетерминированных средах. Проведена классификация алгоритмов по типу обучения (обучение с подкреплением, глубокое обучение, гибридные методы) и по решаемым задачам (построение роя, избегание препятствий, распределение задач). Выполнен сравнительный анализ ключевых подходов, включая multi-agent reinforcement learning (MARL), Deep Q-Networks (DQN) и их модификации, а также алгоритмы на основе теории игр и биологических принципов (рой). Особое внимание уделено вопросам масштабируемости, устойчивости связи и переносу обученных политик (transfer learning) в реальные условия.
Ключевые слова: беспилотные летательные аппараты (БПЛА), групповое управление, машинное обучение, многоагентные системы, обучение с подкреплением, глубокое обучение с подкреплением, управление формацией, автономная навигация.

MACHINE LEARNING ALGORITHMS FOR UAV SWARM CONTROL

Е.А. Logunova

Moscow Polytechnic University,
Russian Federation, Moscow, e.a.logunova@mospolytech.ru

Abstract. The paper presents an overview of modern machine learning (ML) methods applied to solve the problems of autonomous control for groups of unmanned aerial vehicles (UAVs). The transition from traditional control methods to intelligent approaches capable of operating effectively in dynamic and non-deterministic environments is considered. A classification of algorithms is carried out by the type of learning (reinforcement learning, deep learning, hybrid methods) and by the tasks being solved (swarm formation, obstacle avoidance, task allocation). A comparative analysis of key approaches is performed, including multi-agent reinforcement learning (MARL), Deep Q-Networks (DQN) and their modifications, as well as algorithms based on game theory and biological principles (swarm). Special attention is paid to the issues of scalability, communication stability, and the transfer of learned policies (transfer learning) to real-world conditions.

Keywords: unmanned aerial vehicles (UAVs), swarm robotics, machine learning, multi-agent systems, multi-agent reinforcement learning (MARL), deep reinforcement learning, formation control, autonomous navigation.

Введение

Развитие технологий беспилотных летательных аппаратов (БПЛА) в последние годы приобрело значительное ускорение благодаря технологическому прогрессу и существенному снижению их стоимости [3]. Беспилотники находят применение в широчайшем спектре областей: от сельского хозяйства и мониторинга окружающей среды до логистики и реше-

ния оборонных задач [5, 6]. Особый интерес представляет использование групп или роев БПЛА, что позволяет решать задачи, непосильные для одного аппарата: масштабный мониторинг территорий, поисково-спасательные операции в труднодоступных районах, совместная транспортировка грузов и обеспечение временного покрытия сетей связи [1, 9].

Однако управление группой БПЛА сталкивается с рядом фундаментальных вызовов. Традиционные методы, такие как прямое ручное управление, программируемые полетные задания или классические регуляторы (например, ПИД), плохо масштабируются, требуют постоянного присутствия оператора и неэффективны в динамически меняющихся и неопределенных средах [1, 6]. Это обуславливает необходимость перехода к интеллектуальным, автономным системам управления, способным адаптироваться к изменениям обстановки в реальном времени.

Машинное обучение (МО), и в особенности его раздел «обучение с подкреплением» (Reinforcement Learning – RL), предлагает мощный фреймворк для решения этих проблем. RL позволяет агентам (БПЛА) обучаться оптимальному поведению через взаимодействие со средой и друг с другом, максимизируя некоторую функцию награды [5]. Интеграция RL с глубокими нейронными сетями (Deep RL — DRL) открыла возможности для обработки многомерных входных данных (например, изображений с камер) и работы в непрерывных пространствах действий [1]. Для многоагентных сценариев был разработан целый класс методов Multi-Agent Reinforcement Learning (MARL), который и является основой для современных систем управления группами БПЛА [6].

Цель данной статьи – систематизировать и проанализировать существующие алгоритмы машинного обучения для управления группами БПЛА, выявить их сильные и слабые стороны, а также обозначить перспективные направления дальнейших исследований.

Классификация методов управления группой БПЛА

Прежде чем перейти к алгоритмам МО, представим классификацию самих методов управления БПЛА по уровню автономности. Как отмечают [6], можно выделить три основные категории (рис. 1).



Рис. 1. Методы управления БПЛА

Ручное управление полностью контролируется оператором. Прост в реализации для одиночных аппаратов, но неприменим для управления роем из-за высокой когнитивной нагрузки и ограничений каналов связи.

При *полуавтоматическом управлении* БПЛА выполняют часть функций автономно (например, удержание высоты или возврат на базу), но ключевые решения принимает человек. Является переходным этапом к полной автономии.

При *полностью автономном управлении* системы, способные выполнять миссию без вмешательства человека, адаптируясь к изменениям среды. Именно здесь методы машинного обучения находят свое основное применение [10].

В рамках автономного управления группами БПЛА решаются следующие ключевые задачи [1, 3] (рис. 2).



Рис. 2. Вызовы, возникающие при организации группового взаимодействия БПЛА

Формирование роя – объединение БПЛА в группы и поддержание заданного строя.

Навигация и планирование траекторий – построение безопасных и эффективных маршрутов до целей с учетом статических и динамических препятствий.

Избегание столкновений – критически важная функция как для предотвращения столкновений между БПЛА, так и с препятствиями.

Распределение задач – назначение различных подзадач (например, слежение за конкретной целью, патрулирование зоны) между членами группы.

Управление ресурсами – оптимизация использования энергии, каналов связи и вычислительных мощностей.

Алгоритмы машинного обучения для управления группой БПЛА

Современные методы МО, применяемые для решения перечисленных задач, можно разделить на несколько основных категорий (рис. 3, табл. 1).

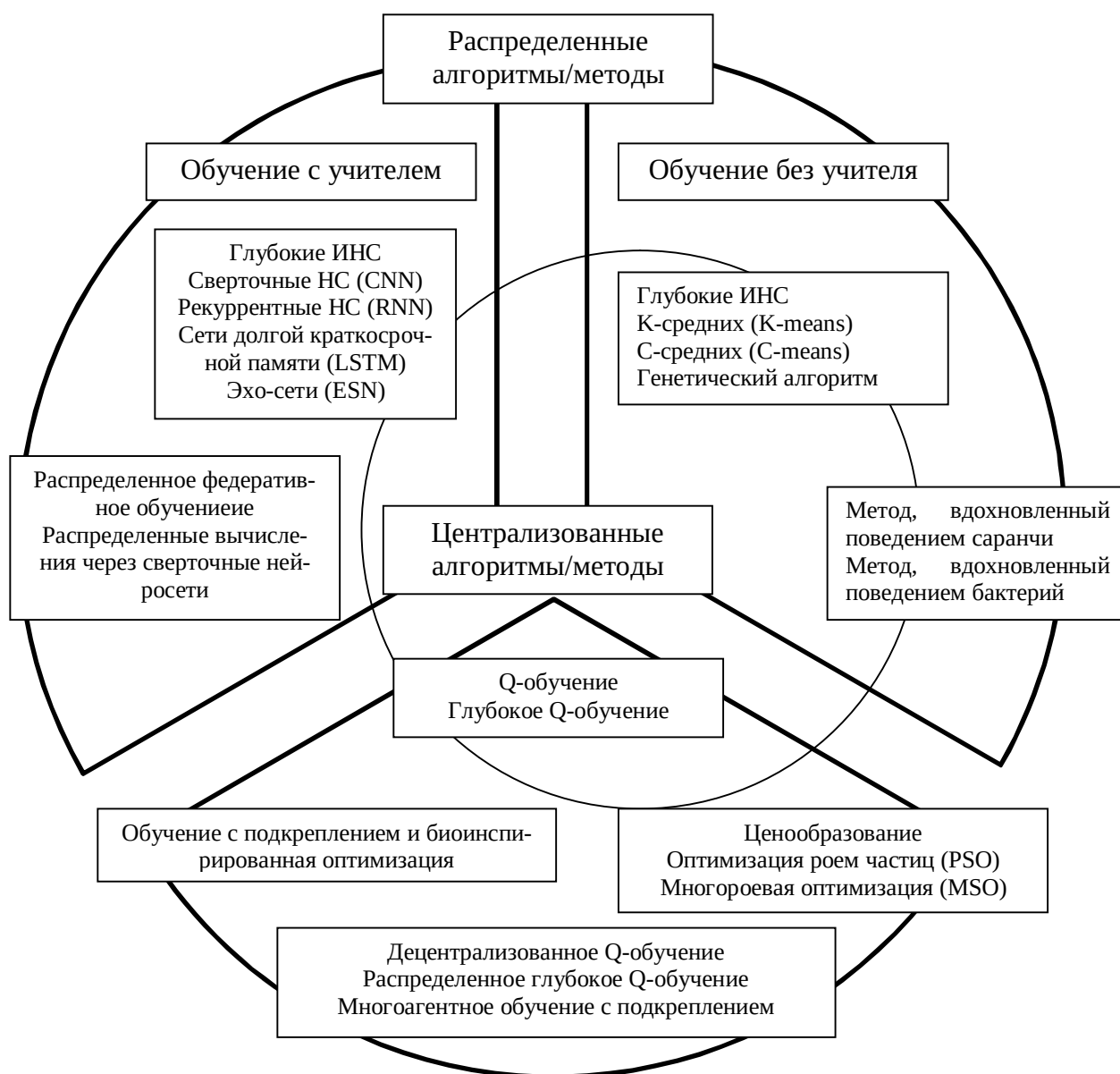


Рис. 3. Основные методы МО

Таблица 1. Основные методы машинного обучения (Ц – централизованный, Р – распределенный)

Метод машинного обучения	Тип обучения	Подход	Основные применения в задачах БПЛА
Глубокое обучение	С учителем / без учителя	Ц/Р	Распределение ресурсов
Сверточная нейросеть	С учителем	Ц	Кластеризация, распределение ресурсов
Рекуррентная нейросеть	Без учителя	Ц	Распределение задач
Сети долгой краткосрочной памяти	С учителем	Ц	Развертывание
Федеративное обучение	С учителем	Ц/Р	Распределение ресурсов
Q-обучение, Обучение с подкреплением	Обучение с подкреплением	Ц/Р	Кластеризация, формирование роя, развертывание, позиционирование, распределение ресурсов
Глубокое Q-обучение, Глубокое обучение с подкреплением	Обучение с подкреплением	Ц/Р	Формирование роя, развертывание, позиционирование, распределение ресурсов, распределение задач
Оптимизация роя частиц	Биоинспирированный	Р	Кластеризация, маршрутизация, позиционирование, распределение задач
Генетический алгоритм	Биоинспирированный	Ц	Кластеризация, развертывание, распределение задач
Метод К-средних	Без учителя	Ц	Кластеризация, развертывание, распределение ресурсов, распределение задач
Метод ценообразования	Без учителя	Р	Распределение ресурсов

Классические и биоинспирированные алгоритмы

До широкого распространения глубокого обучения для оптимизации поведения роя использовались методы, вдохновленные природой. Наиболее популярным является метод роя частиц (Particle Swarm Optimization – PSO). В контексте БПЛА, PSO используется для поиска оптимального расположения аппаратов в пространстве, например, для максимизации покрытия территории при мониторинге или для оптимизации траекторий в работе [9]. PSO моделирует поведение стаи птиц или косяка рыб, где каждая частица (БПЛА) корректирует свое положение на основе собственного лучшего опыта и лучшего опыта своих соседей.

Другие биоинспирированные подходы включают:

- Алгоритм пчелиной колонии (Bee Colony Algorithm) используется для эффективного распределения задач и исследования территории, имитируя поведение пчел-разведчиков и сборщиц нектара [9].
- Муравьиный алгоритм (Ant Colony Optimization – ACO) применяется для задач планирования пути, где «муравьи» оставляют феромоны, указывая на оптимальные маршруты [3].
- Генетические алгоритмы (Genetic Algorithms – GA) используются для оптимизации параметров управления и планирования миссий, оперируя популяцией решений и применяя операторы скрещивания и мутации [3].

Эти методы просты в реализации и не требуют предварительного обучения, однако их эффективность может снижаться в сложных, высокодинамических средах из-за проблем сходимости и отсутствия механизма обучения на прошлом опыте [6].

Обучение с подкреплением (Reinforcement Learning – RL)

RL стало доминирующим подходом для задач автономного управления. В основе RL лежит взаимодействие агента со средой, описываемое марковским процессом принятия решений (MDP). Агент в состоянии ***st*** выполняет действие ***at***, получает награду ***rt*** и переходит в новое состояние ***st + 1***. Цель агента – максимизировать кумулятивную дисконтированную награду (Azar et al., 2021). Для групп БПЛА этот процесс обобщается до много-

агентного обучения с подкреплением (MARL), где среда становится нестационарной с точки зрения отдельного агента, так как другие агенты также обучаются и меняют свое поведение [6].

На рисунке 4 показана схема взаимодействия в типовом алгоритме одноагентного обучения с подкреплением.

На рисунке 5 показана схема взаимодействия в типовом алгоритме многоагентного обучения с подкреплением.

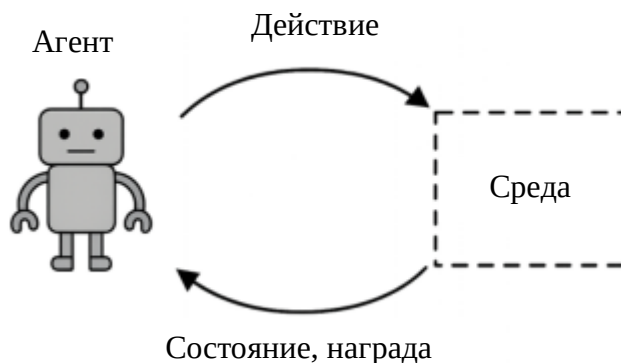


Рис. 4. Схематичное представление одноагентного обучения с подкреплением

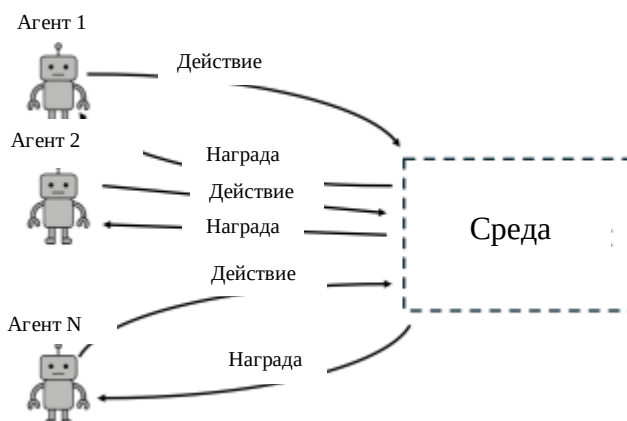


Рис. 5. Схематичное представление многоагентного обучения с подкреплением

MARL-алгоритмы можно классифицировать по типу обучения на value-based (метод ценности – оценивают ценность состояний/действий) и policy-based (метод политики – напрямую оптимизируют политику).

Методы ценности (Value-Based)

Эти методы стремятся оценить функцию ценности $Q(s, a)$ (Q-функцию), которая показывает ожидаемую будущую награду за выполнение действия a в состоянии s . Оптимальная политика затем извлекается из Q-функции (выбирается действие с максимальным Q).

– Deep Q-Networks (DQN)

Комбинация Q-обучения с глубокими нейросетями для аппроксимации Q-функции. В многоагентном контексте (MADQN) каждый агент обучает свою DQN. Применяется для задач, где пространство действий дискретно, например, перемещение по сетке или выбор типа задачи [1, 13]. Однако независимое обучение агентов усугубляет проблему нестационарности среды.

– QMIX

Более продвинутый алгоритм, использующий централизованное обучение с децентрализованным исполнением (CTDE). Он обучает разделимую Q-функцию, где глобальная Q_{tot} является монотонной комбинацией индивидуальных Q-функций агентов. Это позволяет агентам координировать свое поведение во время обучения, но действовать автономно на основе своих локальных наблюдений [6]. QMIX хорошо зарекомендовал себя в симуляционных средах типа StarCraft, но его монотонное ограничение может не охватывать все типы кооперативного поведения.

Методы политики (Policy-Based)

Эти методы оптимизируют непосредственно политику $\pi\theta(a | s)$, представленную параметрической моделью (например, нейросетью). Они особенно эффективны для непрерывных пространств действий, таких как управление скоростью и направлением полета БПЛА [2].

– Deep Deterministic Policy Gradient (DDPG)

Алгоритм типа "actor-critic", где актер (actor) учится политике, а критик (critic) оценивает ее. Его многоагентная версия – MADDPG – стала одним из стандартов MARL. MADDPG использует централизованных критиков, которые во время обучения имеют доступ к наблюдениям и действиям всех агентов, что позволяет каждой паре актер-критик обучаться в более стационарной обстановке. Во время исполнения актеры используют только локальные наблюдения. MADDPG успешно применялся для задач совместной навигации, целераспределения и уклонения от препятствий [6, 12].

– Trust Region Policy Optimization (TRPO)

Алгоритм, обеспечивающий стабильность обучения путем ограничения размера обновления политики с помощью KL-дивергенции. Однако он сложен в реализации и требует больших вычислительных ресурсов [2].

– Proximal Policy Optimization (PPO)

Упрощенная и более эффективная версия TRPO. PPO использует «клиппированную» целевую функцию, которая штрафует за слишком большие изменения политики. MAPPO, его многоагентное расширение, показывает отличные результаты в кооперативных сценариях, часто превосходя MADDPG по стабильности и скорости сходимости [6]. PPO широко применяется для обучения как одиночных дронов [7], так и целых роев [1, 8].

Сравнение PPO, DDPG и TRPO показывает, что PPO обеспечивает наилучший баланс между скоростью сходимости, стабильностью и вычислительной эффективностью, что делает его предпочтительным выбором для многих задач управления БПЛА [2].

Глубокое обучение для извлечения признаков и навигации

Помимо RL, методы глубокого обучения играют ключевую роль в обработке сенсорной информации. Сверточные нейронные сети (CNN) широко используются для анализа изображений с камер БПЛА [4]. Это необходимо для:

- Распознавания объектов – идентификация целей, препятствий, посадочных площадок.
- Классификации сцен – определение типа местности (лес, поле, город) для адаптации стратегии навигации.
- Семантической сегментации – попиксельная классификация изображения для точного построения карты местности и выделения проходимых зон [10].

Интеграция CNN с RL (Deer RL) позволяет создавать системы, которые напрямую принимают решения на основе «сырых» визуальных данных. Например, в работе [1] обученные политики используют информацию о положении и векторах направления для выполнения задач поиска и слежения.

Федеративное обучение (*Federated Learning — FL*)

Относительно новое направление, объединяющее FL и MARL – Multi-Agent Federated Reinforcement Learning (MAFRL). Этот подход особенно актуален для роев БПЛА, где важны конфиденциальность данных и пропускная способность каналов связи. В MAFRL каждый БПЛА обучает свою локальную модель на собранных данных и передает на центральный сервер не сами данные, а только обновления весов модели (градиенты). Сервер усредняет их и рассылает улучшенную глобальную модель обратно агентам [6]. Это позволяет масштабировать систему и сохранять конфиденциальность, хотя и требует решения проблем, связанных с гетерогенностью данных и асинхронностью обновлений.

Сравнительный анализ и тенденции

Основные современные тенденции [2, 6] приведены в таблице 2 и заключаются в следующем:

- Переход к CTDE (Centralized Training Decentralized Execution): Это доминирующая парадигма, позволяющая обучать сложное кооперативное поведение в симуляторе, а затем развертывать легковесные политики на реальных дронах.
- Гибридизация подходов: Комбинация модельных и безмодельных методов RL, использование графовых нейронных сетей (GNN) для учета топологии роя.
- Учет реальных ограничений: Все больше исследований фокусируются на энергоэффективности, устойчивости к отказам и работе в условиях ограниченной связи [1, 11].

Таблица 2. Сравнительный анализ МО

Методология	Ключевая идея	Применение в группах БПЛА	Преимущества	Недостатки
Биоинспирированные (PSO, ACO, GA)	Имитация природных процессов	Оптимизация покрытия, планирование пути, распределение задач	Простота, не требуют обучения, понятная интерпретация	Плохая адаптация к динамике, проблемы сходимости
Value-Based MARL (QMIX, DQN)	Оценка функции ценности действия	Кооперативные задачи с дискретными действиями	Эффективны в дискретных пространствах, QMIX улучшает координацию	Монотонные ограничения QMIX, проблемы масштабируемости у DQN
Policy-Based MARL (MADDPG, MAPPO)	Прямая оптимизация политики	Задачи с непрерывным управлением (навигация, слежение)	Работа с непрерывными действиями, MAPPO стабилен и эффективен	Требовательны к настройке гиперпараметров, высокие вычислительные затраты на обучение
Federated MARL	Обучение без централизации данных	Распределенные рои с ограниченной связью и требованиями приватности	Масштабируемость, сохранение конфиденциальности	Сложности с гетерогенностью и асинхронностью, более медленная сходимость

Открытые проблемы и вызовы

Несмотря на значительный прогресс, существует ряд проблем, требующих дальнейшего исследования [3, 6]:

- Масштабируемость – обучение и координация роев из сотен и тысяч агентов остается крайне сложной задачей.

- Переход из симуляции в реальность – политики, обученные в симуляторах, часто неэффективны в реальном мире из-за несоответствия динамики, сенсорного шума и задержек.
- Устойчивость к сбоям и атакам – обеспечение отказоустойчивости и безопасности в условиях потери связи или выхода из строя отдельных агентов.
- Объяснимость – понимание причин, по которым обученная политика принимает те или иные решения, критически важно для сертификации систем в гражданской авиации.
- Энергоэффективность – алгоритмы обучения должны учитывать жесткие ограничения по энергопотреблению бортовых компьютеров.

Заключение

Управление группами БПЛА представляет собой сложную и междисциплинарную проблему, эффективное решение которой невозможно без применения современных методов машинного обучения. Обучение с подкреплением, особенно в его многоагентной и глубокой формах (MARL, DRL), предлагает мощный инструмент для создания автономных, адаптивных и координированных роев.

Алгоритмы на основе ценностей (QMIX) и политик (MADDPG, MAPPO) успешно решают задачи от планирования траекторий до распределения задач, при этом MAPPO в последних исследованиях демонстрирует наилучший баланс производительности и стабильности. Биоинспирированные методы остаются полезными для оптимизационных задач, а федеративное обучение открывает новые горизонты для масштабируемых и приватных систем.

Несмотря на впечатляющие успехи, в основном достигнутые в симуляционных средах, ключевым вызовом остается переход к надежным, безопасным и энергоэффективным решениям, пригодным для реальной эксплуатации. Дальнейшие исследования должны быть направлены на преодоление разрыва между симуляцией и реальностью, обеспечение отказоустойчивости и разработку методов, позволяющих роям БПЛА работать в сложных, неструктурированных и динамичных условиях, выполняя критически важные миссии.

Библиографический список

1. Arranz, R., Carramiñana, D., de Miguel, G., Besada, J. A., & Bernardos, A. M. (2023). Application of Deep Reinforcement Learning to UAV Swarming for Ground Surveillance. *Sensors*, 23(21), 8766.
2. Azar, A. T., Koubaa, A., Ali Mohamed, N., Ibrahim, H. A., Ibrahim, Z. F., Kazim, M., ... & Casalino, G. (2021). Drone deep reinforcement learning: A review. *Electronics*, 10(9), 999.
3. Azoulay, R., Haddad, Y., & Reches, S. (2021). Machine Learning Methods for UAV Flocks Management-A Survey. *IEEE Access*, 9, 139146-139175.
4. Carrio, A., Sampedro, C., Rodriguez-Ramos, A., & Campoy, P. (2017). A review of deep learning methods and applications for unmanned aerial vehicles. *Journal of Sensors*, 2017(1), 3296874.
5. Choi, S. Y., & Cha, D. (2019). Unmanned aerial vehicles using machine learning for autonomous flight; state-of-the-art. *Advanced Robotics*, 33(6), 265-277.
6. Ekechi, C. C., Elfouly, T., Alouani, A., & Khattab, T. (2025). A Survey on UAV Control with Multi-Agent Reinforcement Learning. *Drones*, 9(7), 484.
7. Koch, W., Mancuso, R., West, R., & Bestavros, A. (2019). Reinforcement learning for UAV attitude control. *ACM Transactions on Cyber-Physical Systems*, 3(2), 1-21.
8. Liu, Y., Liu, H., Tian, Y., & Sun, C. (2020). Reinforcement learning based two-level control framework of UAV swarm for cooperative persistent surveillance in an unknown urban area. *Aerospace Science and Technology*, 98, 105671.
9. Londikov, V. A., Lukanov, S. Y., & Timoshevskaya, O. Y. (2024). Development of an intelligent control algorithm for a group of unmanned aerial vehicles. *Computational Nanotechnology*, 11(2), 86-92.
10. Lukianov, S. Y., Khryshkevich, G. A., Gorelov, M. A., Timoshevskaya, O. Y., & Londikov, V. A. (2024). Разработка модели управления группой беспилотных летательных аппаратов с помощью глубокого обучения с подкреплением. In *Сборник трудов конференции* (стр. 75-83).

11. Ma, H., Yang, G., Sun, X., Qu, D., Chen, G., Jin, X., ... & Liu, X. (2024). Improved DRL-based energy-efficient UAV control for maximum lifecycle. *Journal of the Franklin Institute*, 361(15), 106718.
12. Qie, H., Shi, D., Shen, T., Xu, X., Li, Y., & Wang, L. (2019). Joint optimization of multi-UAV target assignment and path planning based on multi-agent reinforcement learning. *IEEE Access*, 7, 146264-146272.
13. Venturini, F., Mason, F., Pase, F., Chiariotti, F., Testolin, A., Zanella, A., & Zorzi, M. (2020). Distributed reinforcement learning for flexible uav swarm control with transfer learning capabilities. In *Proceedings of the 6th ACM Workshop on Micro Aerial Vehicle Networks, Systems, and Applications* (pp. 1-6).

УДК 004.89; ГРНТИ 27.41.17

ИНТЕЛЛЕКТУАЛЬНЫЕ МЕТОДЫ ФОРМИРОВАНИЯ ПЕРСОНАЛИЗИРОВАННЫХ РЕКОМЕНДАЦИЙ ПРИ ВЫБОРЕ ОДЕЖДЫ С УЧЁТОМ МЕТЕОРОЛОГИЧЕСКИХ УСЛОВИЙ

А.П. Кирсанов, А.В. Крошили

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, sanya.kirsanov2003@yandex.ru*

Аннотация. В статье рассматриваются интеллектуальные методы формирования персонализированных рекомендаций в информационных системах подбора одежды базового гардероба с учётом метеорологических условий. Особое внимание уделено использованию методов машинного обучения и гибридных алгоритмов для анализа погодных данных и пользовательских предпочтений. Описывается архитектура рекомендательной системы, источники входных данных и перспективы развития подобных решений.

Ключевые слова: интеллектуальные системы, персонализированные рекомендации, прогноз погоды, машинное обучение, анализ данных.

INTELLIGENT METHODS FOR GENERATING PERSONALIZED RECOMMENDATIONS IN CLOTHING SELECTION BASED ON WEATHER CONDITIONS

A.P. Kirsanov, A.V. Kroshilin

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, sanya.kirsanov2003@yandex.ru*

The summary. The article examines intelligent methods for generating personalized recommendations in information systems designed to select basic wardrobe items based on weather conditions. Special attention is paid to machine learning techniques and hybrid approaches for processing meteorological data and user preferences. The system architecture, data sources, and future development prospects are analyzed.

Keywords: intelligent systems, personalized recommendations, weather forecast, machine learning, data analysis.

В условиях высокой вариативности климатических факторов и увеличения скорости повседневной жизни выбор одежды становится задачей, требующей учета множества параметров. Существенные колебания температуры воздуха, изменение влажности, усиление ветра и вероятность атмосферных осадков оказывают непосредственное влияние на комфорт человека и его самочувствие [9]. В связи с этим возрастает интерес к интеллектуальным информационным системам, способным автоматизировать процесс подбора одежды с учетом текущих и прогнозируемых метеорологических условий.

Разработка систем поддержки принятия решений в области выбора одежды базового гардероба представляет собой перспективное направление, объединяющее методы анализа данных, искусственного интеллекта и персонализированных рекомендаций. Основной целью таких систем является формирование обоснованных и адаптивных рекомендаций, учитывающих не только погодные условия, но и индивидуальные особенности пользователя, включая стиль, уровень активности и контекст использования одежды (деловая среда, повседневная деятельность, отдых).

Использование интеллектуальных рекомендательных систем позволяет сократить временные затраты на выбор одежды, повысить уровень комфорта и снизить риски, связанные с неблагоприятным воздействием окружающей среды. Кроме того, подобные системы способствуют более рациональному использованию гардероба за счет выявления оптимальных сочетаний элементов одежды и их соответствия текущим погодным условиям [1, 2].

Архитектура интеллектуальной системы персонализированных рекомендаций

Интеллектуальная система формирования рекомендаций при выборе одежды с учетом метеорологических условий строится на модульной архитектуре, обеспечивающей гибкость и масштабируемость решения. В состав системы входят следующие ключевые компоненты [3, 4]:

- *Модуль сбора метеорологических данных* – обеспечивает получение информации из открытых источников, таких как специализированные погодные API и сервисы мониторинга климата. Получаемые данные включают значения температуры воздуха, влажности, скорости ветра, вероятности осадков и других параметров, характеризующих состояние окружающей среды;
- *Модуль интеллектуального анализа погодных условий* – выполняет обработку входных данных и определяет значимость отдельных факторов с точки зрения комфортности ношения различных элементов одежды. На данном этапе используются методы машинного обучения, позволяющие учитывать нелинейные зависимости между погодными параметрами;
- *Модуль пользовательского профиля* – хранит сведения о предпочтениях пользователя, его стиле одежды, типе активности, а также информацию о составе базового гардероба. Данный модуль играет ключевую роль в обеспечении персонализации рекомендаций.
- *Хранилище данных гардероба* – содержит структурированную информацию о предметах одежды, их сезонности, материалах и эксплуатационных характеристиках.
- *Рекомендательный модуль* – осуществляет интеграцию результатов анализа погодных условий и пользовательского профиля, формируя итоговые предложения по выбору одежды.

Источники метеорологических данных и их обработка

Качество рекомендаций напрямую зависит от достоверности и полноты используемых погодных данных. Для построения интеллектуальных моделей применяются следующие источники информации:

- данные спутниковых наблюдений, используемые для оценки облачности и атмосферных процессов;
- измерения наземных метеостанций, обеспечивающие высокую точность локальных показателей;
- исторические климатические ряды, позволяющие выявлять сезонные закономерности;
- результаты численного моделирования атмосферы.

Перед использованием данные проходят этапы предварительной обработки, включающие фильтрацию шумов, нормализацию значений и восстановление пропущенных параметров. Это позволяет повысить устойчивость интеллектуальных моделей и снизить влияние аномальных значений.

Интеллектуальные методы формирования рекомендаций

Формирование персонализированных рекомендаций осуществляется с применением различных подходов машинного обучения. В системе используются методы обучения с учителем, без учителя и гибридные алгоритмы.

Обучение с учителем применяется для прогнозирования степени комфортности одежды при заданных погодных условиях. В качестве моделей используются регрессионные алгоритмы, деревья решений и нейронные сети.

Методы обучения без учителя применяются для кластеризации пользователей и элементов гардероба, что позволяет выявлять группы с похожими предпочтениями и поведенческими паттернами. Используются алгоритмы кластеризации и методы снижения размерности данных.

Для повышения качества рекомендаций применяются гибридные подходы, сочетающие контентную и коллаборативную фильтрацию [5, 6]. Это позволяет учитывать как индивидуальные предпочтения пользователя, так и опыт других пользователей системы.

Использование нейросетевых моделей

Глубокие нейронные сети находят широкое применение в анализе метеорологических данных и формировании адаптивных рекомендаций. Сверточные нейронные сети используются для интерпретации визуальных данных, таких как спутниковые изображения, а рекуррентные нейронные сети и трансформеры — для анализа временных рядов погодных параметров.

Нейросетевые модели позволяют учитывать сезонные изменения, прогнозировать краткосрочные и долгосрочные погодные условия, а также адаптировать рекомендации под индивидуальные особенности пользователя на основе истории его взаимодействия с системой.

Дальнейшее развитие интеллектуальных систем рекомендаций связано с интеграцией новых источников данных и технологий. Перспективными направлениями являются использование информации с носимых устройств, внедрение технологий дополненной реальности для виртуальной примерки одежды, а также развитие самообучающихся алгоритмов, способных учитывать изменение предпочтений пользователя во времени. Несмотря на преимущества интеллектуальных рекомендательных систем, существует ряд проблем [7, 8], связанных с обработкой больших объемов данных, объединением разнородных источников информации и ограниченной точностью метеорологических прогнозов. Существенным вызовом остается задача оптимизации вычислительных ресурсов и обеспечения устойчивости моделей при изменении внешних условий.

Математическая постановка задачи формирования рекомендаций

Задача формирования персонализированных рекомендаций при выборе одежды может быть формализована как задача отображения множества входных параметров в множество рекомендуемых элементов гардероба. Пусть

$W = \{w_1, w_2, \dots, w_n\}$ – вектор метеорологических параметров (температура, влажность, скорость ветра, вероятность осадков и др.),

$U = \{u_1, u_2, \dots, u_n\}$ – вектор характеристик пользователя (предпочтения, стиль, уровень активности, контекст использования),

$G = \{g_1, g_2, \dots, g_k\}$ – множество элементов гардероба пользователя.

Требуется определить функцию рекомендаций

$$R: (W, U, G) \rightarrow G^*,$$

где $G^* \subseteq G$ – оптимальное подмножество элементов одежды, обеспечивающее максимальный уровень комфортности и соответствие пользовательским предпочтениям.

Оптимизационная задача может быть представлена в виде максимизации функции полезности:

$$G^* = \arg \max_{S \in G} F(W, U, S),$$

где F – функция оценки комфортности и релевантности выбранного набора одежды, учитывающая погодные условия, пользовательские предпочтения и совместимость элементов гардероба.

В практических реализациях функция F аппроксимируется с использованием моделей машинного обучения, обучаемых на исторических данных пользовательских выборов и метеорологических параметров. Такая постановка задачи позволяет применять широкий спектр алгоритмов – от регрессионных моделей и ансамблевых методов до нейросетевых архитектур – обеспечивая адаптивность и масштабируемость интеллектуальной системы рекомендаций.

Оценка качества рекомендаций и метрики эффективности

Для объективной оценки качества интеллектуальной системы персонализированных рекомендаций необходимо использовать комплекс количественных и качественных показателей. Эффективность работы рекомендательного модуля может оцениваться с применением стандартных метрик машинного обучения и анализа рекомендаций.

К числу основных метрик относятся:

- **Accuracy** (точность классификации) — доля корректно предсказанных рекомендаций;
- **Precision** и **Recall** — показатели точности и полноты релевантных рекомендаций;
- **F1-мера** — гармоническое среднее точности и полноты;
- **MAE** и **RMSE** — средняя абсолютная и среднеквадратичная ошибка при прогнозировании уровня комфортности одежды;
- **NDCG** (Normalized Discounted Cumulative Gain) – показатель качества ранжирования рекомендаций.

Дополнительно применяется пользовательская оценка удовлетворенности рекомендациями, получаемая посредством обратной связи в интерфейсе системы. Интеграция явной (оценки, рейтинги) и неявной (время взаимодействия, частота выбора элементов) обратной связи позволяет реализовать механизм адаптивного обучения моделей.

Комплексная система оценки обеспечивает не только контроль качества, но и возможность динамической корректировки параметров интеллектуальных алгоритмов.

Контекстно-зависимая персонализация рекомендаций

Одним из ключевых направлений развития интеллектуальных систем подбора одежды является внедрение механизмов контекстно-зависимой персонализации. В отличие от классических рекомендательных систем, учитывающих ограниченный набор параметров, контекстно-ориентированные модели принимают во внимание дополнительные факторы:

- геолокацию пользователя;
- время суток;
- тип планируемой активности;
- социальный контекст (деловая встреча, прогулка, спортивное мероприятие);
- транспортный способ передвижения.

Учет контекста позволяет существенно повысить релевантность рекомендаций. Например, одинаковые погодные условия могут требовать различного набора одежды в зависимости от уровня физической активности пользователя. Для моделирования контекста применяются методы факторизации тензоров, графовые модели и байесовские сети.

Контекстная персонализация обеспечивает более точную адаптацию рекомендаций к реальным условиям эксплуатации одежды и повышает практическую ценность системы.

Вопросы безопасности и защиты персональных данных

Функционирование интеллектуальной системы персонализированных рекомендаций связано с обработкой персональных данных пользователя, включая сведения о его предпочтениях, геолокации и поведенческих характеристиках. В связи с этим особое значение приобретает обеспечение информационной безопасности.

К основным направлениям защиты данных относятся:

- использование методов шифрования при передаче и хранении информации;
- разграничение прав доступа к данным;
- анонимизация пользовательских профилей;
- минимизация объема собираемых данных;
- применение принципов privacy-by-design при проектировании архитектуры системы.

Дополнительно рекомендуется внедрение механизмов локальной обработки части пользовательских данных на клиентском устройстве, что снижает риски утечки информации. Соблюдение требований законодательства в области защиты персональных данных является обязательным условием внедрения подобных интеллектуальных решений в практическую деятельность.

Экономическая и социальная значимость внедрения системы

Разработка и внедрение интеллектуальных систем персонализированного подбора одежды имеет не только технологическую, но и социально-экономическую значимость.

С экономической точки зрения система способствует:

- рационализации использования гардероба;
- снижению импульсивных покупок;
- оптимизации ассортимента в сфере электронной коммерции;
- повышению уровня персонализации онлайн-сервисов.

С социальной позиции система повышает уровень комфорта и качества жизни пользователя, способствует формированию осознанного потребления и снижению избыточного использования ресурсов текстильной промышленности.

В перспективе подобные интеллектуальные решения могут интегрироваться в экосистемы «умного города», обеспечивая синхронизацию погодных сервисов, транспортной инфраструктуры и персонализированных цифровых помощников.

Библиографический список

1. Сафронов А.В. Искусственный интеллект и большие данные. — М.: Юрайт, 2022. — 368 с.
2. Крошилин А.В. Предметно-ориентированные информационные системы: учебное пособие / А.В. Крошилин, С.В. Крошилина, Г.В. Овечкин. — Москва: КУРС, 2023. — 176 с. — (Естественные науки).
3. Перепелкин Д.А., Попова А.А., Крошилин А.В., Крошилина С.В. Компьютерное моделирование процессов поддержки принятия решений врача-стоматолога на основе семантических сетей // Вестник РГРТУ. №89 - Рязань: РГРТУ, 2024. – 155 с. (127-140) DOI: 10.21667/1995-4565-2024-89-127-140
4. Котельников И.А., Мельников В.И. Анализ данных и машинное обучение в среде Python. — СПб.: Питер, 2021. — 400 с.
5. Крошилина С.В., Крошилин А.В., Саморукова О.Д., Жулева С.Ю. Задачи разработки систем медицинского назначения при выборе схемы медикаментозного лечения // Вестник РГРТУ. №88 - Рязань: РГРТУ, 2024. – 142 с. (106-114) DOI: 10.21667/1995-4565-2024-88-106-114
6. Овечкин Г.В., Крошилин А.В., Крошилина С.В., Саморукова О.Д. Моделирование процессов управления в организационных системах на основе теории нечетких когнитивных карт // Вестник РГРТУ. №91 - Рязань: РГРТУ, 2025. – 210 с.
7. Шарп Г. Введение в информационные системы: управление и разработка / Г. Шарп. — Москва: Вильямс, 2020. — 512 с. — ISBN 978-5-9908677-7-2. — EDN BBVIQZ.

8. Власов В. Г. Алгоритмы обработки данных и машинное обучение / В. Г. Власов. — Санкт-Петербург: Питер, 2019. — 352 с. — ISBN 978-5-4461-0921-8. — EDN PZDFYU.

9. Перепад температуры <https://weather.fandom.com/ru/> Перепад температуры

УДК 004.85; ГРНТИ 50.41.25

ПРИМЕНЕНИЕ ГРАФОВЫХ НЕЙРОННЫХ СЕТЕЙ С МЕХАНИЗМОМ ВНИМАНИЯ В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ

П.А. Попов, Н.И. Цуканова

*Рязанский государственный радиотехнический университет,
РФ, Рязань*

Аннотация. В данной работе исследуется применение графовых нейронных сетей с механизмом внимания (Graph Attention Networks – GAT) для построения рекомендательных систем. Особое внимание уделяется анализу влияния выбора исходных данных на качество итоговых рекомендаций. Показано, что использование механизма внимания позволяет динамически взвешивать влияние соседних узлов в графе «пользователь–объект взаимодействия». Проведен сравнительный анализ работы модели на классическом наборе данных Amazon Software, содержащем только бинарные связи, и на современном многоаспектном датасете VK-LSVD. Эксперименты демонстрируют, что обогащение графовой структуры дополнительными признаками взаимодействий позволяет GAT значительно повысить точность рекомендаций по сравнению с классическими графовыми сверточными сетями (GCN).

Ключевые слова: графовая нейронная сеть, рекомендательная система, механизм внимания, GAT, машинное обучение, качество данных

USING GRAPH ATTENTION NETWORKS FOR RECOMMENDER SYSTEMS

P.A. Popov, N.I. Tsukanova

*Ryazan State Radio Engineering University,
RF, Ryazan*

Abstract. This paper investigates the application of Graph Attention Networks (GAT) for building recommender systems. Special attention is paid to analyzing the impact of the choice of initial data on the quality of the final recommendations. It is shown that the use of an attention mechanism allows for dynamically weighting the influence of neighboring nodes in the user–item graph. A comparative analysis of the model's performance is conducted on the classic Amazon Software dataset, containing only binary links, and on the modern, multi-faceted VK-LSVD dataset. Experiments demonstrate that enriching the graph structure with additional interaction features enables GAT to significantly improve the accuracy of recommendations compared to classical Graph Convolutional Networks (GCN).

Keywords: graph neural network, recommendation system, attention mechanism, GAT, machine learning, data quality

Введение

Современные рекомендательные системы работают в условиях постоянного роста объемов данных и усложнения паттернов поведения пользователей. Традиционные методы коллаборативной фильтрации часто неспособны уловить тонкие контекстные зависимости, возникающие в процессе взаимодействия. Графовые нейронные сети (GNN) используют эффективный способ моделирования таких взаимосвязей, представляя данные в виде графа, узлы которого соответствуют пользователям и товарам, а ребра – фактам взаимодействия [1, 2].

Однако ключевым фактором, определяющим успех применения GNN, является не только архитектура модели, но и качество, а также структура исходного графа. Подавляющее большинство доступных датасетов (например, используемый здесь Amazon Software), представляют собой разреженные двудольные графы, где ребро лишь сигнализирует о фак-

те покупки или отзыва. Это ограничивает способность модели к обучению, поскольку игнорируется контекст взаимодействия: исследуется вопрос, почему товар был куплен, насколько интенсивно с ним взаимодействовали, при каких обстоятельствах.

Появление механизмов внимания в графовых нейросетях (GAT) [3] открыло новые возможности. GAT позволяет модели назначать веса ребрам, выявляя наиболее значимые связи. Но потенциал этого механизма раскрывается в полной мере только тогда, когда граф содержит неоднородную информацию. В данной статье исследуется, как обогащение данных о взаимодействиях (т. е. добавление дополнительной информации в признаки узлов) влияет на качество рекомендаций, сравнивая работу GAT на классическом датасете Amazon Software и на новом, детализированном датасете VK-LSVD.

Решение задачи построения рекомендательной системы

Процесс построения рекомендательной системы на основе графовых нейронных сетей включает в себя несколько ключевых этапов, начиная от подготовки данных и заканчивая генерацией рекомендаций.

1. Выбор и анализ набора данных – критически важный этап, определяющий верхнюю границу качества модели. Необходимо оценить количество взаимодействий, плотность графов, а также наличие признаков узлов и ребер, их значение и информативность.
2. Построение графа взаимодействий – преобразование исходных данных в графовую структуру.
3. Предобработка данных – при необходимости привести набор данных к требуемому виду, добавить необходимые признаки, нормализовать векторы.
4. Обучение GAT-модели – обучение графовой нейронной сети с механизмом внимания для получения обогащенных (контекстом) векторных представлений узлов.
5. Формирование рекомендаций – вычисление мер сходства между векторными представлениями пользователей и товаров для генерации персональных рекомендаций.

Важность выбора набора данных

Выбор набора данных является фундаментальным этапом, от которого зависит, сможет ли модель выявить реальные закономерности. Простые датасеты с бинарными взаимодействиями (купил/не купил) формируют разреженный граф, где каждое ребро имеет одинаковый вес. В таких условиях GAT, хоть и может обучать веса внимания и выдать релевантные рекомендации, но из-за недостатка информации не сможет найти скрытые (и более явные) связи.

В отличие от них, датасеты с богатой контекстной информацией, такие как VK-LSVD, предоставляют многомерную обратную связь: время просмотра, лайки, репосты, комментарии, тип платформы и геолокацию. Это позволяет строить не просто двудольный, а мультиграф, где каждое взаимодействие описывается вектором признаков. Именно в таких условиях механизм внимания GAT способен проявить свою силу, обучаясь различать, например, «поверхностный» просмотр и «глубокое» вовлечение, и соответствующим образом взвешивать их влияние на эмбединг (обучаемый вектор) пользователя.

Построение графа взаимодействий

Для датасета Amazon Software граф строится как стандартный двудольный граф: узлы пользователей соединяются с узлами товаров, если был оставлен отзыв. При этом отсутствуют признаки узлов, поэтому мы заполняем их начальными данными (единичным вектором) и обучаем эти векторы в процессе. Признаки ребер содержат оценку пользователем товара (1-5) и могут быть использованы при формировании коэффициентов внимания

Для датасета VK-LSVD подход к построению графа принципиально иной. Каждое ребро между пользователем и видео содержит богатый набор атрибутов: timespent (время просмотра), бинарные флаги like, share, bookmark, click_on_author, open_comments, а также контекстные признаки place, platform, agent. Это позволяет преобразовать простое ребро во взвешенное взаимодействие. При инициализации графа эти признаки могут быть использованы для расчета начальных коэффициентов внимания.

Обучение GAT-модели

В работе используется архитектура GATv2 как более выразительная версия оригинальной GAT. На каждом свёрточном слое вычисляются коэффициенты внимания α_{ij} для пар соседних узлов i и j по формуле (1):

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W}h_i \parallel \mathbf{W}h_j \parallel \Theta e_{ij}]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W}h_i \parallel \mathbf{W}h_k \parallel \Theta e_{ik}]))}$$

где h_i, h_j – признаки узлов, где $\mathbf{W}_h$ и Θ_e – обучаемые матрицы весов для узлов и ребер соответственно, $\mathbf{a}$ – обучаемый вектор внимания, $\parallel$ – операция конкатенации, $\mathcal{N}_i$ – множество соседей узла i .

Для повышения стабильности обучения используется многоголовое внимание (multi-head attention) [3, 5], при котором выходы нескольких независимых механизмов внимания агрегируются. Это позволяет искать важные признаки с нескольких сторон. Затем для каждого слоя применяется функция активации ELU и слой Dropout для регуляризации.

Экспериментальные исследования

Целью эксперимента являлось сравнение влияния структуры и насыщенности данных на качество работы GAT. Для этого были выбраны два набора данных: классический Amazon Software (разреженный, бинарные связи) и современный VK-LSVD (плотный, взаимодействия описаны многомерным вектором). Чтобы доказать, что использование GAT целесообразно, обучение для каждого набора проводилось на двух архитектурах – классической GCN (где вклад соседей определяется структурой графа, а не обучается динамически) и GAT (где важность соседа вычисляется динамически во время обучения).

Характеристики наборов данных

- Amazon Software содержит 375 тысяч пользователей, 210 тысяч товаров и 450 тысяч взаимодействий (отзывов). Граф крайне разрежен (плотность около 0.00057%). Единственный атрибут взаимодействия – рейтинг (1-5), который часто интерпретируется как вес ребра, но не как многомерный контекст.
- VK-LSVD предоставляет несколько подвыборок. Здесь используется выборка up0.001_ip0.001 с 10 тысячами пользователей и 19.6 тысячами видео, но с 47.9 миллионами взаимодействий, что дает высокую плотность (около 24%). Ключевое отличие – каждое взаимодействие описывается 11 признаками (включая timespent и 6 типов явной/неявной обратной связи). Это позволяет строить граф с богатой семантической связью.

Эксперименты проводились на двух архитектурах: базовая графовая сверточная сеть (GCN) и GAT (с 2 слоями, 4 головами внимания, размерностью эмбедингов 64). Обучение велось до 100 эпох с использованием функции потерь MSE. Для датасета VK-LSVD в модели GAT использовались признаки ребер, как показано в формуле (1) выше. Для набора Amazon Software, не имеющего богатых признаками векторов ребер эта часть опускалась как несущественная (т.к. есть всего лишь один признак оценки товара, и его использование не продемонстрировало улучшения качества рекомендаций)

Характеристики набора данных и моделей приведены в таблице 1.

Таблица 1. Характеристики экспериментального набора данных и модели

Характеристика	Amazon Software	VK-LSVD
Плотность графа	Крайне низкая (<0.001%)	Высокая (24% для выборки)
Атрибуты ребер	Только рейтинг (1-5)	11 признаков (timespent, лайки, репосты и др.)
Число пользователей	375147	10000
Число товаров/видео	210623	19600
Число взаимодействий	450 тысяч	47 миллионов
Потери для модели GCN(MSE)	0.1412	0.0979
Потери для модели GAT(MSE)	0.1059	0.0547
Лучшая потеря в тесте (MSE)	0.1059	0.0547

На разреженном Amazon Software GAT показала улучшение (0.1059) по сравнению с GCN (0.1412) за счет использования весов внимания. На насыщенном датасете VK-LSVD преимущество GAT становится еще более очевидным – значение функции потерь (0.0547) практически в два раза лучше, чем у GCN на том же датасете (так как GCN игнорирует полезную информацию, выраженную в признаках ребер). Это подтверждает, что GAT способна эффективно извлекать сигнал из многомерных данных о взаимодействиях, обучаясь игнорировать шум (например, случайные клики) и усиливать важные сигналы (лайки, большое время просмотра).

Эксперименты проводились на ПЭВМ с операционной системой Windows 11, на базе процессора AMD Ryzen 9 5950X 16-Core Processor с оперативной памятью объемом 16 Гб. Программа выполнялась в среде Python 3.11.

Выводы

Проведенное исследование доказывает, что прогресс в архитектурах графовых нейронных сетей, таких как внедрение механизма внимания в GAT, должен идти рука об руку с повышением качества входных данных. Использование "плоских" датасетов с бинарными взаимодействиями не позволяет раскрыть потенциал механизма внимания, который наиболее эффективен при работе с контекстно-обогащенными графами.

Датасет VK-LSVD, благодаря своей детальной проработке и включению множества аспектов пользовательского поведения, представляет собой идеальный полигон для моделей типа GAT. Эксперименты показали, что переход от простой фиксации факта взаимодействия к его более подробному описанию позволяет значительно улучшить точность предсказаний. Это открывает путь к созданию по-настоящему персонализированных рекомендательных систем, способных понимать не только то, что выбрал пользователь, но и почему и в каком контексте он это сделал.

Библиографический список

1. Zhou J., Cui G., Hu S. et al. Graph neural networks: A review of methods and applications Published at AI Open 2021
2. Wu Z., Pan S., Chen F. et al. A Comprehensive Survey on Graph Neural Networks
3. Мингалев А. В., Метод и алгоритмы обработки тепловизионных изображений на основе сверточных нейронных сетей для автоматического обнаружения заданных типов объектов // Вестник РГРТУ, 2025 № 91.
4. Veličković P., Cucurull G., Casanova A. et al. Graph Attention Networks // International Conference on Learning Representations (ICLR). 2018.
5. Демидова Л. А., Журавлев В. Е. Подход к анализу регулярных выражений с применением графовых нейронных сетей и контрастного обучения // Вестник РГРТУ, 2025 № 92.
6. Brody S., Alon U., Yahav E. How Attentive are Graph Attention Networks? // International Conference on Learning Representations (ICLR). 2022.

7. Цуканова Н. И., Александров В. В., Головкин Н. В., Шурыгина О. В. Применение искусственных нейронных сетей и машинного обучения к оценке качества коллективно-договорных актов в сфере образования // Вестник РГРТУ, 2023 № 86.
8. VK-LSVD: Large Short-Video Dataset. Hugging Face. URL: <https://huggingface.co/datasets/deepvk/VK-LSVD>

УДК 004.032.26; ГРНТИ 28.23.37

МЕТОД КОРРЕКТИРОВКИ ЦЕЛЕВЫХ МЕТОК ПРИ ОБУЧЕНИИ НЕЙРОННЫХ СЕТЕЙ ОЦЕНКИ КАЧЕСТВА ИЗОБРАЖЕНИЙ С УЧЁТОМ ВЗАИМОДЕЙСТВИЯ ДЕФЕКТОВ

В.А. Анохин

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, anohin.va@yandex.ru*

Аннотация. В работе рассматривается проблема обучения нейронных сетей для оценки качества изображений при наличии множественных дефектов. Предложен метод корректировки целевых меток, учитывающий взаимосвязь между параметрами деградации изображения. Проведён анализ соотношения размытия Гаусса и масштабирования при генерации обучающей выборки. Исследовано взаимодействие дефектов при последовательном применении. В заключении сделаны выводы по дальнейшему исследованию проблемы.

Ключевые слова: нейронные сети, оценка качества изображений, целевые метки, дефекты изображений, размытие Гаусса, масштабирование, взаимодействие дефектов.

METHOD OF TARGET LABEL CORRECTION IN TRAINING NEURAL NETWORKS FOR IMAGE QUALITY ASSESSMENT CONSIDERING DEFECT INTERACTION

V.A. Anokhin*

**Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, anohin.va@yandex.ru*

The summary. The paper considers the problem of training neural networks for image quality assessment in the presence of multiple defects. A method for correcting target labels is proposed, taking into account the relationship between image degradation parameters. An analysis of the ratio of Gaussian blur and scaling in the generation of the training sample is carried out. The interaction of defects during sequential application is investigated. In conclusion, conclusions are drawn on further investigation of the problem.

Keywords: neural networks, image quality assessment, target labels, image defects, Gaussian blur, scaling, defect interactions.

Введение

Широкое внедрение технологий глубокого обучения в задачи обработки визуальных данных актуализирует проблему создания надежных модулей объективной оценки качества изображений. Эффективность работы таких систем напрямую зависит от точности определения параметров деградации.

В предыдущих исследованиях [1] была предложена архитектура нейронной сети повышения качества изображения, включающая в себя модуль оценки качества входного изображения.

Данный модуль предсказывает:

1. Бинарную классификацию – определение принадлежности изображения к классу «качественное» или «дефектное». Данная задача решается с использованием релятивистской функции потерь, позволяющей модели сравнивать не абсолютные значения, а разность оценок между качественным и некачественным изображениями. Также данный выход можно использовать для генеративно-сопоставительного обуче-

ния (GAN). В этом случае модуль оценки качества выступает в роли дискриминатора, который анализирует качество выходного изображения и используется в качестве функции потерь.

2. Регрессию параметров дефектов – предсказание числовых значений, характеризующих силу различных типов искажений:

- Сила шума (стандартное отклонение);
- Степень JPEG-сжатия (качество от 0 до 100%);
- Коэффициент масштабирования (далее выход был объединен с предсказанием размытия);
- Сила размытия Гаусса (стандартное отклонение).

Такой подход позволяет нейронной сети целенаправленно обучаться определять тип и интенсивность дефектов, что способствует более эффективному устранению искажений в последующих слоях модели.

В ходе экспериментов было выявлено, что при достижении высокой точности бинарной классификации точность регрессии параметров дефектов оставалась достаточно низкой. Это означает, что модель успешно определяла наличие дефектов, но не могла точно определить какого они типа и силы.

Постановка задачи

Анализ причин показал, что основная проблема заключается в отсутствии учета взаимодействия дефектов при последовательном наложении на изображение. Из-за этого целевые метки становятся неточными и не отражают реальную «видимость» каждого дефекта в итоговом изображении.

Данная проблема связана со следующими факторами:

1. Взаимодействие дефектов – при последовательном применении нескольких типов искажений более «сильные» дефекты могут маскировать или подавлять проявления других.
2. Взаимосвязь параметров деградации – некоторые дефекты (например, масштабирование и размытие) имеют схожий визуальный эффект, что приводит к неоднозначности в определении их индивидуальных вкладов.
3. Нелинейность целевых меток – при комбинировании дефектов в одной области изображения их метки складываются нелинейно, что усложняет обучение модели регрессии.

Задачей данной работы является исследование метода корректировки целевых меток при обучении нейронной сети оценки качества изображений, учитывающего взаимодействие дефектов и взаимосвязь между параметрами модели деградации.

Подавление дефектов размытием

Экспериментально установлено, что при применении размытия после других типов дефектов (шума, JPEG-артефактов) видимость последних снижается. Данный эффект обусловлен усреднением значений пикселей в окрестности, что «размазывает» высокочастотные компоненты, характерные для шума и блочных артефактов.

Для учёта данного эффекта в целевых метках была разработана методика корректировки значений при генерации обучающей выборки.

При применении размытия к области, содержащей шум, целевая метка силы шума корректируется с учётом коэффициента подавления, основанного на свойствах гауссовского ядра (1).

$$S_{noise}(\sigma, \alpha) = \frac{1}{2\sqrt{\pi} \cdot \sigma_{major} \cdot \sigma_{minor}}, \#(1)$$

где σ_{major} и σ_{minor} – стандартные отклонения анизотропного гауссовского ядра вдоль главной и побочной осей.

Данная формула выведена из нормировки двумерного гауссовского распределения и характеризует долю высокочастотной энергии, сохраняющейся после применения фильтра. При увеличении силы размытия коэффициент подавления уменьшается, что отражает снижение видимости шума.

Подавление JPEG-артефактов при размытии имеет более выраженный характер: при достижении определённого порога размытия блочная структура становится неразличимой. Для моделирования данного эффекта использована функция гауссова затухания (2).

$$S_{jpeg}(\sigma) = \exp\left(-\left(\frac{\sigma}{0.5}\right)^2\right) \cdot \#(2)$$

Параметр 0.5 был подобран экспериментально на основе визуального анализа изображений с различными комбинациями сжатия и размытия. При $\sigma > 1.0$ коэффициент подавления падает ниже 0.02, что соответствует практической невозможности детектирования JPEG-артефактов.

Корректировка целевых меток применяется только когда размытие применяется после других дефектов. Шум и артефакты сжатия вносят высокочастотные компоненты в изображение, но не повышают резкость размытых элементов.

Эквивалентность эффектов размытия и масштабирования

При масштабировании изображения с последующим увеличением до исходного размера неизбежно возникает размытие, обусловленное интерполяцией пикселей. Данный эффект визуально схож с применением фильтра размытия Гаусса, что создаёт проблему для обучения нейронной сети – модель не может однозначно определить, вызвано ли размытие операцией масштабирования или отдельным применением гауссовского фильтра.

Для обеспечения стабильности обучения было принято решение объединить предсказания по масштабированию и размытию при финальной оценке качества

Для количественной оценки взаимосвязи между коэффициентом масштабирования K и эквивалентным параметром размытия σ был проведён эксперимент.

Методика эксперимента:

1. Создавалось тестовое изображение размером 201×201 пикселей с единичным импульсом в центре (дельта-функция);
2. Изображение сжималось с коэффициентом K с использованием разных видов интерполяции, затем увеличивалось до исходного размера тем же методом;
3. Из полученного изображения извлекался горизонтальный профиль яркости через центр;
4. Профиль нормализовался по максимальному значению и аппроксимировался функцией Гаусса (3):

$$g(x) = A \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right), \#(3)$$

где x – координата пикселя относительно центра импульса (расстояние в пикселях);

A – амплитуда сигнала (нормированное максимальное значение яркости профиля);

σ – стандартное отклонение гауссовского распределения, характеризующее ширину ядра размытия и силу эквивалентного искажения;

5. Методом наименьших квадратов определялся параметр σ , характеризующий эквивалентное размытие.

В результате было установлено, что для коэффициентов масштабирования K зависимость эквивалентной сигмы от коэффициента масштабирования аппроксимируется линейной функцией (4).

$$\sigma_{equivalent} \approx 0.3 \cdot K \#(4)$$

При малых значениях $\sigma (< 0.5)$ зависимость становится нелинейной, однако такие значения размытия не использовались при обучении модели, так как соответствуют минимальным уровням деградации, не влияющим существенно на восприятие качества.

Результатом нахождения данной зависимости стала возможность объединить выходы нейронной сети, отвечающие за размытие и масштабирование. Это позволило стабилизировать обучение и повысить точность регрессии параметров деградации.

Взаимодействие дефектов одного типа

При генерации обучающей выборки дефекты применяются к изображению последовательно в случайном порядке. Поскольку сложные деградации могут включать несколько преобразований одного и того же типа, следует рассмотреть правила сложения дефектов одинакового типа.

1. *Шум*. При последовательном добавлении гауссовского шума с параметрами σ_1 и σ_2 итоговая сила шума определяется по правилу сложения дисперсий (5):

$$\sigma_{total} = \sqrt{\sigma_1^2 + \sigma_2^2} \#(5)$$

2. *Размытие*. Последовательное применение двух гауссовых фильтров эквивалентно одному фильтру с параметром σ_{total} , который также рассчитывается по формуле (5).
3. *JPEG-артефакты*. При повторном сжатии часто остаются специфические артефакты, но общий уровень дефектов не складывается аддитивно, если сетка блоков DCT не совпадает. При смещении области применения сжатия относительно исходной сетки блоков эффект усиливается, однако при полном совпадении повторное сжатие не вносит существенных изменений.

Заключение

В ходе выполнения работы был разработан метод корректировки целевых меток при обучении нейронной сети оценки качества изображений. Предложенный подход учитывает:

1. Взаимосвязь между масштабированием и размытием Гаусса, для которой выявлена эмпирическая зависимость $\sigma \approx 0.3 \cdot K$;
2. Взаимодействие дефектов при их последовательном применении, включая аддитивность шума и размытия, а также неаддитивность JPEG-артефактов;
3. Подавление высокочастотных дефектов (шум, JPEG) при применении размытия, для учёта которого предложены формулы корректировки целевых меток на основе свойств гауссовского фильтра.

Экспериментальные результаты показали, что корректировка целевых меток позволяет повысить точность регрессии параметров дефектов на 20% по метрике MSE. Разработанная система демонстрирует высокую эффективность в детектировании и оценке интенсивности различных типов дефектов изображений.

Перспективные направления развития:

1. Расширение модели взаимодействия дефектов;
2. Адаптивная корректировка меток на основе анализа локальной текстуры изображения;
3. Создание локальных дефектов моделью деградации, для того чтобы модель определяла не только наличие дефектов, но и их местоположение;

4. Интеграция метода в системы восстановления качества для автоматического подбора параметров обработки.

Библиографический список

1. В.А. Анохин, В.В. Белов. Метод улучшения обобщающей способности нейронных сетей повышения качества изображений // 4 том сборника трудов VIII Международного форума СТНО-2025
2. Р. Гонсалес, Р. Вудс Цифровая обработка изображений; Москва: Техносфера, 2005.– 1072 с

УДК 616-77; ГРНТИ 76.29.41

ПЕРСПЕКТИВЫ ПРИМЕНЕНИЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ В УПРАВЛЕНИИ БИОНИЧЕСКИМИ ПРОТЕЗАМИ КИСТИ

Я.В. Долгов, Д.Е. Судаков

*Тамбовский государственный технический университет,
Российская Федерация, Тамбов, 2yaroslav00000@mail.ru
Российская Федерация, Тамбов, sudakov.dima1702@yandex.ru*

Аннотация. Данная работа посвящена анализу текущих ограничений бионических протезов верхних конечностей и поиску путей их преодоления. Современные миоэлектрические протезы, управляемые сигналами мышц, стали отраслевым стандартом, однако их функционал зачастую сводится к примитивным операциям (хват/разжим), что не позволяет в полной мере компенсировать утрату руки. Ключевая мысль исследования заключается в том, что прорыв в данной сфере станет возможным благодаря внедрению технологий искусственного интеллекта, а именно – глубоких нейронных сетей. В материале описывается механизм работы многоканальных ЭМГ-датчиков в тандеме с нейросетевыми алгоритмами, способными расшифровывать сложные паттерны биоэлектрической активности мышц. Это позволяет преобразовывать тонкие моторные намерения человека (например, определенные жесты или движения пальцев) в команды для протеза. Применение машинного обучения трансформирует протез из грубого механического инструмента в органичное продолжение тела, что напрямую влияет на автономность пациента и его уровень жизни. В заключении обозначены главные векторы развития отрасли: создание энергоэффективных алгоритмов для работы на компактных процессорах и улучшение методов сбора биометрических данных. Применение машинного обучения трансформирует протез из грубого механического инструмента в органичное продолжение тела, что напрямую влияет на автономность пациента и его уровень жизни.

Ключевые слова: бионическое протезирование, электромиография (ЭМГ), нейронные сети, распознавание паттернов, качество жизни, человеко-машинное взаимодействие.

PROSPECTS OF USING DEEP NEURAL NETWORKS IN THE CONTROL OF BIONIC HAND PROSTHESES

Y.V. Dolgov, D.E. Sudakov

*Tambov State Technical University,
Russia, Tambov, 2yaroslav00000@mail.ru
Russia, Tambov, sudakov.dima1702@yandex.ru*

The summary. This paper analyzes the current limitations of upper limb bionic prostheses and explores ways to overcome them. Modern myoelectric prostheses, controlled by muscle signals, have become the industry standard; however, their functionality is often limited to primitive operations (grip/release), which fails to fully compensate for the loss of a hand. The key idea of the study is that a breakthrough in this field will become possible through the implementation of artificial intelligence technologies, namely deep neural networks. The material describes the mechanism of operation of multichannel EMG sensors in tandem with neural network algorithms capable of deciphering complex patterns of bioelectrical muscle activity. This allows for the translation of subtle motor intentions (e.g., specific gestures or finger movements) into commands for the prosthesis. The application of machine learning transforms the prosthesis from a crude mechanical tool into a natural extension of the body, directly impacting patient autonomy and quality of life. In conclusion, the main directions for further development in the field are outlined: the creation of energy-efficient algorithms for operation on compact processors and the improve-

ment of methods for collecting biometric data. The application of machine learning transforms the prosthesis from a crude mechanical tool into a natural extension of the body, directly impacting patient autonomy and quality of life.

Keywords: bionic prosthetics, electromyography (EMG), neural networks, pattern recognition, quality of life, human-machine interaction.

Потеря кисти вследствие травмы или врожденной патологии существенно снижает способность человека к самообслуживанию и выполнению рутинных задач, ухудшая общее качество жизни. На сегодняшний день «золотым стандартом» замещения отсутствующей конечности служат активные бионические протезы, функционирующие за счет считывания электромиографических (ЭМГ) сигналов с мышц культи предплечья [1]. Несмотря на широкое распространение, данные устройства имеют ряд эксплуатационных недостатков. Помимо проблем с износостойкостью механических компонентов и необходимостью частой подзарядки, главным минусом остается крайне скудный репертуар движений – как правило, пациенту доступны лишь базовые функции сжатия и разжатия. Это превращает протез в приспособление для выполнения ограниченного перечня действий, а не в полноценную замену руки.

Если вопросы автономности и механической надежности решаются путем совершенствования элементной базы (аккумуляторов, микродвигателей и полимерных материалов), то проблема точного и интуитивного позиционирования остается острой. Традиционные алгоритмы управления, основанные на линейной зависимости между амплитудой ЭМГ-сигнала и усилием сжатия, неспособны интерпретировать сложные моторные команды, требующие скоординированной работы пальцев и запястья [2]. Возникает необходимость в интеллектуальных системах, способных распознавать многообразие двигательных намерений человека.

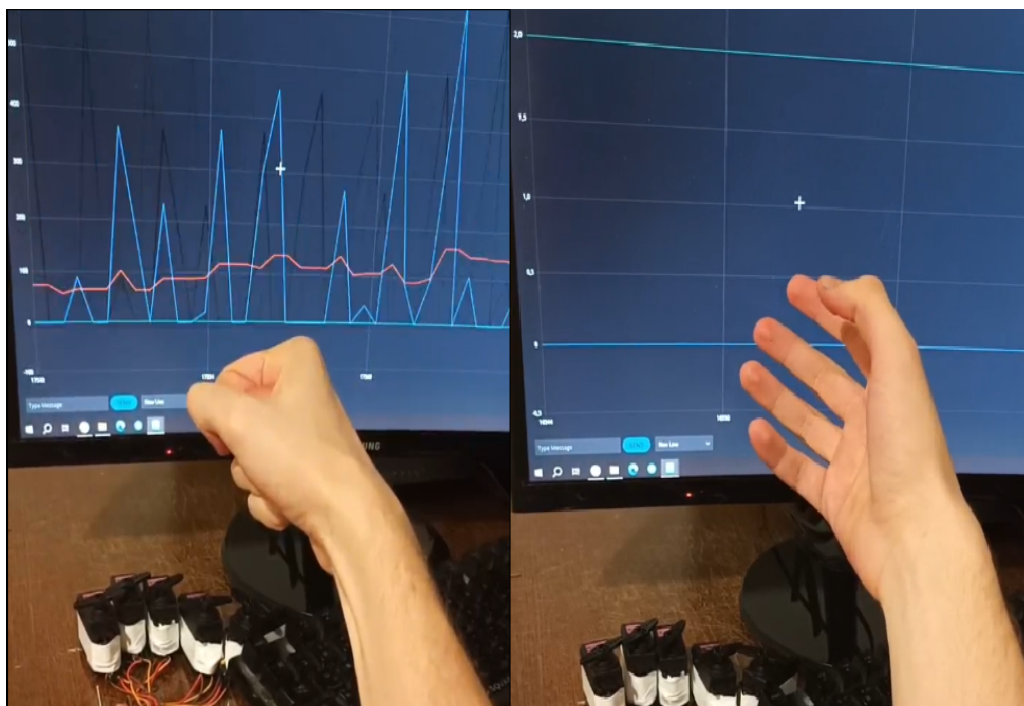


Рис. 1. Активность мышцы при сокращении и расслаблении.
Слева – сокращена, справа – расслаблена

Эффективным решением этого противоречия является интеграция алгоритмов искусственного интеллекта и глубинного машинного обучения непосредственно в управляющую электронику протеза. Развитие микроэлектроники сделало возможным создание миниатюрных многоканальных ЭМГ-сенсоров с высоким разрешением. Формирование мини-

мально рабочей конфигурации системы сбора данных требует применения прецизионных инструментальных усилителей, таких как AD627 или AD620, обеспечивающих необходимое подавление синфазной помехи и усиление слабых биоэлектрических потенциалов. Критически важным звеном выступает выбор микроконтроллера: использование относительно мощных решений, например ESP32-S3, построенного на архитектуре Xtensa LX7 с аппаратной поддержкой векторных операций, позволяет выполнять предварительную обработку сигналов непосредственно на борту устройства. Для достижения высокой точности распознавания двигательных намерений целесообразно применять системы с количеством каналов более четырех и параллельными аналого-цифровыми преобразователями (АЦП), обеспечивающими синхронность выборок. В этом контексте особую перспективность демонстрируют интегральные схемы ADS1299 или ADS1298, которые объединяют в себе до восьми и более каналов программируемого усиления и многоканальные АЦП с высоким разрешением, гарантируя тем самым высокое качество регистрируемой электромиограммы с культи и формируя надежную базу для последующих алгоритмов декодирования [3].

Такие датчики способны улавливать широкий спектр биоэлектрических сигналов от различных мышечных групп культи. Однако сырая электромиограмма требует тщательной предварительной обработки: фильтрации для удаления сетевых наводок (50/60 Гц), подавления электродных артефактов движения и выделения активной составляющей сигнала в диапазоне 10–500 Гц. Далее выполняется сегментация сигнала на перекрывающиеся временные окна длительностью от 150 до 300 мс, что обеспечивает компромисс между поддержкой управления и стабильностью распознавания. В каждом окне из многоканальных данных извлекаются как временные, так и частотные признаки – среднеквадратичное значение, коэффициенты авторегрессии, параметры вейвлет-преобразования. Именно на этом этапе традиционные системы используют амплитуду сигнала как аналоговую меру усилия, тогда как нейросетевые подходы позволяют обойтись без ручного выделения признаков, передавая сырые или минимально обработанные временные ряды на вход глубинных моделей [4].

Критическое значение при построении системы сбора ЭМГ имеет топология входного каскада. Биоэлектрические потенциалы, регистрируемые с поверхности культи, характеризуются крайне низкой амплитудой (десятки микровольт — единицы милливольт) и накладываются на мощные синфазные помехи, прежде всего наводки сети переменного тока (50/60 Гц) и электродные потенциалы постоянного тока. В таких условиях применение простой двухэлектродной схемы (сигнальный электрод и «земля») оказывается несостоятельным. Поскольку импеданс кожи и качество контакта электродов неидеальны, синфазная помеха трансформируется в дифференциальную составляющую, которая после усиления может полностью перекрыть полезный сигнал. Более того, двухэлектродная топология не позволяет реализовать эффективное подавление синфазного сигнала, что приводит к дрейфу нуля, высокой чувствительности к движениям пациента и, как следствие, к нестабильности работы всего тракта.

Именно поэтому стандартом электромиографии является дифференциальное включение с использованием трех электродов: активного (инвертирующего), референтного (неинвертирующего) и «земляного» (дополнительного, часто называемого *reference* или *driven-right-leg*). В такой конфигурации полезным сигналом служит разность потенциалов между активным и референтным электродами, тогда как «земля» выступает в роли точки стабилизации потенциала тела пациента относительно схемы усилителя. Ключевое преимущество достигается за счет применения инструментального усилителя с высоким коэффициентом ослабления синфазного сигнала (CMRR — *Common Mode Rejection Ratio*), который подавляет идентичные помехи на обоих входных электродах, выделяя лишь дифференциальную составляющую мышечной активности. Промышленные микросхемы, такие как AD620 или AD627, реализуют именно эту топологию, обеспечивая CMRR на уровне 80–100 дБ в рабочем диапазоне частот.

Дополнительным преимуществом трехэлектродной схемы является возможность организации активного экранирования и снижения влияния импеданса электрод-кожа. Современные многоканальные системы (например, ADS1298) интегрируют в себе не только дифференциальные усилители и параллельные АЦП, но и цепи формирования опорного потенциала, что позволяет синхронно регистрировать сигналы с нескольких мышечных групп, сохраняя при этом высокое соотношение сигнал/шум. Без применения дифференциальной топологии даже при использовании высококачественных микроконтроллеров и нейросетевых алгоритмов неизбежно накопление артефактов, которые невозможно будет компенсировать цифровой фильтрацией, поскольку спектр помех зачастую перекрывается с частотным диапазоном ЭМГ (10–500 Гц). Следовательно, закладывание корректной дифференциальной схемы с тремя электродами на этапе проектирования является необходимым условием для получения чистого сигнала, пригодного для обучения и стабильной работы нейросетевых моделей, интерпретирующих двигательные намерения пользователя [3].

Нейросетевые модели, предварительно обученные на обширных массивах данных ЭМГ, научаются выявлять уникальные пространственно-временные паттерны мышечной активности. Для решения этой задачи применяются различные архитектуры. Сверточные нейронные сети (CNN) эффективно извлекают пространственные корреляции между каналами, действуя как адаптивные пространственные фильтры. Рекуррентные сети (LSTM, GRU) учитывают динамику мышечных сокращений во времени, что критически важно для распознавания непрерывных движений и переходных режимов. Перспективным направлением становится применение трансформеров, которые благодаря механизму само-внимания способны моделировать долгосрочные зависимости в последовательности ЭМГ-кадров без ограничений, присущих рекуррентным слоям. Комбинированные CNN-LSTM и CNN-Transformer архитектуры позволяют объединить преимущества пространственной и временной обработки, обеспечивая надежное декодирование таких тонких намерений, как выпрямление указательного пальца, щипковый захват или супинация кисти [2,3].

В процессе калибровки под конкретного пользователя алгоритм создает персонализированную и самообучающуюся модель управления, которая со временем адаптируется к особенностям пациента. Однако одной лишь точной классификации жестов недостаточно для создания ощущения естественности. Важнейшим компонентом интеллектуального протеза является замкнутый контур управления, включающий сенсорную обратную связь. Современные исследования направлены на интеграцию тактильных датчиков в захватные поверхности протеза с последующим транскодированием тактильной информации в стимулы, подаваемые на сохраненные нервные окончания культи или через механическую вибрацию на проксимальные участки конечности. Такая бидирекциональная связь позволяет пользователю не только управлять протезом, но и «чувствовать» усилие сжатия, текстуру объекта, что существенно повышает точность манипуляций и снижает когнитивную нагрузку [5].

Внедрение нейросетей и методов машинного обучения в контур управления бионическими протезами знаменует собой сдвиг парадигмы: от примитивного силового контроля к интеллектуальному распознаванию намерений. Такой подход нивелирует ключевой недостаток современных систем – жесткую ограниченность функционала. Реализация данной концепции позволяет создавать протезы будущего, которые воспринимаются пользователем как естественная часть тела, а не инородный предмет. Это существенно расширяет физические возможности человека, способствует его социальной и бытовой независимости и, как следствие, улучшает качество его жизни.

Ключевыми направлениями дальнейших научных изысканий являются оптимизация нейросетевых архитектур для функционирования на маломощных бортовых процессорах, разработка новых методов сбора и синтетического обогащения обучающих выборок ЭМГ-сигналов (включая использование генеративно-сопоставительных сетей и симуляторов биомеханики), а также создание замкнутых систем с тактильной обратной связью, способных адаптироваться к изменяющимся условиям использования в реальном времени. Достиже-

ние этих целей позволит перевести бионическое протезирование из разряда вспомогательных технических средств в категорию полноценных нейроинтерфейсов, возвращающих утраченные функции на качественно новом уровне.

Библиографический список

1. Micera, S. Control of hand prostheses using peripheral information / S. Micera, J. Carpaneto, S. Raspopovic // IEEE Reviews in Biomedical Engineering. – 2010. – Vol. 3. – P. 48–68.
2. Яцун, С. Ф. Разработка системы управления кистевого протеза с тактильной обратной связью / С. Ф. Яцун, А. В. Мальчиков, А. Н. Рукавицын // Медицинская техника. – 2024. – № 1. – С. 7–9.
3. Градиентный метод нейросетевого управления многосвязными нелинейными нестационарными стохастическими системами / С. В. Фролов, А. А. Коробов, А. Ю. Потлов, К. С. Савинова // Приборы и системы. Управление, контроль, диагностика. – 2021. – № 5. – С. 41-48. – DOI 10.25791/pribor.5.2021.1262. – EDN NHLYAD.
4. Акопян, Ф. Г. Нейросеть в биопротезировании / Ф. Г. Акопян, В. В. Дубровин // Энергосбережение и эффективность в технических системах : Материалы IX Всероссийской научно-технической конференции с международным участием для студентов, молодых ученых и специалистов, Тамбов, 19 октября 2022 года. – Тамбов: Издательский центр ФГБОУ ВО "Тамбовский государственный технический университет", 2022. – С. 154-156. – EDN IKXOFX.
5. Ширкина, Е. С. Актуальность разработки системы для тренировки управления бионическим протезом предплечья / Е. С. Ширкина // Радиотехника. Проблемы и перспективы развития : Сборник трудов IX Всероссийской научно-практической конференции с международным участием, Тамбов, 07 мая 2024 года. – Тамбов: Тамбовский государственный технический университет, 2024. – С. 253-254

УДК 004.02; ГРНТИ 28.23.19

ОПТИМИЗАЦИЯ АЛГОРИТМА MINIMAX В ПАРАМЕТРИЗОВАННЫХ ШАХМАТАХ: ИССЛЕДОВАНИЕ ОЦЕНОЧНЫХ ФУНКЦИЙ

Ф.А. Табаков, Н.К. Петрова

Казанский государственный энергетический университет,
Российская Федерация, Казань, fedortab7@gmail.com

Аннотация. Исследовано влияние трёх типов оценочных функций на эффективность алгоритма Minimax с альфа-бета в параметризованных шахматах – варианте, где у фигур изменяются числовые характеристики (здоровье, атака, защита, мобильность). Реализованы материальная, позиционная и адаптивная комбинированная функции; проведены 150 экспериментов в Unity (C#). Проанализированы процент побед, время хода, глубина поиска и средняя оценка позиции. Адаптивная комбинированная функция статистически превосходит монофункции (60% vs 44% и 52%), подтверждая её эффективность для разработки игровых ИИ.

Ключевые слова: Minimax, альфа-бета отсечение, параметризованные шахматы, оценочная функция, игровой ИИ, адаптивная эвристика, Unity, C#.

OPTIMIZATION OF THE MINIMAX ALGORITHM IN PARAMETERIZED CHESS: RESEARCH OF EVALUATION FUNCTIONS

F.A. Tabakov, N.K. Petrova

Kazan State Power Engineering University,
Russian Federation, Kazan, fedortab7@gmail.com

Abstract. The influence of three types of evaluation functions on the efficiency of the Minimax algorithm with alpha-beta pruning in parametrized chess — a variant where pieces have modified numerical characteristics (health, attack, defense, mobility) — was investigated. Material, positional, and adaptive combined evaluation functions were implemented; 150 experiments were conducted in Unity (C#). The win rate, move time, search depth, and average position score were analyzed. The adaptive combined function statistically outperforms the single-focus functions (60 % vs. 44 % and 52 %), confirming its effectiveness for developing game AI.

Keywords: Minimax, alpha-beta pruning, parametrized chess, evaluation function, game AI, adaptive heuristic, Unity, C#.

Разработка интеллектуальных агентов для стратегических игр остаётся актуальной задачей искусственного интеллекта. Шахматы и их модификации являются удобной средой для тестирования методов поиска, эвристического оценивания и принятия решений в условиях полной информации.

В последнее время внимание исследователей смещается от классических правил к параметризованным моделям. Параметризованные шахматы – это расширение классических правил, в котором каждая фигура наделяется числовыми характеристиками: здоровьем, силой атаки, защитой, мобильностью. Эти характеристики, в данном контексте, называются *юнитами*. В процессе игры эти параметры изменяются: фигуры получают повреждения, теряют боевую ценность, могут быть временно усилены или ослаблены. Такая модель приближает игру к тактическим симуляторам и требует от ИИ-агентов учёта не только статического материального баланса, но и динамики состояний фигур.

Классическим алгоритмом решения задач в играх с полной информацией является Minimax с альфа-бета отсечением [1-3]. Однако его эффективность критически зависит от качества оценочной функции. В параметризованных шахматах традиционные оценки, ориентированные на материальное соотношение и типовые позиционные факторы, оказываются недостаточно информативными.

Цель работы – сравнительный анализ влияния различных типов оценочных функций (количественной, позиционной, комбинированной) на эффективность алгоритма Minimax с альфа-бета отсечением в среде параметризованных шахмат.

Теоретические основы и методология

Алгоритм Minimax обеспечивает поиск оптимальной стратегии при полной информации. Альфа-бета отсечение позволяет сократить количество просматриваемых узлов, отсекая заведомо бесперспективные ветви. В классической реализации ключевыми факторами, определяющими практическую эффективность, являются порядок генерации ходов, глубина поиска и, прежде всего, оценочная функция.

В параметризованных шахматах ценность фигуры становится динамической, появляются временные преимущества, а позиционная значимость зависит от состояния соседних юнитов. Это требует адаптации оценочных функций.

В рамках исследования были выделены три класса функций [4,5]:

1. Количественные (материальные) – оценивают сумму числовых характеристик фигур (здоровье, атака, защита). Преимущество: скорость вычисления. Недостаток: игнорирование позиционной составляющей;

2. Позиционные – учитывают контроль полей, мобильность, безопасность короля, угрозы и защищённость. Требуют более сложных эвристик, но лучше отражают стратегическую составляющую;

3. Комбинированные – взвешенная сумма материального и позиционного компонентов с адаптивными коэффициентами, зависящими от фазы игры.

Методология эксперимента включала:

- интеграцию модуля оценочных функций в единую архитектуру;
- проведение серий партий с фиксированными параметрами поиска;
- сбор метрик: процент побед, средняя оценка позиции, количество ходов, время на ход, эффективность отсечений;
- статистический анализ результатов.

Экспериментальная часть

В качестве платформы для реализации, визуализации и управления экспериментами был выбран Unity –кросс платформенный игровой движок (среда разработки) для создания 2D, 3D и AR/VR приложений. Это позволило эффективно прототипировать пара-

метризованные правила и наглядно проводить тестовые партии. В рамках данной платформы использовался подход на C# с применением ООП: для фигур, состояний и стратегий определены интерфейсы и абстрактные классы, что упрощает расширение набора оценочных функций и добавление новых правил. Модульная архитектура позволила изолировать оценочные функции от ядра игры и поискового алгоритма, что обеспечило возможность их гибкой замены и сравнительного тестирования.

В ходе эксперимента сравнивались три типа ИИ-агентов (ботов), каждый из которых использовал алгоритм Minimax с альфа-бета отсечением, но отличался оценочной функцией:

Реализованные оценочные функции (ИИ-агенты- боты):

- MaterialEvaluator (M) – материальная (количественная) оценка: взвешенная сумма здоровья, атаки и защиты всех фигур.

- PositionalEvaluator (P) – позиционная оценка: учитывает контроль центра, мобильность, угрозы, безопасность короля, пороговые значения здоровья.

- CombinedEvaluator (K) – комбинированная адаптивная оценка: линейная комбинация материальной и позиционной оценок с адаптивными весами, изменяющимися в зависимости от этапа игры (дебют, миттельшпиль, эндшпиль).

Качественный анализ игрового поведения ботов:

- MaterialEvaluator – тактически агрессивен, эффективно уничтожает ослабленные фигуры, но не видит позиционных угроз, не занимает центр, не защищает короля;

- PositionalEvaluator – стратегически грамотен, контролирует ключевые поля, но пассивен в тактике, упускает выгодные размены;

- CombinedEvaluator – адаптивен, сочетает стратегическое планирование с тактической точностью, демонстрирует сбалансированный стиль игры.

Методика тестирования:

Проведено 150 партий (по 50 для каждой пары функций: М-Р, М-К, Р-К) с фиксированной глубиной поиска 2–3 полухода. Боты играли друг против друга автоматически до гибели короля. Запись логов позволила провести детальный количественный и качественный анализ.

В ходе эксперимента фиксировались также вычислительные характеристики работы алгоритма: время на ход, количество обработанных узлов и процент α - β отсечений. Сводные данные по производительности каждой оценочной функции приведены в таблице 1.

Таблица 1. Производительность вычислений

Тип функции	Время/ход (мс)	Узлов/ход	Отсечений α - β (%)	Эффективность поиска (узлов/мс)
Материальная	420 ± 150	690 ± 320	68 ± 8%	1.64
Позиционная	780 ± 280	1035 ± 480	54 ± 9%	1.33
Комбинированная	620 ± 220	862 ± 400	61 ± 8%	1.39

Материальная функция ожидаемо оказалась самой быстрой, однако комбинированная продемонстрировала наилучший баланс скорости и качества оценки.

На рисунке 1 представлена динамика процента побед каждого из трёх ботов в зависимости от стадии партии. Данные получены по всем проведённым играм (150 партий) и отражают совокупную статистику побед каждой функции против обоих соперников.

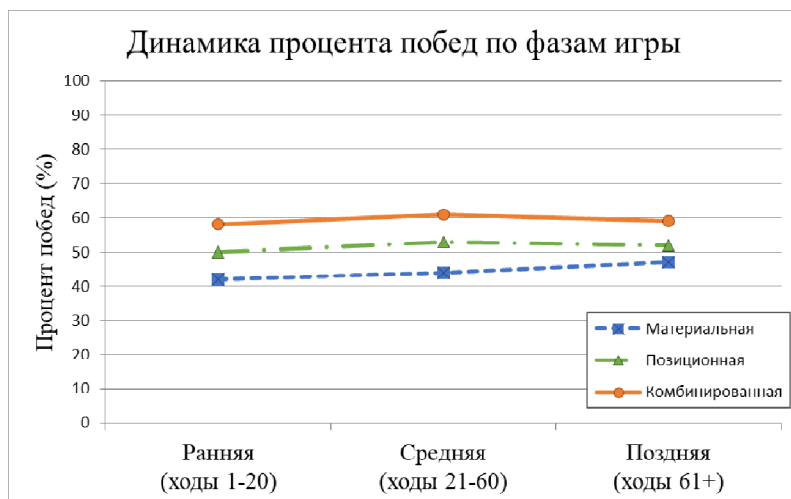


Рис. 1. Динамика процента побед по фазам игры

Количественные результаты эксперимента, включая процент побед и среднюю оценку позиции для каждой оценочной функции, представлены в таблице 2. Из приведённых данных видно, что комбинированная функция обеспечивает статистически значимое преимущество по проценту побед (60% против 44% и 52%), наилучшую среднюю оценку позиции и устойчивую результативность на всех стадиях игры.

Таблица 2. Эффективность оценочных функций

Тип функции	Побед/100 игр	Побед (%)	Среднее число ходов	Средняя оценка позиции (y.e.)	Средняя глубина поиска
Материальная	44 ± 6	44 ± 4%	85.5 ± 9.6	521.25±187.65	2.8
Позиционная	52 ± 6	52 ± 5%	102.6 ± 12.3	354.45±133.44	2.5
Комбинированная	60 ± 6	60 ± 5%	95.4 ± 11.4	675.7±249.24	2.6

Проведённое экспериментальное исследование позволило выполнить сравнительный анализ трёх типов оценочных функций – материальной, позиционной и комбинированной – в среде параметризованных шахмат. Все функции были реализованы в единой модульной архитектуре на платформе Unity и протестированы в сериях из 50 партий для каждой пары соперников при фиксированной глубине поиска 2–3 полухода.

Таким образом, в ходе исследования проведён сравнительный анализ трёх типов оценочных функций для алгоритма Minimax с альфа-бета отсечением в среде параметризованных шахмат. Полученные результаты подтверждают, что в условиях динамически изменяющихся характеристик фигур ни материальная, ни позиционная оценка по отдельности не обеспечивают устойчивого преимущества.

Наиболее эффективной оказалась комбинированная адаптивная функция, динамически изменяющая веса компонентов в зависимости от фазы игры. Её применение обеспечило 60% побед, что на 16 п.п. выше, чем у материальной, и на 8 п.п. выше, чем у позиционной функции. Комбинированная функция также показала наилучшую среднюю оценку позиции и оптимальный баланс между вычислительной сложностью и качеством принимаемых решений.

Предложенная модульная архитектура и методика сравнительного анализа могут быть использованы при разработке игровых ИИ для стратегий с пошаговой тактикой, а также в учебных целях для демонстрации принципов интеграции алгоритмов поиска и эвристических оценок. Полученные результаты найдут применение в дальнейших исследованиях, связанных, например, с автоматической настройкой весов комбинированной функции, методами машинного обучения.

Библиографический список

1. Жданов Н.А., Бурханова Ю.М., Воронецкий Ю.О. Алгоритм «минимакс» и его реализация на примере игры «крестики-нолики» // Политехнический молодёжный журнал. — 2019. — №5.
2. Алгоритм minimax в шахматах [Электронный ресурс] // Habr. — 2024. — URL: <https://habr.com/ru/companies/otus/articles/785512/> (дата обращения: 10.02.2026).
3. Fuller S.H., Gaschnig J.G., Gillogly J.J. Analysis of the alpha-beta pruning algorithm. — Pittsburgh: Carnegie-Mellon University, Dept. of Computer Science, 1973. — Technical Report.
4. Arbiser A. Towards the Unification of Intuitive and Formal Game Concepts with Applications to Computer Chess // Proceedings of DiGRA 2005 Conference. — Tampere: DiGRA, 2005. — DOI: 10.26503/dl.v2005i1.165.
5. Sadikov A., Bratko I., Kononenko I. Search versus Knowledge: An Empirical Study of Minimax on KRK // Advances in Computer Games / H.J. van den Herik, H. Iida, E.A. Heinz (eds.). — Boston: Springer, 2004. — P. 33–46. — DOI: 10.1007/978-0-387-35706-5_3.

СЕКЦИЯ «ЭВМ И СИСТЕМЫ»

УДК 004.75; ГРНТИ 20.23.19, 50.41.25

ОСНОВНЫЕ КОНЦЕПЦИИ РАСПРЕДЕЛЁННЫХ ВЫЧИСЛЕНИЙ: АРХИТЕКТУРЫ, МОДЕЛИ СОГЛАСОВАННОСТИ И УПРАВЛЕНИЕ РЕСУРСАМИ

Р.В. Хруничев

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, hrunichev_robert@mail.ru*

Аннотация. Распределённые вычисления лежат в основе современных сервисов обработки данных. В статье рассмотрены базовые понятия распределённых вычислений: типовые архитектуры, модели программирования, механизмы отказоустойчивости, а также компромиссы согласованности данных. Отдельное внимание уделено управлению ресурсами и планированию задач в гетерогенных системах, включая краевые/туманные вычисления.

Ключевые слова: распределённые вычисления, распределённые системы, кластер, облако, IoT, планирование задач, отказоустойчивость, согласованность данных.

BASIC CONCEPTS OF DISTRIBUTED COMPUTING: ARCHITECTURES, COORDINATION MODELS, AND RESOURCE MANAGEMENT

R.V. Khrunichev

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, hrunichev_robert@mail.ru*

The summary. Distributed computing is the basis of modern data processing services. The article discusses the basic concepts of distributed computing, including typical architectures, programming models, fault-tolerance mechanisms, and data consistency trade-offs. Special attention is given to resource management and task scheduling in heterogeneous systems, including edge/fog computing.

Keywords: distributed computing, distributed systems, cluster, cloud, IoT, task scheduling, fault-tolerance, data consistency.

Под распределёнными вычислениями обычно понимают выполнение вычислительной задачи на множестве связанных узлов (серверов/устройств), которые обмениваются данными по сети и совместно формируют результат [5, 6]. В отличие от обычного распараллеливания вычислений на одном компьютере, распределённым системам свойственны сетевые задержки, частичные отказы и неоднородность ресурсов, поэтому ключевыми становятся масштабируемость, отказоустойчивость и управляемость [5, 6].

Практические предпосылки распределения: рост объёмов данных, необходимость обработки как можно ближе к источнику (edge/fog для IoT), экономия за счёт горизонтального масштабирования и повышение доступности сервисов [7]. В то же время распределение усложняет разработку: появляется необходимость согласования состояния, обработки отказов и управления нагрузкой.

Классы распределённых вычислений и базовые определения

На прикладном уровне удобно различать несколько классов:

1. **Кластеры** – набор узлов в едином центре обработки данных (ЦОД), управляемых общей системой планирования (свойственно для корпоративных платформ и аналитики больших данных) [1].

2. **Облачные вычисления** – предоставление виртуализированных ресурсов «как сервиса», где распределение скрыто за API, а пользователь управляет логикой и масштабированием на уровне приложений [5, 6].

3. **Grid/научные распределённые вычисления** – объединение ресурсов разных организаций для расчётов (виртуальный суперкомпьютер на кластерном принципе), ему свойственны пакетные задания, очереди и федеративное управление [8].

4. **Краевые/туманные вычисления** – перенос части обработки на периферию сети (шлюзы, устройства, микроЦОДы) для снижения задержек и разгрузки магистрали [7].

Распределённая система описывается множеством узлов N , каналами связи и протоколами обмена. Её важнейшими свойствами являются:

- **Масштабируемость:** рост производительности/пропускной способности при добавлении ресурсов.
- **Отказоустойчивость:** способность продолжать работу при отказах части узлов/каналов.
- **Прозрачность распределения:** анонимизация сложности (местоположение данных, миграция, репликация) от пользователя/разработчика [6].

Типовые архитектуры и схемы взаимодействия

1. Архитектура master–worker

Одна из самых распространённых схем – **координатор (master/dispatcher)** и множество исполнителей (**workers**). Координатор принимает задания, разбивает их на подзадачи, распределяет по исполнителям, отслеживает состояние и собирает результат.

Схематично архитектура показана на рисунке 1.

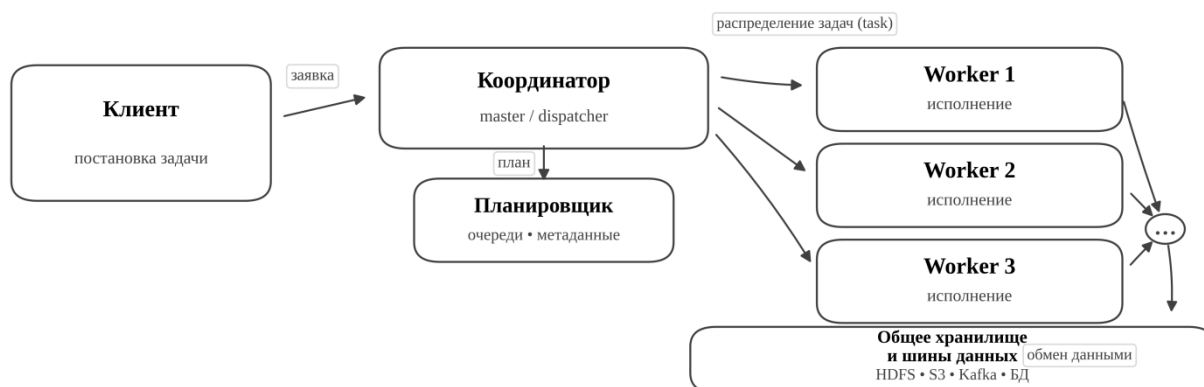


Рис. 1. Архитектура master–worker

Преимущества архитектуры – простота управления и мониторинга, централизованное планирование. Недостатки – риск единой точки отказа, требования к устойчивым хранилищам и журналированию.

2. Многоуровневые и распределённые схемы управления

В IoT и больших корпоративных инфраструктурах часто применяют многоуровневые схемы (edge → регион → облако), а в сетевой части – программно-конфигурируемые сети (SDN), где контроллеры могут быть одиночными, распределёнными или иерархическими [2]. Это снижает задержки управления и повышает устойчивость, но требует корректного протоколирования согласованного состояния (например, между контроллерами).

Для IoT-окружений интересен подход, где архитектура сети и метрики качества каналов формализуются и используются при адаптивной конфигурации. В работе [1] предложена пятиуровневая архитектура SDN для устройств интернета вещей и система «нечётких» метрик для конфигурирования каналов связи.

Модели программирования и обработка данных

1. Потокковая и пакетная обработка

По характеру данных различают следующие виды обработки:

- **Batch (пакетная):** данные накапливаются заранее, расчёт запускается массивом.
- **Streaming (потокковая):** данные поступают непрерывно, обработка событий должна осуществляться с малой задержкой.

В обоих случаях важны: разделение данных на фрагменты, обеспечение близости данных к месту обработки, равномерное распределение вычислительной нагрузки и эффективное управление узлами с низкой производительностью.

2. MapReduce как базовая модель распределённой обработки

Классическая и простая для понимания модель – **MapReduce** (рис. 2): функция *map* порождает промежуточные пары ключ–значение, а функция *reduce* агрегирует значения по ключу [9].

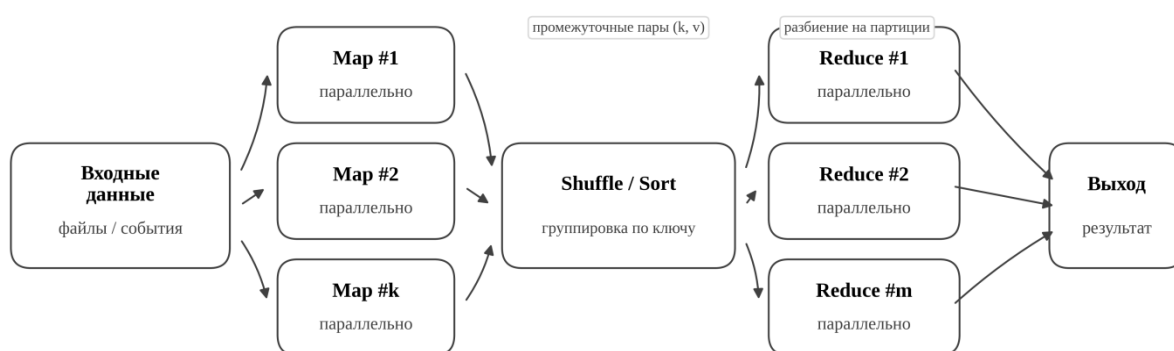


Рис. 2. Модель MapReduce

MapReduce характерна тем, что инфраструктура включает на себя параллелизм, пересылку данных и устойчивость к сбоям (повтор выполнения задач). Однако для итеративных алгоритмов и интерактивной аналитики подход с сохранением результатов на диске оказывается недостаточно эффективным.

3. In-memory вычисления и устойчивость наборов данных

Чтобы ускорить выполнение итеративных операций, применяют методы временного хранения промежуточных данных непосредственно в оперативной памяти кластеров. Например, концепция устойчивых распределённых наборов данных (RDD – Resilient Distributed Datasets) обеспечивает надёжность путём отслеживания истории преобразований («родословной») наряду с созданием контрольных точек (чекпоинтов) [10].

Отказоустойчивость и согласованность данных

1. Отказы как норма и базовые механизмы

В распределённых системах частичные отказы неизбежны: узел может выйти из строя, сетевые соединения разрываются на изолированные сегменты (partition), передача сообщений замедляется, поэтому применяются:

- **Репликация** данных/сервисов (active-active или active-passive);
- **Heartbeat и failure detector** (обнаружение отказов по тайм-аутам);
- **Повтор выполнения** задач (retries), идемпотентность операций;
- **Журналирование** (лог) для восстановления состояния.

2. Компромиссы CAP и практические модели consistency

Известная постановка **CAP** говорит, что при сетевых разделениях нельзя одновременно гарантировать сильную согласованность и доступность для всех запросов, а система выбирает компромисс в зависимости от целей [11]. На практике это выражается в спектре

моделей: от строгой согласованности (удорожание по задержкам) до eventual consistency (быстрее и устойчивее, но возможны временные расхождения).

3. Консенсус и упорядочивание событий

Для согласованной репликации и выбора ведущего узла применяются протоколы консенсуса. На практике алгоритм Raft воспринимается как более понятный способ обеспечения консенсуса с лидером и реплицируемым логом [12]. Для логического упорядочивания событий используются логические часы и отношение «happens-before» [13].

Математическое описание эффективности распределения

1. Ускорение и эффективность

Пусть T_1 – время выполнения задачи на одном узле, а T_p – на p узлах. Тогда:

$$S(p) = \frac{T_1}{T_p}, \quad E(p) = \frac{S(p)}{p}.$$

Классическая оценка верхней границы ускорения при наличии доли строго последовательной работы $(1-\alpha)$ (закон Амдала):

$$S(p) \leq \frac{1}{(1-\alpha) + \alpha/p}.$$

Практически это означает: чем выше доля коммуникаций, синхронизаций и «последовательных» участков, тем меньше эффект от добавления узлов.

2. Учёт обмена данными и накладных расходов

В простейшей модели время на p узлах:

$$T_p \approx \frac{T_{\text{пер}}}{p} + T_{\text{пер}} + T_{\text{синх}} + T_{\text{пл}},$$

где $T_{\text{пер}}$ – время передачи по сети, $T_{\text{синх}}$ – время на барьеры/блокировки, $T_{\text{пл}}$ – время работы планировщика. Поэтому за счёт грамотного выбора архитектуры и программного обеспечения распределённых вычислений стремятся: повышать локальность данных, сокращать интервалы или частоту глобальных операций синхронизации, использовать эффективные стратегии планирования.

В публикациях последних лет подчёркивается два практических направления: управление распределёнными ресурсами в динамических средах и распределённые IoT системы. В работе [1] предложена пятиуровневая архитектура программно-конфигурируемой сети устройств IoT и математическая модель с применением нечёткого вывода для метрик каналов связи. Это иллюстрирует важную тенденцию: качество распределённых вычислений и принятия решений всё чаще зависит от адаптивной сети и её метрик, а не только от вычислителей. В работе [2] проведены моделирование и верификация топологий SDN с одиночным, распределённым и иерархическим контроллерами. В статье [3] рассматривается проектирование кластерных вычислителей на базе ПЛИС для массово-параллельной обработки данных, включая роль управляющего ядра и частичную реконфигурацию. В работе [4] выполнен анализ методов принятия решений в распределённых IoT-системах на основе мультиагентного подхода, что отражает рост роли автономности узлов и локального принятия решений, когда централизация приводит к задержкам и уязвимостям.

Таким образом распределённые вычисления – это не только параллельная обработка, но и набор архитектурных решений, обеспечивающих масштабирование, устойчивость к отказам и приемлемые задержки. Базовые концепции включают: архитектуры управления и

исполнения, модели обработки данных, механизмы отказоустойчивости, выбор компромисса согласованности, планирование ресурсов в гетерогенных динамических средах.

Библиографический список

1. Перепелкин Д. А., Анисимов К. В. Разработка архитектуры и системы нечетких метрик каналов связи программно-конфигурируемой сети устройств интернета вещей // Вестник Рязанского государственного радиотехнического университета. 2022. № 80. С. 53–66. URL: https://vestnik.rsreu.ru/images/archive/2022/80/2.1_.pdf (дата обращения: 21.01.2026).
2. Никишин К. И. Моделирование и верификация топологий программно-конфигурируемых сетей // Вестник Рязанского государственного радиотехнического университета. 2022. № 80. С. 67–74. DOI: 10.21667/1995-4565-2022-80-67-74. URL: https://vestnik.rsreu.ru/images/archive/2022/80/2.2_.pdf (дата обращения: 23.01.2026).
3. Дуксин Н. А. Проектирование вычислительных кластеров программируемой логики // Вестник Рязанского государственного радиотехнического университета. 2025. № 92. С. 92–99. DOI: 10.21667/1995-4565-2025-92-92-99. URL: <https://vestnik.rsreu.ru/images/archive/2025/92/3.2-.pdf> (дата обращения: 03.02.2026).
4. Благирев М. М., Костыренков А. О. Анализ методов принятия решений в распределённых IoT-системах на основе мультиагентного подхода // Вестник Рязанского государственного радиотехнического университета. 2025. № 92. С. 146–159. DOI: 10.21667/1995-4565-2025-92-146-159. URL: <https://vestnik.rsreu.ru/images/archive/2025/92/4.4-.docx.pdf> (дата обращения: 05.02.2026).
5. Курочкин И. И. (ред.) Облачные и распределенные вычислительные системы: труды 4-й междунар. науч.-техн. конф. (Переславль-Залесский, 28 ноября – 1 декабря 2023 г.). Переславль-Залесский, 2023. URL: https://library.ict.az/elektron/2023/Sravnitelnyy_analiz_metodov_raspoznaniya_jestov_na_osnove_ispolzovaniya_kamer_qlubini.pdf (дата обращения: 07.02.2026).
6. Таненбаум Э., ван Стеен М. Распределённые системы. Принципы и парадигмы / пер. с англ. СПб.: Питер, 2003. 877 с.
7. Клименко А. Б., Баринев А. А. Метод управления вычислительными ресурсами распределенных систем на основе «жадной» стратегии и онтологии эффективных алгоритмов // Моделирование, оптимизация и информационные технологии. 2024. Т. 12, № 1. DOI: 10.26102/2310-6018/2024.44.1.018. URL: <https://moitvvt.ru/ru/journal/pdf?id=1508> (дата обращения: 10.02.2026).
8. Колесов Н. В., Литуненко Е. Г., Толмачева М. В., Тюльников В. С. Планирование распределенных вычислений в недетерминированных системах // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 1. С. 106–113. DOI: 10.17586/2226-1494-2025-25-1-106-113. URL: <https://ntv.ifmo.ru/file/article/23325.pdf> (дата обращения: 23.01.2026).
9. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. 2004. URL: <https://research.google.com/archive/mapreduce-osdi04.pdf> (accessed: 26.01.2026).
10. Zaharia M., Chowdhury M., Das T. et al. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing // Proc. NSDI. 2012. URL: <https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf> (accessed: 03.02.2026).
11. Gilbert S., Lynch N. Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services // ACM SIGACT News. 2002. URL: <https://users.ece.cmu.edu/~adrian/731-sp04/readings/GL-cap.pdf> (accessed: 03.02.2026).
12. Ongaro D., Ousterhout J. In Search of an Understandable Consensus Algorithm (Raft). 2014. URL: <https://raft.github.io/raft.pdf> (accessed: 13.02.2026).
13. Lamport L. Time, Clocks, and the Ordering of Events in a Distributed System // Communications of the ACM. 1978. Vol. 21, No. 7. URL: <https://lamport.azurewebsites.net/pubs/time-clocks.pdf> (accessed: 05.02.2026).
14. Ильин В. П. Программирование ближайшего будущего: концепция и прагматика // Вестник РАН. 2023. Т. 93, № 2. С. 150–161. DOI: 10.31857/S086958732302007X. URL: <https://sciencejournals.ru/view-article/?a=Vestnik2302007Ilin&j=vestnik&n=2&v=93&y=2023> (дата обращения: 05.02.2026).

УДК 004.62; ГРНТИ 20.23.17, 50.41.25

ОТЛИЧИЯ ТРАДИЦИОННЫХ МЕТОДОВ ОБРАБОТКИ ДАННЫХ ОТ ПОДХОДОВ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Р.В. Хруничев*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, hrunichev_robert@mail.ru*

Аннотация. Рассмотрены отличия традиционных методов обработки данных (реляционные СУБД, ETL-хранилища данных, OLAP/BI, регламентная отчётность) от подходов анализа больших данных (распределённое хранение и обработка, потоковые конвейеры, Data Lake, масштабируемые вычисления, машинное обучение). Приведены систематизированные различия по характеристикам данных, архитектурам, технологическим стекам, методам обработки, требованиям к качеству и управлению данными. Приведена иллюстрация типовых архитектур и базовые математические соотношения производительности распределённых вычислений.

Ключевые слова: традиционная обработка данных, большие данные, Big Data, хранилище данных, Data Lake, ETL, ELT, OLAP, потоковая обработка, распределённые вычисления, MapReduce, Spark, качество данных, Data Governance.

DIFFERENCES BETWEEN TRADITIONAL DATA PROCESSING METHODS AND BIG DATA ANALYSIS APPROACHES

R.V. Khrunichev*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, hrunichev_robert@mail.ru*

The summary. The article discusses the differences between traditional data processing methods (relational DBMS, ETL data warehouses, OLAP/BI, and regular reporting) and big data analysis approaches (distributed storage and processing, streaming pipelines, Data Lake, scalable computing, and machine learning). The article provides a systematic comparison of data characteristics, architectures, technology stacks, processing methods, quality requirements, and data management. It also illustrates typical architectures and basic mathematical relationships for distributed computing performance.

Keywords: traditional data processing, big data, Big Data, data warehouse, Data Lake, ETL, ELT, OLAP, streaming processing, distributed computing, MapReduce, Spark, data quality, Data Governance.

В классических корпоративных информационных системах данные традиционно обрабатывались в рамках **централизованных** архитектур: транзакционные базы (Online Transaction Processing – OLTP) обслуживают бизнес-операции, а аналитические хранилища (Data Warehouse – DWH) – управленческую отчётность и план-факт анализ. Такой подход предполагает относительно **структурированные** данные, «схему-до-данных» (схема/структура данных определяется и применяется перед загрузкой данных в систему хранения) (schema-on-write) и заранее заданные сценарии использования.

Однако с развитием цифровых технологий, технических средств, телеметрии, мультимедийного контента и истории поведения пользователей возникли задачи обработки массивов, которые характеризуются большими объёмами, **скоростью поступления и разнообразием** (в зарубежной литературе эти признаки зафиксированы как «3V»/«V V V» и далее расширяются дополнительными измерениями, такими как достоверность, ценность и др.) [1–4].

Покажем, **чем** различаются традиционная обработка данных и подходы Big Data не только по объёму данных, но и по архитектуре, методам работы с данными и критериям эффективности.

Под **традиционными методами обработки данных** как правило понимается совокупность практик и технологий, характерных для классических ИТ-инфраструктур:

- реляционные СУБД и нормализованные модели данных;

- извлечение-преобразование-загрузка (ETL) в централизованное хранилище данных;
- аналитика на основе OLAP-кубов, витрин и BI-отчётности;
- акцент на целостность, консистентность и регламентные показатели (KPI) [9, 13].

Под **подходами анализа больших данных** понимается набор принципов и средств, обеспечивающих экономически эффективную обработку данных больших объёмов/скоростей/разнообразия (в том числе слабоструктурированных):

- распределённое хранение (Data Lake, распределённые файловые системы);
- распределённая пакетная и потоковая обработка;
- масштабирование «горизонтально» (кластеры) и перенос логики ближе к данным;
- применение статистики, Data Mining и ML-моделей для поиска закономерностей и прогнозирования [1, 5–8, 10–12, 15].

Традиционные методы обработки данных: сильные стороны и ограничения

1. Архитектурная логика: *OLTP* → *ETL* → *DWH* → *BI*

Классическая схема строится на строгой интеграции данных. Первоначально источники (*OLTP*-системы, файлы, внешние справочники) проходят **ETL**: очистку, приведение типов, согласование справочников, проверку ограничений. Далее данные попадают в *DWH* (обычно на реляционной платформе), где формируются витрины и *OLAP*-представления для отчётности [13].

Достоинства:

- высокая повторяемость и объяснимость отчётов;
- управляемая модель качества (валидируемые правила);
- удобство аудита и SVOT (техническая концепция в компьютеризированном управлении бизнесом).

Недостатки/ограничения:

- высокая стоимость изменений (любая новая сущность/атрибут требует доработки схем, ETL и витрин);
- сложность работы с полу- и неструктурированными данными;
- ограниченная скорость изменения/обновления (часто пакетная загрузка по расписанию, а не мгновенно).

2. Схема-до-данных (*schema-on-write*)

В традиционном подходе данные подлежат аналитике после приведения к согласованной структуре, что является преимуществом для KPI и соответствия (комплаенса), но недостатком для случаев, когда заранее неизвестно, **какие** признаки окажутся полезными (например, для поиска аномалий или обучения модели на логах и текстах).

Подходы анализа больших данных: ключевые идеи

1. Данные как поток и как «озеро»

В Big Data-архитектурах данные зачастую первоначально организуются в «озеро» (Data Lake), где допустимы их разные форматы и уровни «зрелости», а обработка выполняется по схеме получение данных → их интерпретация (очистка, приведение к требуемому формату) → выполнение вычислений. Это «роднит» аналитику с инженерией данных и снижает барьер добавления новых источников.

2. Горизонтальное масштабирование и распределённые системы

Большие данные своим масштабом и организацией требуют применять распределённую обработку. Важным этапом стала модель **MapReduce**, где пользователь задаёт функции *map* и *reduce*, а платформа берёт на себя параллелизм, распределение по узлам и обработку отказов [5].

Развитие экосистемы привело к появлению фреймворков, поддерживающих DAG-вычисления, обработку в памяти, объединение пакетной (batch) и потоковой (streaming) обработки. Например, **Apache Spark** позиционируется как консолидированная аналитическая платформа для крупномасштабной обработки данных и включает модули для SQL, стриминга, ML и графов [7].

3. Ресурсное управление кластерами

Чтобы множество задач и пользователей могли одновременно использовать кластер, применяются диспетчеры ресурсов. В экосистеме Hadoop эта роль связана с YARN (разделение управления ресурсами и управления приложениями) [8].

Сравнительный анализ отличий

Ниже приведены ключевые различия, которые на практике и определяют фундаментальное отличие между традиционной обработкой данных и Big Data-подходами.

1. Характеристики данных: объём, скорость, разнообразие и «неидеальность»

- В традиционном подходе доминируют **структурированные** записи и справочники; типовые объёмы укладываются в вертикально масштабируемую СУБД.
- В Big Data-подходе источники часто дают **разнотипные** данные: логи, тексты, события, изображения, телеметрию; поток может быть непрерывным и высокочастотным. Именно это соответствует определению big data, требующих новых форм обработки [1].

2. Архитектуры и хранение: DWH / Data Lake (и гибридов)

- **DWH**: строгая интеграция, предметная модель, управление изменениями; типовая цель – отчётность и принятие решений.
- **Data Lake**: хранение в исходном виде с последующей обработкой; удобен для экспериментов, ML-проектов, быстрого подключения новых источников.

На практике распространены также **гибриды**: «lakehouse», витрины поверх озера, потоковые витрины для real-time BI и др., но это не замена, а расширение инструментов.

3. Процессы подготовки: ETL в сравнении с ELT/streaming-конвейерами

- Традиционный ETL делает данные структурированными до загрузки в хранилище.
- В Big Data часто встречается ELT (сначала загрузка, потом преобразование), а для потоков – непрерывные конвейеры приём → обработка → витрина/скоринг.

4. Вычисления: одна машина/кластер

Традиционные системы анализа часто оптимизируют запросы в пределах одного сервера или SMP-узла. Big Data анализ предполагает вычисления на **кластерах**. Производительность здесь удобно описывать через ускорение:

$$S(p) = \frac{T_1}{T_p},$$

где T_1 – время выполнения на одном узле, T_p – на p параллельных узлах. Практическое ускорение ограничено последовательной долей работы (закон Амдала):

$$S(p) \leq \frac{1}{(1 - \alpha) + \alpha/p},$$

где α – доля вычислений, выполняемых параллельно.

Отсюда следует важное отличие: главная задача в Big Data – не только сформировать запрос, но и уменьшить последовательные участки, оптимизировать обмен данными и устойчивость к сбоям (сбой узла в большом кластере – статистически нормальное событие).

5. Модели данных и запросы: SQL-запросы/мультязыковой подход

Традиционные подходы к аналитике строятся преимущественно на SQL-запросах. В Big Data же чаще используются различные языки программирования. SQL остаётся важным в таких фреймворках, как Spark SQL и др. [7], но дополняется NoSQL-моделями, графовыми БД, поисковыми индексами, стриминговыми Digital Subscriber Line (DSL) и ML-пайплайнами.

6. Аналитика: отчётность/предиктивная аналитика и рекомендации

- Традиционная аналитика отвечает на вопрос «что произошло?» (descriptive) и «почему?» (diagnostic) посредством KPI, срезов, детализацию данных (drill-down).
- Big Data-аналитика чаще отвечает на вопросы «что будет?» и «что делать?» – прогноз, оптимизация, рекомендации, автоматизация решений.

В исследованиях по интеллектуальному анализу данных, таких, например, как [17] рассматриваются вопросы обработки больших данных нейросетевыми архитектурами, методы векторизации текстов как элемент предобработки для ML [16], а также модели многокритериального анализа данных для систем поддержки решений [18].

7. Качество данных и доверие: «строгая валидация» против вероятностных гарантий

В DWH важны формальные проверки и согласование справочников. В Big Data данные могут быть «зашумленными», неполными, с задержками и, поэтому часто применяются вероятностные подходы (например, робастные оценки устойчивые к наличию в данных аномальных значений (выбросов), методы обработки пропусков, выявление выбросов), а качество измеряется метриками в процессе (data observability).

8. Управление данными и правовые аспекты

Стремительный рост массивов данных усиливает требования к управлению доступом к ним, журналированию, обезличиванию. В РФ с 1 сентября 2025 г. вступили в силу новые требования к обезличиванию персональных данных (приказ Роскомнадзора № 140) [14], который накладывает значительные ограничения и на традиционные хранилища, и на Big Data-озёра, но особенно это критично для lake-подходов, где данные часто хранятся «сырыми».

На рисунке 1 схематично показано различие типовых последовательностей обработки.

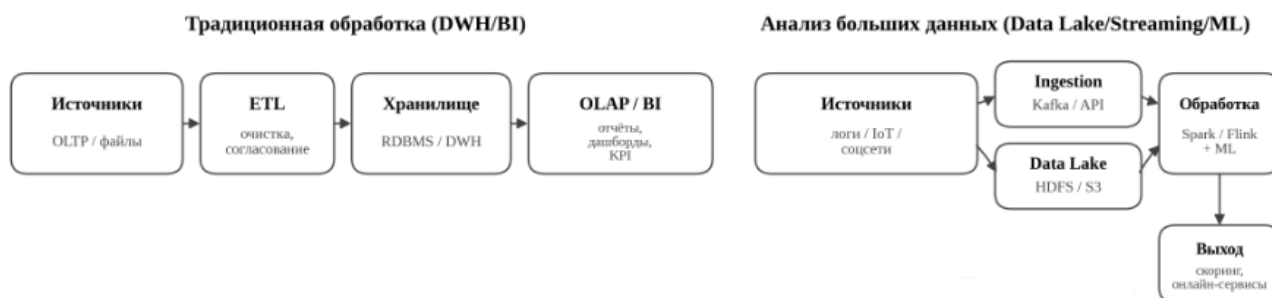


Рис. 1. Этапы обработки в традиционной аналитике (слева) и аналитике больших данных (справа)

Сравнение методов традиционной обработки данных и подходов анализа больших данных представлено в таблице 1.

Таблица 1. Сравнение методов традиционной обработки данных и подходов анализа больших данных

Критерий / методический аспект	Традиционная обработка данных (OLTP/ETL/DWH/OLAP/BI)	Анализ больших данных (Data Lake/Distributed/Streaming/ML)
Типичные данные	Структурированные таблицы, справочники	Разнотипные: логи, события, тексты, графы, мультимедиа
Принцип организации	Schema-on-write : сначала схема данных, потом загрузка	Часто schema-on-read : храним данные «как есть», интерпретируем при чтении
Основной конвейер	OLTP → ETL → DWH → витрины/OLAP → отчёты	Источники → ingestion → Data Lake → распределённая обработка → сервисы/витрины
Подготовка данных	Жёсткая очистка/согласование до аналитики	Итеративная подготовка, feature engineering, возможны «сырые» данные
Пакетная обработка	Основной режим	Есть, но чаще вместе с потоками
Потоковая обработка	Ограничена и часто отдельными решениями	Базовый сценарий – потоковая обработка
Хранилище	DWH на системе управления реляционными базами данных, витрины, индексирование под запросы	Data Lake, распределённые форматы (Parquet/ORC), lakehouse-подходы
Масштабирование	Чаще вертикальное	Горизонтальное, кластер/облако при необходимости
Параллелизм	Внутри СУБД: оптимизатор, параллельные планы	Явный/неявный в движке: распределение по узлам, направленный ациклический граф (DAG), партиции
Типовые модели вычислений	SQL-запросы, OLAP-кубы, процедуры	MapReduce/DAG, Spark jobs, streaming pipelines, ML-пайплайны
Задержка получения результата	Минуты/часы	От миллисекунд/секунд (потоковая онлайн) до часов (пакетная)
Отказоустойчивость	Надёжность обеспечивается СУБД/СХД, сбой – редкое событие	Сбой узла статистически нормален: репликация, контрольные точки (checkpoint), однократная доставка (retry), повтор сообщений (exactly-once/at-least-once)
Языки и инструменты	SQL, ETL-инструменты, BI-платформы	SQL+DSL (stream), Python/Scala, ML-фреймворки, оркестрация
Основной результат	Отчёты, KPI, план-факт, разовые (ad-hoc) аналитические запросы	Модели, онлайн-скоринг, рекомендации, выявление аномалий, интерактивные (real-time) витрины
Подход к качеству данных	Жёсткие правила качества до публикации в витрины	Наблюдаемость данных, метрики качества в потоке, допуски/обработка «зашумленных» данных
Обновления и изменения	Дорогостоящее изменение схемы/ETL, долгосрочный цикл	Добавление источников и признаков быстрее, но выше риск получения Data Swamp (болота данных)
Стоимость	Часты капитальные затраты, лицензии, дорогостоящее оборудование	Часты операционные расходы, облако, оплата за ресурсы. Есть экономия на масштабе, но дороже сопровождение
Управление доступом	Ролевые модели в СУБД, централизованный контроль	Сложнее из-за множества слоёв: lake + движки + каталоги + политики
Типовые риски	Очередь обработки в ETL/хранилище, медленные изменения	Сложность эксплуатации, рост стоимости доработок, не гарантирована доставка/повторы

Справедливо рассматривать подходы традиционной аналитики и аналитики Big Data как дополняющие друг друга:

- DWH/BI остаётся «мощным» инструментом, когда нужна управляемая отчётность, аудит, единые справочники, стабильные показатели.
- Big Data-подходы важны, когда источники быстро меняются, много неструктурированных данных, требуется обработка в реальном времени (real-time) и ML-продукты.

Поэтому «зрелое» управление ИТ-ландшафтом часто включает и контур DWH, и Data Lake/поток, с согласованными политиками качества, доступа и жизненного цикла данных.

Библиографический список

1. Gartner. Big Data. URL: <https://www.gartner.com/en/information-technology/glossary/big-data> (дата обращения: 10.02.2026).
2. Laney D. Deja VVVu: Gartner's Original "Volume-Velocity-Variety" Definition of Big Data. 2012. URL: <https://community.aiim.org/blogs/doug-laney/2012/08/25/deja-vvvu-gartners-original-volume-velocity-variety-definition-of-big-data> (дата обращения: 02.02.2026).
3. Ward J. S., Barker A. A Survey of Big Data Definitions. arXiv. 2013. URL: <https://arxiv.org/pdf/1309.5821> (дата обращения: 02.02.2026).
4. De Mauro A. и др. What is Big Data? A Consensual Definition and a Review of Key Research Topics. 2014. URL: https://www.researchgate.net/profile/Andrea-De-Mauro-2/publication/265775800_What_is_Big_Data_A_Consensual_Definition_and_a_Review_of_Key_Research_Topics/links/54e61d170cf277664ff2f0b4/What-is-Big-Data-A-Consensual-Definition-and-a-Review-of-Key-Research-Topics.pdf (дата обращения: 09.02.2026).
5. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. OSDI'04. 2004. URL: <https://research.google.com/archive/mapreduce-osdi04.pdf> (дата обращения: 10.02.2026).
6. Apache Hadoop. Project information and modules (HDFS, YARN, MapReduce). URL: <https://hadoop.apache.org/> (дата обращения: 25.01.2026).
7. Apache Spark. Overview. Documentation. URL: <https://spark.apache.org/docs/latest/> (дата обращения: 25.01.2026).
8. Vavilapalli V. K. и др. Apache Hadoop YARN: Yet Another Resource Negotiator. 2013. URL: https://www.cse.ust.hk/~weiwa/teaching/Fall15-COMP6611B/reading_list/YARN.pdf (дата обращения: 07.02.2026).
9. Вольфсон М. Б. Анализ данных. Учебное пособие. СПб.: СПбГУТ, 2015. 82 с. URL: https://www.sut.ru/doci/umu/disc/3471/3_Posobie.pdf (дата обращения: 07.02.2026).
10. Кобзаренко Д. Н. Анализ больших данных. Учебное пособие. Махачкала, 2019. URL: <https://www.apap.kg/uploads/Images%20content/biblioteka/E-books/%D0%9A%D0%BE%D0%B1%D0%B7%D0%B0%D1%80%D0%B5%D0%BD%D0%BA%D0%BE%20%D0%94%D0%9D.%20%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7%20%D0%B1%D0%BE%D0%B%D1%8C%D1%88%D0%B8%D1%85%20%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85.%20-%20%D0%9C%D0%B0%D1%85%D0%B0%D1%87%D0%BA%D0%B0%D0%BB%D0%B0%2C%202019.pdf> (дата обращения: 06.02.2026).
11. Тесленко И. Б. Big Data (Большие данные). Учебное издание. 2021. URL: <https://dspace.www1.vlsu.ru/bitstream/123456789/9407/1/02292.pdf> (дата обращения: 07.02.2026).
12. Самсонов А. Д. Обзор методик обработки и хранения больших данных. 2022. URL: <https://sanse.ru/index.php/sanse/article/download/557/506> (дата обращения: 06.02.2026).
13. Хранилище данных (Data Warehouse). Loginom Wiki. URL: <https://wiki.loginom.ru/articles/%D0%A5%D1%80%D0%B0%D0%BD%D0%B8%D0%BB%D0%B8%D1%89%D0%B5%20%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85%20%28Data%20Warehouse%29.%20Loginom%20Wiki.pdf> (дата обращения: 29.01.2026).
14. Роскомнадзор. Приказ от 19.06.2025 № 140 «Об утверждении требований к обезличиванию персональных данных...» (применяется с 01.09.2025). URL: <https://www.consultant.ru/law/hotdocs/90337.html> (дата обращения: 10.02.2026).
15. Зейналлы Т. Э., Демидов Д. Г. Прогнозирование нагрузки с использованием рекуррентных нейронных сетей для определения приоритета узла распределенной системы. Вестник Рязанского государственного радиотехнического университета. 2024. № 90. С. 54–66. DOI: 10.21667/1995-4565-2024-90-54-66. URL: <https://vestnik.rsreu.ru/images/archive/2024/90/2.2-.pdf> (дата обращения: 12.02.2026).

16. Попова О. А. Анализ методов векторизации текстовых документов. Вестник Рязанского государственного радиотехнического университета. 2023. № 85. С. 96–102. DOI: 10.21667/1995-4565-2023-85-96-102. URL: https://vestnik.rsreu.ru/images/archive/2023/85/3.5_.pdf (дата обращения: 09.02.2026).

17. Демидова Л. А., Морошкин Н. А. Аспекты разработки архитектуры вопросно-ответной системы для обработки больших данных на основе нейросетевого моделирования. Вестник Рязанского государственного радиотехнического университета. 2023. № 86. С. 145–155. URL: <https://vestnik.rsreu.ru/ru/archive/2023/vypusk-86> (дата обращения: 06.02.2026).

18. Щенёва Ю. Б., Пылькин А. Н., Щенёв Е. С. Модель многокритериального анализа данных для системы поддержки принятия решений. Вестник Рязанского государственного радиотехнического университета. 2025. № 94. С. 124–138. DOI: 10.21667/1995-4565-2025-94-124-138. URL: <https://vestnik.rsreu.ru/images/archive/2025/94/4.2-.pdf> (дата обращения: 07.02.2026).

УДК 004.932; ГРНТИ 28.23.15

ИССЛЕДОВАНИЕ АЛГОРИТМОВ МАСКИРОВАНИЯ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ ЯДЕР РАЗМЫТИЯ СЛОЖНОЙ ФОРМЫ

К.А. Васильев

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, cornholio62@yandex.ru*

Аннотация. В работе приведено исследование процессов несанкционированного восстановления третьими лицами маскированной информации. Маскирование производится с помощью размытия изображений ядрами сложной формы. Свертка изображения с матричным ядром описывается интегральным уравнением Фредгольма первого рода. Приведено краткое описание алгоритма восстановления размытого изображения с помощью регуляризационного алгоритма. В экспериментальной части приведены результаты размытия двух типов изображений ядрами сложной формы и попыток восстановления размытых изображений регуляризационным методом, но с простыми ядрами. Подтверждена невозможность корректного восстановления информации при таком способе несанкционированного вмешательства в работу системы.

Ключевые слова: размытие, маскирование, ядро размытия, регуляризационный метод.

RESEARCH OF IMAGE HIDING ALGORITHMS USING COMPLEX-SHAPED BLUR KERNELS

K.A. Vasilev

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, cornholio62@yandex.ru*

The summary. The paper presents a study of the processes of unauthorized recovery of masked information by third parties. Masking is performed by blurring images with complex-shaped kernels. The convolution of an image with a matrix kernel is described by the Fredholm integral equation of the first kind. The paper provides a brief description of the algorithm for restoring a blurred image using a regularization algorithm. The experimental section presents the results of blurring two types of images with complex-shaped kernels and attempts to restore the blurred images using a regularization method with simple kernels. The impossibility of correctly restoring information in this way of unauthorized interference with the system has been confirmed.

Keywords: blurring, masking, blurring kernel regularization method.

В цифровой обработке изображений размытие обычно рассматривается как нежелательный артефакт, возникающий из-за движения камеры, дефокусировки или атмосферных помех. Однако намеренное размытие изображений также является мощным инструментом для защиты конфиденциальной информации, такой как лица людей, номерные знаки автомобилей, текстовые данные на документах или чувствительные элементы ландшафта на спутниковых снимках. В этом контексте размытие выступает как метод маскирования, позволяющий скрыть детали, сохранив общую структуру сцены. Эффективность такой маскировки напрямую зависит от формы и параметров ядра размытия и(функции, описывающей, как одна точка изображения «расплывается» по соседним пикселям).

Классические подходы, такие как равномерное или гауссово размытие, хотя и просты в реализации, обладают существенным недостатком с точки зрения безопасности: при наличии соответствующих компетенций их можно подобрать при несанкционированном доступе к маскированной информации. Зная или восстановив простое ядро, злоумышленник может применить обратную фильтрацию и частично или полностью восстановить скрытую информацию. Это делает стандартные методы уязвимыми для атак, особенно в эпоху развития глубокого обучения и алгоритмов слепого восстановления.

Данное исследование фокусируется на ядрах размытия сложной формы как на перспективном направлении для создания более устойчивых методов маскирования визуальных данных. В отличие от симметричных и гладких ядер, сложные ядра позволяют эффективнее скрывать информацию за счет нарушения предсказуемой структуры искажений. Они затрудняют слепое восстановление, так как задача оценки параметров такого ядра становится существенно сложнее.

Таким образом, переход от простых к сложным ядрам размытия открывает новые возможности не только в традиционной задаче восстановления изображений, но и в обратной задаче – создании необратимого или устойчивого к восстановлению маскирования, что является критически важным для обеспечения приватности в цифровую эпоху.

Размытие изображений

Процесс размытия сигнала описывается интегральным уравнением Фредгольма 1-го рода [1, 2]

$$\int_a^b K(x, s) \cdot \phi(s) ds = f(x), \quad a \leq x, s \leq b \quad (1)$$

– для одномерного сигнала и

$$\int_{ca}^{db} \int K(x, y, t, s) \cdot \varphi(t, s) dt ds = f(x, y), \quad x, y \in [a, b] \times [c, d] \quad (2)$$

– для двумерного (изображения).

В (1) и (2) $K(x, s)$, $K(x, y, t, s)$ – ядра интегрального уравнения (аппаратная функция прибора);

$f(x)$, $f(x, y)$ – наблюдаемые сигналы (изображения);

$\phi(s)$, $\varphi(t, s)$ – истинные сигналы, которые должны быть найдены в результате решения уравнений (1) и (2).

Для формирования ядра размывающего оператора использовались порождающие функция вида:

$$K_1(x, s) = \lambda e^{-\frac{(x-s)^2}{\sigma^2}}, \quad K_2(x, s) = \lambda e^{-\frac{x}{d}} * \cos(\pi * \frac{x}{T}) \quad (2)$$

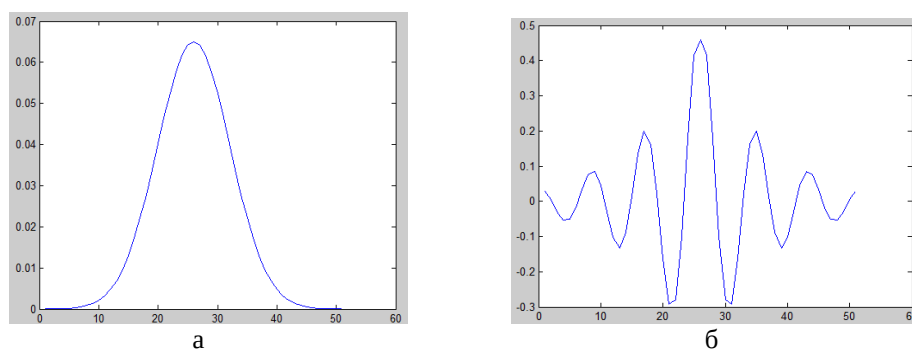
где $\lambda = 1$. Функция (2) определена на всей числовой оси, но при этом $K(x, s) \rightarrow 0$ при $x \rightarrow \infty$.

Поэтому «хвосты» функции $K(x, s)$ обрезаются. В результате получается вектор-строка

$(r_k, r_{k-1}, \dots, r_1, r_0, r_1, \dots, r_{k-1}, r_k)$ конечной длины $(2k+1)$. Подбор параметров в формуле (2), обеспечивающих формирование вектор-строки такой длины достигается условием $r_k = 10^{-6}$ и

условием $\sum_{i=-k}^k r_i = 1$. В итоге, разностные ядра (2) длины $2k+1=51$ имеют следующий вид

(рис. 1):

Рис. 1. Ядро размытия K_1 (а) и K_2 (б)

Для восстановления и имитации «атаки» на размытое изображение использовался регуляризационный метод восстановления. Алгоритм сводится к решению параметрического семейства систем линейных алгебраических уравнений (СЛАУ):

$$(\mathbf{K}^T \mathbf{K} + \alpha \mathbf{E}) \cdot \mathbf{I} = \mathbf{K}^T \mathbf{I}^{sgl} \quad (4)$$

В составе этой системы $\mathbf{K}$ – дискретный аналог ядра интегрального уравнения, $\mathbf{E}$ – единичная матрица, α – параметр регуляризации, $\mathbf{I}^{sgl}$ – размытое изображение.

Строки матрицы $\mathbf{K}$ образованы элементами порождающего вектора $(k_{-n}, k_{-n+1}, \dots, k_{-1}, k_0, k_1, \dots, k_{n-1}, k_n)$, в составе которого $k_j = \mu \cdot \exp(-\beta \cdot j^2)$ и $k_{-j} = k_j$, $j = \overline{1, n}$. Каждая следующая строка матрицы $\mathbf{K}$ получается из предыдущей сдвигом приведенного вектора длины $2n+1$ на одну позицию вправо. Остальные элементы матрицы заполнены нулями. В результате матрица является симметричной и ленточной, а матрица $\mathbf{K}^T \mathbf{K}$ – еще и положительно определенной теплицевой матрицей [2]. СЛАУ (4) целесообразно решать методом, описанным в [3] специально для решения СЛАУ с теплицевой основной матрицей.

Общая схема алгоритма:

- задается монотонно убывающая последовательность $\alpha_0, \alpha_1, \dots, \alpha_m$ значений параметра регуляризации α ;
 - для каждого значения α_k , $k = 0, 1, \dots, m$ параметра регуляризации α_k формируется и решается СЛАУ (4). В результате получается параметрическое семейство решений $\hat{\mathbf{I}}_k = \hat{\mathbf{I}}(\alpha_k)$, $k = 1, 2, \dots$;
 - начиная со второго шага (со второй СЛАУ) вычисляется в евклидовой метрике уклонение $\delta_k = \|\hat{\mathbf{I}}^{(k)} - \hat{\mathbf{I}}^{(k-1)}\|$ текущего вектора $\hat{\mathbf{I}}^{(k)}$ от предыдущего $\hat{\mathbf{I}}^{(k-1)}$;
 - начиная с четвертого шага проверяется выполнение неравенств $\delta_{k-1} > \delta_k < \delta_{k+1}$
- Если для некоторого k_0 эти неравенства выполняются, то соответствующее значение параметра регуляризации α_{k_0} принимается за оптимальное, а решение $\hat{\mathbf{I}}_{opt} = \hat{\mathbf{I}}(\alpha_{k_0})$ – за квазиоптимальное решение.

Экспериментальная часть

Для исследований использовались изображения спутникового снимка и структурной схемы с текстом, как показано на рисунках 2,а и 2,б соответственно.

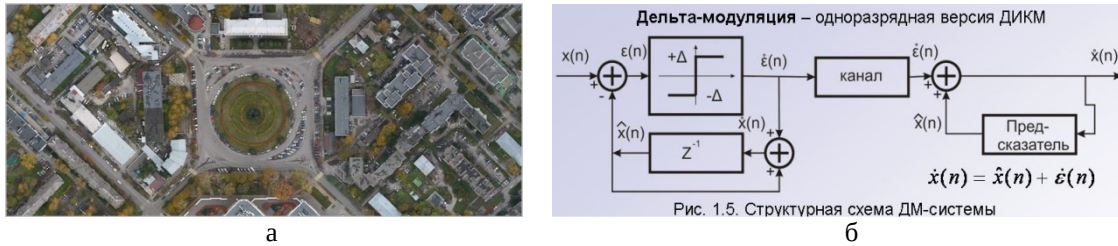


Рис. 2. Исходные изображения спутникового снимка (а) и структурной схемы (б)

Далее моделировался несанкционированный доступ к изображениям размытым сложным ядром K_2 с попыткой восстановления их обычным гауссовским ядром K_1 того же размера. На рисунке 3 приведены размытое спутниковое изображение (рис. 3,а) и результат его восстановления простым гауссовским ядром той же длины (рис. 3,б).

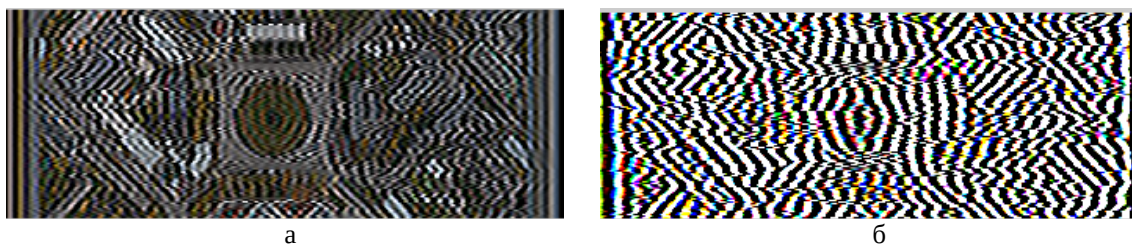


Рис. 3. Размытое изображение спутникового снимка сложным ядром (а) и попытка восстановления простым гауссовским ядром той же длины (б)

На рисунках 4,а и 4,б приведены аналогичные по смыслу результаты размытия (рис. 4,а) и восстановления (рис.4,б) изображения схемы.

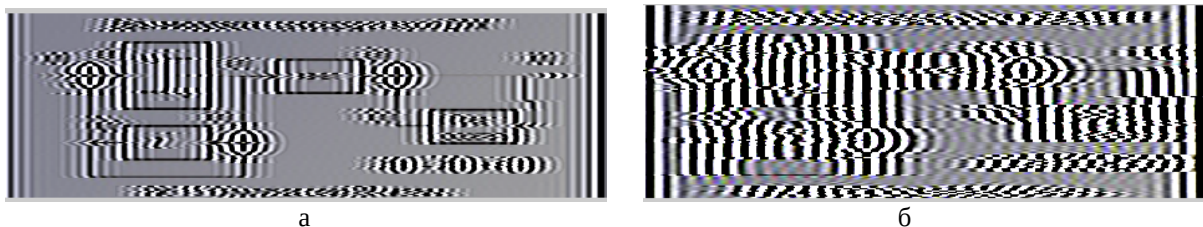


Рис. 4. Размытое изображение схемы сложным ядром (а) и попытка восстановления простым гауссовским ядром той же длины (б)

Даже визуальный анализ результатов восстановления как в первом, так и во втором случаях позволяет заключить, что попытка несанкционированного восстановления оказалась безуспешной. Подтверждается этот вывод и оценками среднеквадратического отклонения σ :

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (I_i^{\text{восст}} - I_i^{\text{ориг}})^2}.$$

В обоих случаях оценка СКО оказалась равной приблизительно 88.

Теперь оценим восстанавливающую способность простого ядра, но с другим размером, отличным от размера сложного ядра, которым размывали изображения. Для примера возьмем размер восстанавливающего ядра примерно в 2 раза меньше, чем размер размывающего. То есть смоделируем ситуацию, когда злоумышленнику совсем ничего не известно о ядре размытия.

На рисунках 5 и 6 показан результат такого восстановления. Визуально восстановление прошло еще хуже, чем в предыдущем варианте. Среднеквадратичное отклонение при этом выросло на 35% и составило 118 для каждого изображения.

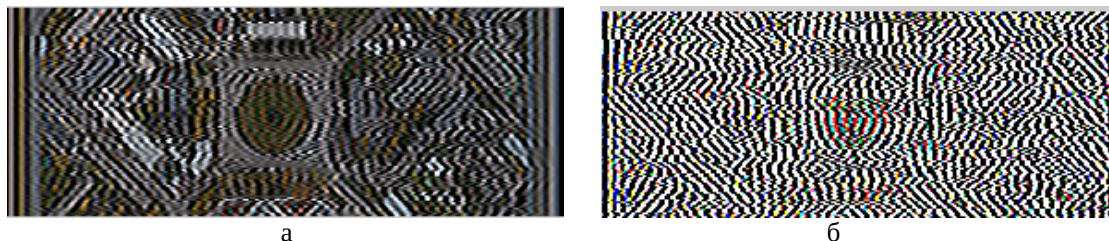


Рис. 5. Размытое изображение спутникового снимка сложным ядром (а) и попытка восстановления простым гауссовским ядром другой длины (б)

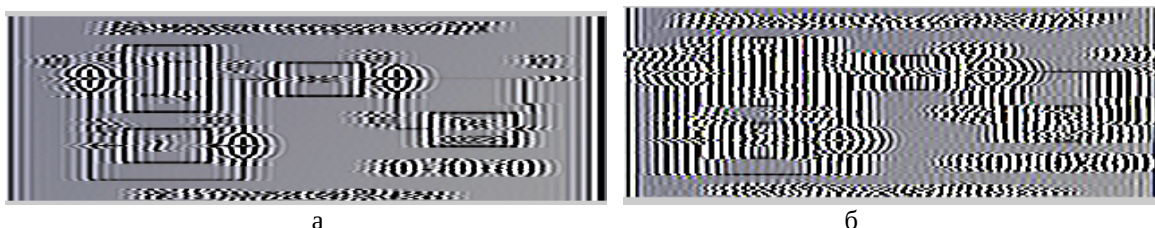


Рис. 6. Размытое изображение схемы сложным ядром (а) и попытка восстановления простым гауссовским ядром другой длины (б)

На рисунках 5 и 6 показан результат такого восстановления. Визуально восстановление прошло еще хуже, чем в предыдущем варианте. Среднеквадратичное отклонение при этом выросло на 35% и составляет 118 для обоих изображений.

Заключение

Рассмотрен процесс размытия изображения с помощью сложного ядра, как процесс маскирования информации, с целью защиты ее от злоумышленников. Сначала было произведено размытие изображений сложным ядром, затем имитировалась атака на размытое изображение, путем восстановления обычным гауссовским ядром, при этом размер ядра был одинаков. Далее исследовалась ситуация, когда размер ядра неизвестен. В обоих случаях восстановление не удалось, причем в случае, когда длина ядра неизвестна, результат восстановления оказался еще хуже. Тем самым подтверждена гипотеза о возможности использования размытия изображений как криптографического инструмента в области защиты визуальных данных

Библиографический список

1. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. - М.: Наука, 1986. – 288 с.
2. Васильев К.И., Новиков А.И. Регуляризационный алгоритм восстановления размытого изображения в задаче маскирования информации // XXI век: итоги прошлого и проблемы настоящего плюс. Т. 13, № 3(67), 2024. – С. 8-16.
3. Краснов М.Л. Интегральные уравнения. – М.: Наука, 1975. – 301 с.
4. Бакушинский А.Б., Гончарский А.В. Итеративные методы решения некорректных задач. - М.: Физматлит, 1989. – 130 с.

УДК 004.7; ГРНТИ 50.39.19

ОБЛАЧНЫЕ РЕШЕНИЯ И ИНФРАСТРУКТУРНЫЕ СЕРВИСЫ ДЛЯ АНАЛИЗА ДАННЫХ (AMAZON WEB SERVICES, GOOGLE CLOUD PLATFORM, YANDEX CLOUD)

Р.В. Хруничев

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, hrunichev_robert@mail.ru*

Аннотация. Облачные решения предоставляют организациям масштабируемые вычислительные ресурсы и управляемые сервисы для полного цикла работы с данными: от сбора и хранения до обработки, машинного обучения и визуализации. В работе приводится классификация моделей предоставления сервисов и рассматриваются инфраструктурные сервисы для анализа данных на примере платформ Amazon Web Services (AWS), Google Cloud Platform (GCP) и Yandex Cloud. Выполнен сравнительный анализ по ключевым функциональным возможностям, результаты которого могут быть использованы при выборе облачной платформы для аналитических задач с учётом требований к локализации данных, уровню управляемости и интеграции сервисов.

Ключевые слова: облачные вычисления (cloud computing), инфраструктура как услуга (Infrastructure as a Service, IaaS), платформа как услуга (Platform as a Service, PaaS), озеро данных (data lake), хранилище данных (data warehouse), потоковая обработка (stream processing), извлечение-преобразование-загрузка (Extract-Transform-Load, ETL), машинное обучение (machine learning), бизнес-аналитика (business intelligence, BI).

CLOUD SOLUTIONS AND INFRASTRUCTURE SERVICES FOR DATA ANALYSIS (AMAZON WEB SERVICES, GOOGLE CLOUD PLATFORM, YANDEX CLOUD)

R.V. Khrunichev

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, hrunichev_robert@mail.ru*

The summary. Cloud solutions provide organizations with scalable computing resources and managed services for the full data lifecycle, from collection and storage to processing, machine learning, and visualization. This paper provides a classification of service delivery models and examines infrastructure services for data analysis using the Amazon Web Services (AWS), Google Cloud Platform (GCP), and Yandex Cloud platforms as examples. It also conducts a comparative analysis of key functionalities, which can be used to choose the right cloud platform for analytical tasks based on data localization requirements, service management levels, and integration capabilities.

Keywords: cloud computing, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), data lake, data warehouse, streaming processing, Extract-Transform-Load, (ETL), machine learning, business intelligence (BI).

Рост объёмов и разнообразия данных в информационных системах приводит к усложнению аналитических платформ: требуется одновременная поддержка пакетной обработки (batch processing), потоковой обработки (stream processing), интерактивных запросов (interactive queries) и задач машинного обучения (machine learning). Традиционная модель развёртывания аналитической инфраструктуры (на собственных серверах) сопровождается значительными затратами на закупку, эксплуатацию и масштабирование оборудования, а также на сопровождение программной платформы.

Облачные вычисления (cloud computing) рассматриваются как модель предоставления вычислительных ресурсов с возможностью быстрого выделения и освобождения, а также с измеримостью потребления ресурсов [10]. В определении [1] подчёркиваются их ключевые характеристики: самообслуживание по требованию (on-demand self-service), широкий сетевой доступ (broad network access), объединение ресурсов (resource pooling), эластичность (rapid elasticity) и измеряемое обслуживание (measured service).

В нормативных документах [2] облачные вычисления определяются как парадигма, обеспечивающая сетевой доступ к масштабируемому и эластичному пулу разделяемых ресурсов.

Для задач анализа данных ключевыми являются понятия:

- озеро данных (data lake) – логическая зона хранения данных в «сыром» или слабо структурированном виде, часто поверх объектного хранилища (object storage);
- хранилище данных (data warehouse) – специализированная система хранения и обработки аналитических данных с поддержкой SQL;
- витрина данных (data mart) – тематически ориентированная часть хранилища данных;
- конвейер данных (data pipeline) – последовательность этапов сбора, переноса, преобразования и загрузки данных.

Модели предоставления: IaaS, PaaS, SaaS

Традиционно выделяют:

- инфраструктуру как услугу (Infrastructure as a Service, IaaS) – виртуальные машины, сети, диски, базовые хранилища;
- платформу как услугу (Platform as a Service, PaaS) – управляемые платформенные компоненты (кластеры обработки, базы данных, очереди сообщений, сервисы интеграции);
- программное обеспечение как услугу (Software as a Service, SaaS) – конечные прикладные сервисы (например, BI-платформы, среды разработки).

Для аналитики на практике используется комбинирование моделей [11]: данные хранятся в объектном хранилище (object storage), обработка выполняется на управляемых кластерах (managed clusters) или в бессерверных средах (serverless), визуализация – в BI-сервисах.

Инфраструктурные сервисы для анализа данных

Приведём функциональный минимум, который чаще всего требуется для построения аналитического решения в облаке.

1. Сбор и доставка данных: очереди сообщений (messaging), потоковые шины (streaming), коннекторы переноса (data transfer).

2. Хранение данных: объектное хранилище (object storage), базы данных (databases), хранилища данных (data warehouses).

3. Обработка: пакетная обработка (batch), потоковая обработка (stream), интерактивные SQL-запросы (interactive SQL).

4. Машинное обучение: среда экспериментов (notebooks), обучение (training), развёртывание (serving).

5. Визуализация и доступ: бизнес-аналитика (business intelligence, BI), панели мониторинга (dashboards), отчётность.

6. Управление и безопасность: управление идентификацией и доступом (Identity and Access Management, IAM), журналы (logging), мониторинг (monitoring), политики и соответствие требованиям (compliance) [9].

Платформы и сервисы

1. Yandex Cloud (Яндекс Облако)

Yandex Cloud предоставляет набор управляемых сервисов для аналитического цикла: объектное хранилище (object storage), управляемые кластеры обработки (managed clusters) для задач пакетной обработки (batch processing), средства потоковой передачи (streaming), инструменты интерактивного SQL-анализа и сервисы визуализации. Официальная документация систематизирует сервисы по назначению и сценариям применения [4, 7].

Практико-ориентированный обзор построения аналитических проектов в облаке изложен в методических материалах Yandex Cloud, где сервисы сгруппированы по этапам аналитического проекта [3].

2. Amazon Web Services (AWS)

AWS предоставляет набор инфраструктурных и платформенных сервисов, обеспечивающих построение озера данных (data lake) на базе объектного хранилища (object storage), выполнение управляемой кластерной обработки (managed processing), реализацию извлечения-преобразования-загрузки (ETL), интерактивные запросы к данным (interactive queries), потоковую обработку (stream processing), а также поддержку полного жизненного цикла моделей машинного обучения (machine learning lifecycle). Официальная документация описывает данные сервисы, их ограничения и рекомендации по архитектуре [5].

3. Google Cloud Platform (GCP)

Google Cloud Platform предоставляет инфраструктурные сервисы хранения данных (storage), управляемую обработку (managed processing), потоковую доставку сообщений и событий (messaging/streaming), специализированное хранилище данных (data warehouse), сервисы машинного обучения (machine learning services) и средства визуализации аналитики. В документации отражены принципы построения аналитических контуров и эксплуатационные характеристики ключевых сервисов, включая особенности модели расчёта стоимости обработки запросов (query processing cost) [6].

Сравнительный анализ по критериям

1. Уровень управляемости и эксплуатационные затраты

Для анализа данных наиболее важны управляемые сервисы (managed services), так как они снижают долю трудозатрат на администрирование и повышают воспроизводимость процессов. Все три платформы предоставляют управляемые компоненты хранения, обработки и визуализации, при этом различия проявляются в деталях интеграции сервисов, типовых архитектурных шаблонах и механизмах управления данными [3-6].

2. Поддержка типового аналитического контура

Типовой контур «поток → хранение → обработка → витрины → BI» может быть реализован на каждой платформе средствами её стандартных сервисов и рекомендуемых архитектурных схем [4-6].

Для проектирования предъявляются существенные требования к задержке получения результата (latency), объёму данных (data volume), режиму обработки (batch/stream), а также к уровню автоматизации эксплуатации (operations automation).

3. Локализация данных и нормативные требования

Для организаций, на которые распространяются требованиями по размещению данных и инфраструктуры на территории РФ, существенным фактором выбора является соответствие требованиям обработки и хранения данных. В таких случаях отечественные облачные платформы часто рассматриваются как приоритетный (а чаще единственный) вариант развертывания инфраструктуры. При проектировании также необходимо учитывать требования к обезличиванию персональных данных и связанные нормативные документы [8].

4. Упрощенная экономическая модель

Суммарные затраты на облачную аналитику можно представить в виде:

$$C_{\Sigma} = C_{\text{хран.}} + C_{\text{вычисл.}} + C_{\text{сет.}} + C_{\text{эксп.}},$$
 где $C_{\text{хран.}}$ – затраты на хранение данных (storage) (диски, репликации, резервные копии и т.д.), $C_{\text{вычисл.}}$ – затраты на вычислительные процессы (compute) (CPU/GPU, кластеры Spark/Flink и т.д.), $C_{\text{сет.}}$ – затраты на сеть и передачу данных (transfer) (трафик в облачных сервисах, сетевые устройства и т.д.), $C_{\text{эксп.}}$ – затраты на эксплуатацию и сопровождение (operations).

Экономическая эффективность существенно зависит от архитектурных решений: уровня агрегации, форматов хранения, частоты перерасчётов и выбора между выделенными

и бессерверными ресурсами. Приведённые особенности подробно отражаются в документации платформ, включая описание моделей тарификации отдельных сервисов [4-6].

В таблице 1 приведены функциональные категории для сопоставления по официальной документации соответствующих платформ [4-6].

Таблица 1. Сравнение облачных платформ

Функциональная категория	Amazon Web Services (AWS)	Google Cloud Platform (GCP)	Yandex Cloud
Объектное хранилище (object storage)	объектное хранилище (object storage) платформы	объектное хранилище (object storage) платформы	объектное хранилище (object storage) платформы
Управляемая кластерная обработка (managed Spark/Hadoop)	управляемая обработка (managed processing)	управляемая обработка (managed processing)	управляемая обработка (managed processing)
Потоковая доставка/шина событий (streaming/messaging)	потоковые сервисы (streaming services)	обмен сообщениями/потоки (messaging/streaming)	потоковые сервисы (streaming services)
Интеграция/ETL (ETL)	средства интеграции и ETL (ETL services)	средства построения конвейеров (data pipelines)	средства переноса и конвейеры (data transfer/pipelines)
Интерактивные SQL-запросы (interactive SQL)	интерактивные запросы (interactive queries)	интерактивная аналитика (interactive analytics)	интерактивный анализ (interactive analysis)
Хранилище данных (data warehouse)	хранилище данных (data warehouse)	хранилище данных (data warehouse)	аналитическое хранилище/СУБД (analytic database)
Машинное обучение (machine learning)	платформа машинного обучения (ML platform)	платформа машинного обучения (ML platform)	платформа машинного обучения (ML platform)
Визуализация и BI (business intelligence, BI)	BI-инструменты экосистемы (BI tools)	BI-инструменты экосистемы (BI tools)	BI-инструменты экосистемы (BI tools)

Облачные решения формируют технологическую основу современной аналитики, предоставляя стандартизованные характеристики облачных вычислений и широкий спектр инфраструктурных сервисов: хранение, обработку, потоковую интеграцию, машинное обучение и визуализацию. Сравнение Amazon Web Services, Google Cloud Platform и Yandex Cloud показывает сопоставимость функциональных возможностей по основным категориям задач при различиях в практических сценариях применения, механизмах управления, а также особенностях расчёта стоимости и требований к локализации данных. При выборе платформы целесообразно оценивать нормативные ограничения, характер нагрузок, требования к задержкам и уровню управляемости сервисов.

Библиографический список

1. Mell P., Grance T. The NIST Definition of Cloud Computing. NIST Special Publication 800-145. 2011. URL: <https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-145.pdf> (дата обращения: 28.01.2026).
2. ГОСТ ISO/IEC 17788-2016. Информационные технологии. Облачные вычисления. Общие положения и терминология. URL: <https://protect.gost.ru/document.aspx?control=7&id=205502> (дата обращения: 02.02.2026).
3. Как делать аналитические проекты в облаке. Методические материалы Yandex Cloud. URL: https://storage.yandexcloud.net/yc-materials/pdf/Analytical%20guide_RU.pdf (дата обращения: 10.02.2026).
4. Yandex Cloud. Документация платформы и сервисов для анализа данных: объектное хранилище (Object Storage), управляемая обработка (Data Proc), потоковые сервисы (Data Streams), интерактивный анализ

(Query), аналитические базы данных (Managed Service for ClickHouse), визуализация (DataLens), машинное обучение (DataSphere). URL: <https://yandex.cloud/ru/docs> (дата обращения: 25.01.2026).

5. Amazon Web Services. Документация платформы и сервисов для анализа данных: объектное хранилище (Amazon S3), управляемая обработка (Amazon EMR), ETL (AWS Glue), интерактивные запросы (Amazon Athena), хранилище данных (Amazon Redshift), потоковая обработка (Amazon Kinesis), машинное обучение (Amazon SageMaker). URL: <https://docs.aws.amazon.com/> (дата обращения: 25.01.2026).

6. Google Cloud. Документация платформы и сервисов для анализа данных: объектное хранилище (Cloud Storage), хранилище данных (BigQuery), управляемая обработка (Dataproc), обмен сообщениями и потоки (Pub/Sub), машинное обучение (Vertex AI), визуализация (Looker). URL: <https://cloud.google.com/docs> (дата обращения: 28.01.2026).

7. Баланов А. Н. Облачные технологии. Учебное пособие. СПб.: Лань, 2024.

8. Роскомнадзор. Приказ от 19.06.2025 № 140 «Об утверждении требований к обезличиванию персональных данных...» (применяется с 01.09.2025). URL: <https://www.consultant.ru/law/hotdocs/90337.html> (дата обращения: 29.01.2026).

9. Пименова А. А., Никитин Д. Д., Никишин К. И. Моделирование сценариев безопасности в программно-конфигурируемых сетях. Вестник Рязанского государственного радиотехнического университета. 2022. № 82. С. 60-72. DOI: 10.21667/1995-4565-2022-82-60-72. URL: https://vestnik.rsreu.ru/images/archive/2022/82/2.2_.pdf (дата обращения: 12.02.2026).

10. Перепелкин Д. А., Ткачев Д. Д. Разработка облачной платформы и визуальной программной системы конфигурирования устройств Интернета вещей. Вестник Рязанского государственного радиотехнического университета. 2022. № 82. С. 73–89. DOI: 10.21667/1995-4565-2022-82-73-88. <https://vestnik.rsreu.ru/archives/2022/vypusk-82/1298-razrabotka-oblachnoy-platformy-i-vizualnoy-programmnoy-sistemy-konfigurirovaniya-ustrojstv-interneta-veshchej> (дата обращения: 12.02.2026).

11. Благирев М. М., Костыренков А. О. Анализ методов принятия решений в распределённых IoT-системах на основе мультиагентного подхода. Вестник Рязанского государственного радиотехнического университета. 2025. № 92. С. 146-159. DOI: 10.21667/1995-4565-2025-92-146-159. URL: <https://vestnik.rsreu.ru/images/archive/2025/92/4.4-.docx.pdf> (дата обращения: 29.01.2026).

УДК 004.896; ГРНТИ 20.23.17

ПРИМЕНЕНИЕ АЛГОРИТМА МУРАВЬИНОЙ КОЛОНИИ В СИСТЕМАХ ТЕХНИЧЕСКОГО ДОКУМЕНТИРОВАНИЯ

К.В. Бебко, Д.А. Осокин, А.В. Осокин, Б.В. Костров

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, torn_78@mail.ru*

Аннотация. В работе рассматриваются возможности применения видоизмененного оптимизационного алгоритма муравьиной колонии для систем технического документирования. Использование новых принципов обработки технической документации позволит значительно сократить ее перевод из бумажного в электронный вид, а также последующую ее поддержку в актуальном виде. Приводится пример программного кода построенного на основе алгоритма муравья-самурая. Показывается его преимущества по сравнению с другими подобными решениями.

Ключевые слова: оцифровка технической документации, PDM, алгоритм муравьиной колонии, САПР, эволюционные алгоритмы.

APPLICATION OF THE ANT COLONY ALGORITHM IN TECHNICAL DOCUMENTATION

K.V. Bebko, D.A. Osokin, A.V. Osokin, B.V. Kostrov

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, torn_78@mail.ru*

The summary. The paper discusses the possibility of using a modified optimization algorithm of the ant colony for technical documentation systems. The use of new principles for processing technical documentation will significantly reduce the time required to convert it from paper to electronic form, as well as to maintain it in an up-to-date state. The paper provides an example of a program code based on the samurai ant algorithm and demonstrates its advantages over other similar solutions.

paper considers the possibilities of using a modified ant colony optimization algorithm for *Key-words*: digitization of technical documentation, PDM, ant colony algorithm, CAD, evolutionary algorithms.

В настоящее время возможности эволюционных вычислений применяются при решении не только оптимизационных проблем, но и по представлению сложного поведения в виде результата взаимодействия относительно простых структур. При этом они являются одним из важных направлений исследования и совершенствования методов вычислительного интеллекта.

Следует заметить, что в работе современного инженера часто встречается ситуация, когда спецификации изделия существуют только в бумажном виде. Поэтому появляется необходимость в оцифровке и структурировании документов, для повышения эффективности проектирования и дальнейшего управления жизненным циклом изделия. Чтобы перенести документы в электронный вид сначала проводят их сканирование. Благодаря чему получают цифровые изображения спецификаций с выбором разрешения исходящего файла, который будет распознаваться [1].

Следующее действие оцифровки – оптическое распознавание текста (OCR). С помощью программ ABBYY FineReader, Tesseract OCR и других аналогичных преобразовывают отсканированные изображения в редактируемый текст.

Зачастую сканируемый документ имеет плохое качество изображения исходного текста, поэтому приходится использовать на практике ручной ввод текста.

Также, следует отметить, что и при распознавании документа перечисленные выше программы могут допускать ошибки, поэтому стоит проводить проверку корректности распознанных данных. Для этого можно использовать специализированные алгоритмы проверки, которые позволят сравнить исходный документ с конечным результатом.

После получения цифрового текста необходимо структурировать данные в соответствии со спецификацией изделия, благодаря чему можно повысить качество обработки необходимых документов [2]. Обычно структура включает:

1. обозначение изделия – уникальный идентификатор, который будет связан с нормативной базой;
2. наименование компонентов – описание элементов с учетом стандартов;
3. количество – информация о составе и необходимых ресурсах;
4. материалы – состав и необходимые ресурсы;
5. примечания – технические комментарии, ограничения и ссылки на дополнительную документацию.

Далее, для автоматизации обработки спецификаций и их дальнейшего внедрения, в процессах работы с технической документацией используются PDM-системы.

Стоит отметить, что современные инженерные задачи часто связаны с необходимостью поиска оптимальных решений в сложных многомерных пространствах, где классические методы оптимизации оказываются неэффективными. Целевые функции в этих задачах, как правило, обладают множеством локальных экстремумов, что делает невозможным применение детерминированных алгоритмов, быстро “застревающих” в неоптимальных решениях. Поэтому для решения прикладных производственных задач в таких условиях только одной сквозной интеграцией различных систем не обойтись. На помощь приходят популяционные алгоритмы – методы поиска, вдохновленные природными процессами и эволюционными стратегиями [3].

Основываясь на полученных экспериментальных данных [4, 5], одним из самых перспективных алгоритмов для решения нашей задачи является алгоритм муравьиной колонии. Муравьиный алгоритм основан на наблюдении за поведением муравьев в природе. При поиске пищи муравьи перемещаются случайным образом, оставляя за собой феромонный след [6]. Если путь оказывается удачным, по нему начинает ходить больше муравьев, и след ста-

новится сильнее. Со временем менее удачные маршруты забываются, так как феромон на них испаряется. В результате колония находит наиболее короткий и удобный путь.

Этот принцип был перенесен в вычислительные алгоритмы. В муравьином алгоритме несколько простых агентов независимо строят варианты решения задачи. Чем лучше полученное решение, тем больше оно влияет на дальнейший поиск. Таким образом, алгоритм постепенно концентрируется на наиболее удачных направлениях, но при этом не прекращает исследование других вариантов.

Классическая постановка муравьиного алгоритма традиционно рассматривается на примере задачи коммивояжера (TSP), в которой целью является минимизация суммарной длины маршрута. Формальное описание данной задачи и её математическая модель широко представлены в литературе и в дальнейшем используются как базовая схема роевого поиска [7].

В общем виде задача коммивояжера формулируется следующим образом (1):

$$p_{ij}^k(k) = \begin{cases} \frac{\tau_{ij}^\alpha \eta_{ij}^\beta(t)}{\sum_s \tau_{ij}^\alpha \eta_{ij}^\beta(t)} & \text{если } s, j \notin \text{tabu}_k \\ 0 & \text{иначе} \end{cases} \quad (1)$$

Математическая модель базовой системы муравьёв для задачи TSP выглядит следующим образом (2, 3):

Вероятность перехода:

$$p_{ij}^k(k) = \frac{\tau_{ij}^\alpha \eta_{ij}^\beta(t)}{\sum_s \tau_{ij}^\alpha \eta_{ij}^\beta(t)}, \text{ если город } j \text{ не посещён } (j \notin \text{tabu}_k), \quad (2)$$

$$p_{ij}^k(k) = 0, \text{ в противном случае}$$

Обновление феромона:

$$\tau_{ij}(t+n) = (1-\rho)(t) + \Delta\tau_{ij}, \text{ где } \Delta\tau_{ij} = \sum_{k=1}^m \Delta\tau_{ij}^k \quad (3)$$

где:

m – количество муравьёв в колонии;

$\tau_{ij}(t)$ – концентрация феромона на ребре между городами i и j в момент времени t ;

p_{ij}^k – вероятность перехода k -го муравья из города i в город j в момент t ;

α и β – управляющие параметры, определяющие относительную важность следа феромона (τ) и эвристической информации (η) при выборе пути;

$\eta_{ij}(t)$ – эвристическая желательность перехода из города i в город j (обычно обратно пропорциональна расстоянию между городами);

$\text{tabu}_k(k = 1, 2, \dots, m)$ – список уже посещённых k -м муравьём городов (список запретов), который динамически обновляется;

$\rho \in (0, 1]$ – коэффициент испарения феромона. Параметр $(1-\rho)$ представляет собой долю феромона, которая сохраняется;

$\Delta\tau_{ij}^k$ – откладываемое количество феромона.

Для наглядного представления принципа работы муравьиного алгоритма может быть использована упрощенная структурная схема, отражающая основные этапы его функционирования (рисунок 1).

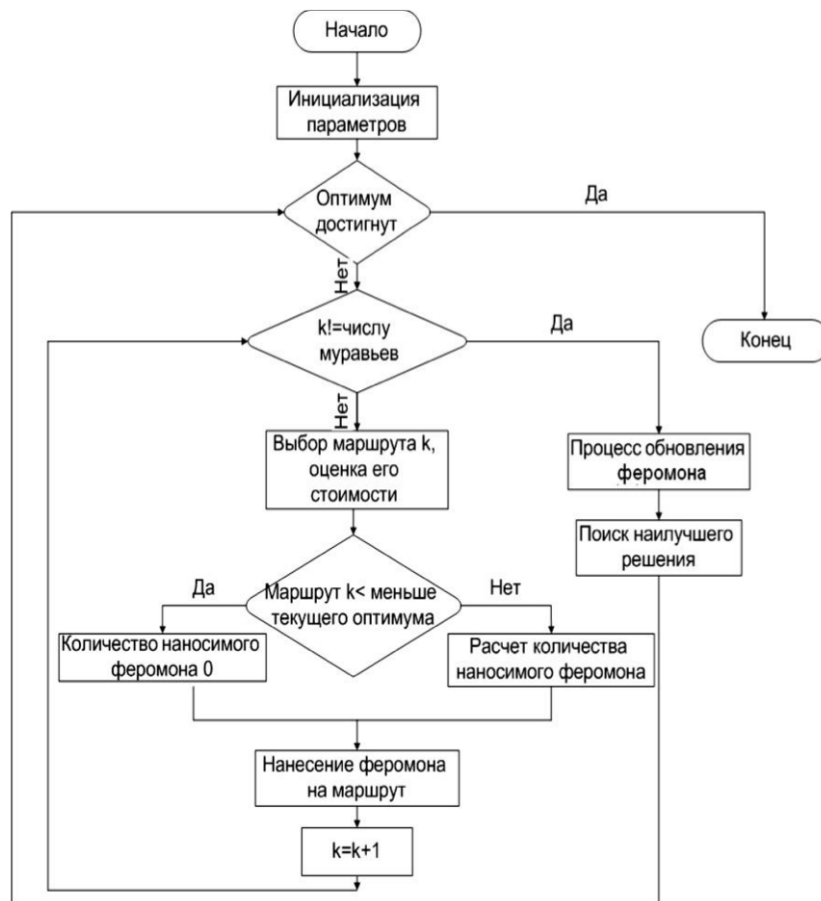


Рис. 1. Схема работы муравьиного алгоритма оптимизации

В задачах формирования структуры изделия, электронной спецификации или конструкторско-технологической документации оптимизация геометрического расстояния не является определяющим фактором. Основную роль играет сам факт допустимого и предпочтительного перехода между элементами структуры.

В связи с этим вводятся следующие допущения:

Эвристическая информация, связанная с расстоянием, исключается из модели (4):

$$\eta_{ij} = 1. \quad (4)$$

Вероятность перехода определяется только уровнем феромона, отражающим устойчивость и целесообразность связи между элементами (5):

$$P_{ij}^k(t) = \begin{cases} \frac{\tau_{ij}^\alpha(t)}{\sum_{i \in J_t^k} \tau_{ij}^\alpha(t) \eta_{ij}^\beta}, & \\ 0, & j \in J_i^k, \\ & j \notin J_i^k. \end{cases} \quad (5)$$

Приращение феромона, зависящее от длины маршрута $\Delta\tau_{ij}^k$ приравняется к нулю (6):

$$\Delta\tau_{ij}^k = 0. \quad (6)$$

Таким образом, феромон не откладывается как функция оптимальности длины, а используется исключительно как накопленная характеристика выбора пути.

Испарение феромона без учёта приращения

Обновление феромона в этом случае сводится только к процессу испарения (7):

$$\tau_{ij}(t+1) = (1 - \rho)\tau_{ij}(t), \quad (7)$$

где $\rho \in (0,1]$ – коэффициент испарения, определяющий скорость «забывания» редко используемых связей.

В результате описанных преобразований формируется упрощённая модель (рисунок 2), условно называемая моделью муравья-самурая, для которой:

- не важна оптимизация длины маршрута;
- феромон отражает значимость пути как логической или технологической связи;
- поведение муравья ориентировано на последовательное построение корректной структуры, а не на поиск кратчайшего пути.

Каждый маршрут муравья соответствует допустимой последовательности элементов изделия (узлов, деталей, документов), а накопленный феромон формирует матрицу связей (8):

$$T = \{\tau_{ij}\} \quad (8)$$

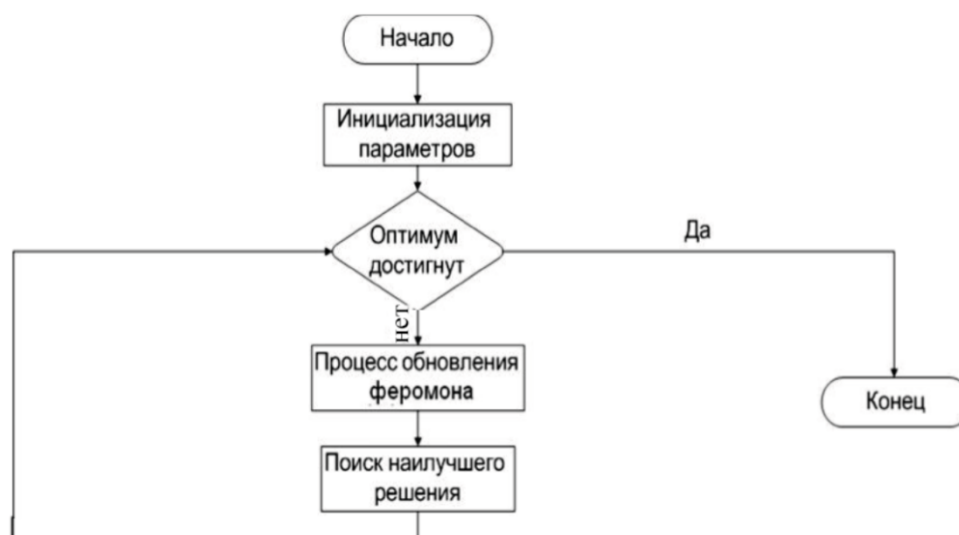


Рис. 2. Упрощенная схема работы муравьиного алгоритма оптимизации

В рамках исследования было разработано программное обеспечение, которое наглядно показывает концепцию муравья-самурая на основе алгоритма муравьиной колонии.

Первоначально программа была создана программа на исходном алгоритме по нахождению кратчайших путей при прохождении всех точек заданного маршрута.

Интерфейс программы представляет из себя:

1. Ввод исходных данных:

- автоматическая генерация координат 10 городов;
- расчет матрицы расстояний между городами.

2. Настройка параметров алгоритма:

- количество муравьев в колонии;
- коэффициент испарения феромонов;
- весовые коэффициенты влияния феромона и эвристики;
- максимальное число итераций.

3. Визуализация в реальном времени:

- динамическое отображение перемещения муравьев;
- цветовая индикация интенсивности феромонных следов;
- режим сравнения маршрутов всех муравьев;

- информационная панель с текущими показателями.
- 4. Отображение результатов:
 - визуализация оптимального маршрута;
 - длина и последовательность конечного пути.

Реализация программы с визуализацией работы муравьиного алгоритма, где программа выводит все пути муравьев + лучший путь представлена на рисунке 3.

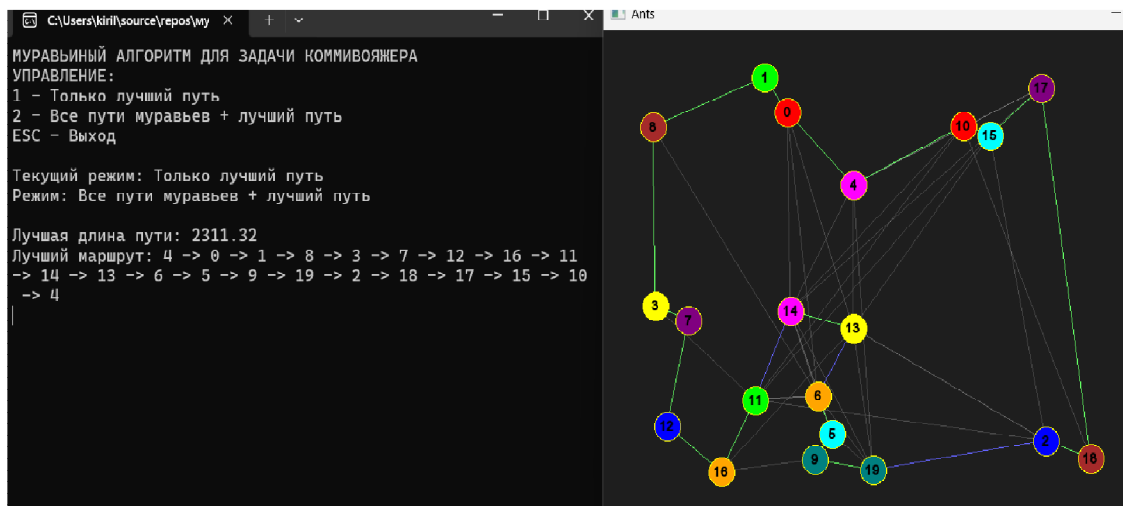


Рис. 3. Реализация все пути муравьев + лучший путь

Реализация программы с визуализацией работы муравьиного алгоритма, где программа выводит только лучший путь представлена на рисунке 4.

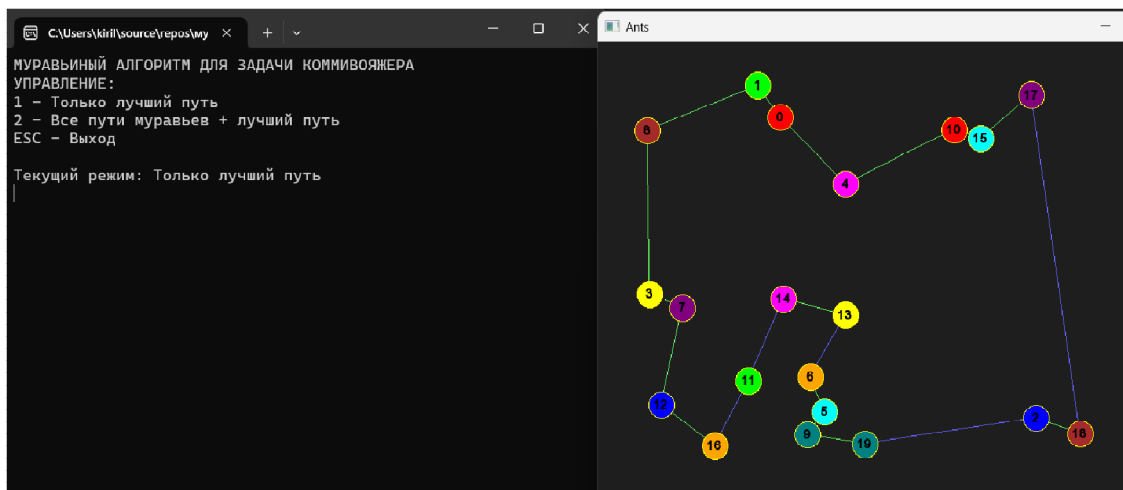


Рис. 4. Реализация только лучшего пути

Следующий этап эксперимента – корректировка программного обеспечения на основании представленного выше алгоритма, построенного на концепции муравья-самурая (рисунок 5). Важно заметить, что внесена возможность удаления муравья, в случае если маршрут ошибочен с целью исключения сбоев в построении структуры изделия в последующем. При этом само программное обеспечение возможно использовать в качестве встраиваемого модуля в существующих системах PLM.

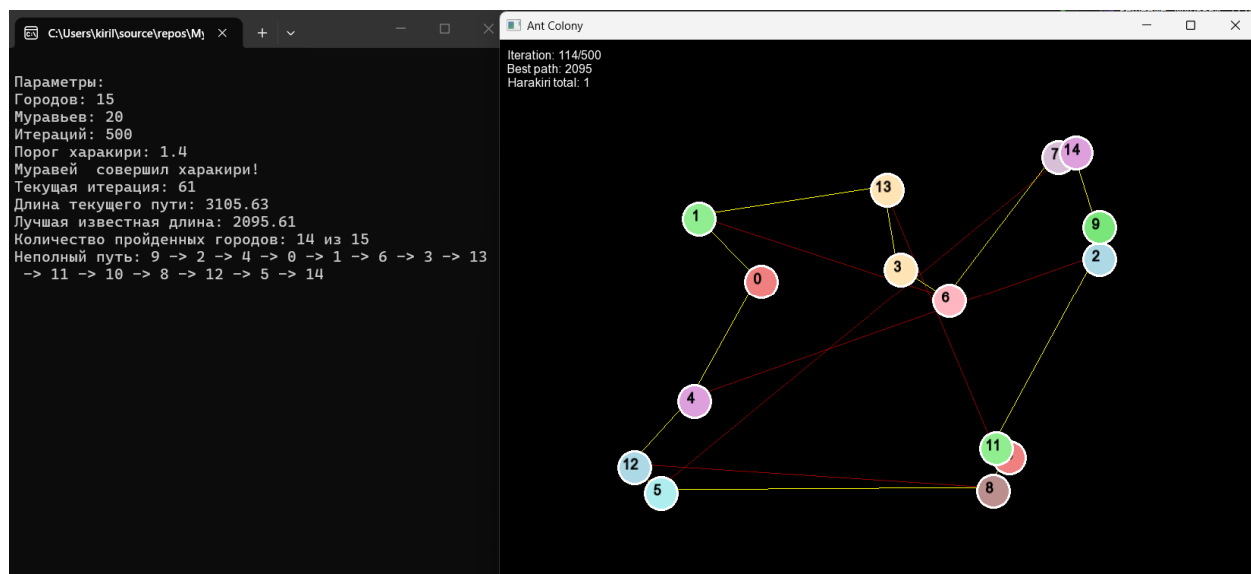


Рис. 5. Программная реализация алгоритма муравья-самурая

Полученные результаты доказывают, что модифицированный алгоритм муравьиной колонии, основанный на предложенной концепции муравья-самурая приспособлен к решению сложных задач проектирования. Он не требует точной математической модели и может работать с большими пространствами решений. Алгоритм устойчив к застреванию в локальных оптимумах и способен адаптироваться к изменению условий задачи, при этом не исключается возможность использовать его как самостоятельно, так и в сочетании с другими системами оптимизации.

Библиографический список

1. Карпухин, Д. С. Обработка спецификаций с бумажных носителей с целью построения электронной структуры изделия / Д. С. Карпухин, А. В. Осокин, И. С. Филиппин // Новые технологии в учебном процессе и производств : Материалы XXIII научно-технической конференции с международным участием, Рязань, 16–18 апреля 2025 года. – Рязань: Московский политехнический университет, 2025. – С. 324-327. – EDN PKIXNV.
2. Картышов К.В. Особенности функционального моделирования процесса автоматизации проектно-конструкторских работ РЭС // Научные труды. – 2020.
3. Корячко, В. П. Основные типы генетических алгоритмов / В. П. Корячко, А. Н. Сапрыкин, А. О. Сапрыкина // Информационные технологии в прикладных исследованиях : Межвузовский сборник научных трудов. – Рязань : ИП Коняхин А.В. (Book Jet), 2023. – С. 128-132. – EDN VTNKIU.
4. Грошев, Н. К. Применения алгоритма муравьиной колонии при Agile разработке продукта / Н. К. Грошев, А. В. Осокин, Б. В. Костров // Актуальные проблемы современной науки и производства : Сборник материалов X Всероссийской научно-технической конференции, посвященной 100-летию со дня рождения Заслуженного деятеля науки и техники Российской Федерации В.П. Миловзорова. В 2-х томах, Рязань, 25–28 ноября 2025 года. – Рязань: Рязанский государственный радиотехнический университет им. В.Ф. Уткина, 2025. – С. 95-101. – EDN EEKYNL.
5. Перепелкин, Д. А. Нейросетевая многопутевая маршрутизация в программно-конфигурируемых сетях на основе алгоритмов оптимизации муравьиной колонии / Д. А. Перепелкин, В. Т. Нгуен // Вестник Рязанского государственного радиотехнического университета. – 2024. – № 89. – С. 39-55. – DOI 10.21667/1995-4565-2024-89-39-55. – EDN YZNUAI.
6. Dorigo, M. Ant Colony Optimization / M. Dorigo, T. Stutzle. - Cambridge: MIT Press, 2004. - 305 p.
7. Bi, Sh. The Design and Analysis of TSP Problem Based on Genetic Algorithm and Ant Colony Algorithm / Sh. Bi, X. Dong, Ya. Ma // International Journal of Education and Management Engineering. – 2012. – Vol. 2, No. 9. – P. 56-60. – DOI 10.5815/ijeme.2012.09.09. – EDN LBEODS

УДК 004.932; ГРНТИ 28.23.15

СПОСОБ ЛОКАЛИЗАЦИИ КОДОВ AZTEC В РЕАЛЬНОМ ВРЕМЕНИ**К.П. Лапин, Е.Р. Муратов***Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, vebstersvfsay@gmail.com*

Аннотация. В работе представлен метод локализации двумерных кодов Aztec в режиме реального времени, основанный на анализе их структурных признаков. Ключевым этапом является обнаружение центрального шаблона «мишени», что позволяет корректно определять границы кода даже при наличии сложного фона. Для выделения линейных элементов применяется алгоритм LSD, а для уточнения принадлежности найденной структуры используется анализ последовательностей длин серий. Предложенный подход характеризуется высокой вычислительной эффективностью и может быть реализован в различных встраиваемых системах.

Ключевые слова: Aztec, алгоритм LSD, bullseye, мишень, локализация, штрих-код.

METHOD FOR LOCALIZING AZTEC CODES IN REAL TIME**K.P. Lapin, E.R. Muratov***Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, vebstersvfsay@gmail.com*

Abstract. The article proposes a method for real-time localization of two-dimensional Aztec codes based on the analysis of their structural features. A key stage is the detection of the central «bullseye» pattern, which makes it possible to correctly determine the boundaries of the code even in the presence of a complex background. The LSD algorithm is used to extract linear elements, and the analysis of run-length sequences is applied to verify the identified structure. The proposed approach is characterized by high computational efficiency and can be implemented in various embedded systems.

Keywords: Aztec, LSD algorithm, bullseye, target pattern, localization, barcode.

Основной метод локализации Aztec-кодов основан на выявлении их ключевой структурной особенности – шаблона «мишени», или bullseye. Этот центральный элемент кода состоит из чередующихся черных и белых модулей, формирующих несколько слоев в виде концентрических квадратов [1]. Количество слоев в «мишени» зависит от объема кодируемых данных: для компактного кодирования используется 2 слоя, а для полноразмерного – 3 слоя (рисунок 1).



Рис. 1. Шаблоны поиска двух типов кодов Aztec

На основе визуального анализа структуры «мишени» можно выявить ключевые линейные элементы – грани концентрических квадратов. В сущности, это прямые линии, образующие стороны каждого слоя шаблона поиска. Использование данных элементов позво-

ляет определить геометрические характеристики кода, такие как положение в кадре, ориентацию и его размер.

В реальных условиях локализация Aztec-кода усложняется из-за присутствия шумов сцены, искажениями перспективы, а также разнообразия и неоднородности типов поверхностей, на которых может находиться код. Поэтому для выделения линейных элементов «мишени» целесообразно использовать высокопроизводительный алгоритм обнаружения линейных сегментов, обладающий устойчивостью к шумам и искажениям, включая различные параболические деформации.

Таким методом, пригодным для внедрения в системы реального времени, является LSD (Line Segment Detector) [2]. Основные этапы работы алгоритма включают:

1. Преобразование изображения в оттенки серого и масштабирование. Исходное изображение приводится к градациям серого и масштабируется для устранения артефактов, таких как эффект лестницы, что предотвращает разбиение одного линейного сегмента на несколько частей.

2. Вычисление градиента изображения. Для определения направления и скорости изменения яркости рассчитывается градиент изображения. Точки упорядочиваются по величине градиента, начиная с наиболее контрастных. Точки с градиентом ниже порога (например, однородные фоны или плавные градиенты) отбрасываются, так как они мало информативны для поиска границ.

3. Формирование областей поддержки линий. От точки с наибольшим градиентом создаются области поддержки, к которым добавляются соседние точки, если угол их градиента согласован с направлением линии в пределах допустимого отклонения.

4. Аппроксимация линий прямоугольниками. Каждая область поддержки описывается прямоугольником, центр которого определяется как центр масс области, а ориентация – по главной оси инерции. Размеры прямоугольника выбираются так, чтобы полностью покрыть область, сохраняя её геометрические свойства.

5. Валидация с использованием NFA (Number of False Alarms). Для каждого прямоугольника вычисляется значение NFA, оценивающее вероятность того, что данный сегмент образован случайным шумом, а не реальной линией.

6. Уточнение и фильтрация результатов. Для повышения точности обнаружения прямоугольники корректируются: уменьшается их размер или корректируется угол ориентации, что улучшает качество выделенных линейных сегментов.

Использование описанного алгоритма позволяет выделить все линейные элементы шаблона bullseye. Демонстрация работы алгоритма LSD представлена на рисунке 2.



Рис. 2. Результат работы алгоритма LSD на изображениях с кодами Aztec

После выделения всех линейных сегментов сцены необходимо выделить линии, которые в совокупности образуют квадратные структуры [3]. Для этого применяется метод фильтрации линейных сегментов, включающий последовательное выполнение следующих этапов [4]:

1. Проверка соседних линий. Конечные точки каждой линии анализируются для обнаружения близлежащих линий в пределах заданной области, радиус которой определяется исходным разрешением изображения.

2. Оценка соотношения длин. Длины соседних линий анализируются для проверки их соответствия сторонам прямоугольника, приближенного к квадрату, с учетом возможных искажений перспективы.

3. Анализ углов. Углы между соседними линиями проверяются на соответствие диапазону от 70° до 120° , что позволяет корректно выявлять квадратные структуры с учетом возможных искажений перспективы.

4. Объединение и пометка. Линии, удовлетворяющие всем условиям, объединяются для формирования прямоугольников (квадратов) и помечаются, чтобы исключить повторное использование на последующих этапах анализа.

Результат выполнения описанного алгоритма фильтрации линейных сегментов представлен на рисунке 3.



Рис. 3. Результат выполнения алгоритма обнаружения концентрических квадратов «мишени»

Для повышения точности локализации Aztec-кода применяется алгоритм идентификации, основанный на дополнительном анализе центрального элемента, который позволяет подтвердить соответствие обнаруженной структуры искомому объекту. Его работа основана на построении средних линий, проходящих через центр концентрических квадратов, найденных на предыдущем этапе локализации. Алгоритм реализуется следующим образом.

1. *Определение центра квадратов.* Для каждого найденного концентрического квадрата вычисляется центр как среднее координат верхнего левого и нижнего правого углов. Этот центр служит начальной точкой для построения средних линий.

2. *Расширение квадратов для построения средних линий.* Для захвата всех слоев «мишени» стороны прямоугольника расширяются в 3 раза относительно центра. Это позволяет линиям проходить через весь поисковый шаблон и фиксировать чередующиеся черные и белые модули кода.

3. *Считывание последовательности пикселей путем построения средней линии.* Для каждого концентрического квадрата строится собственная средняя линия, проходящая че-

рез его центр. Линия проводится из центра одновременно в обе стороны – к одной и к противоположной границе расширенного квадрата, как это представлено на рисунке 4. По этой линии фиксируется последовательность пикселей, классифицируемых как черные или белые в зависимости от выбранного порога яркости. Такой подход обеспечивает точное прохождение линии через центр «мишени» и снижает влияние искажений на формируемую последовательность пикселей.

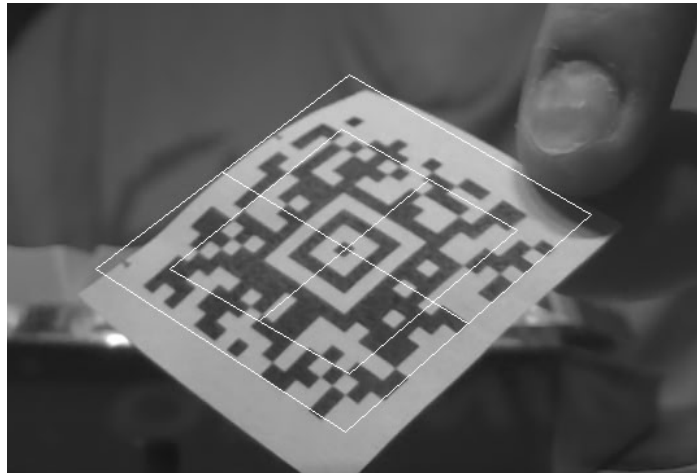


Рис. 4. Визуализация построения средних линий для всех расширенных концентрических квадратов

4. *Построение последовательности длин серий.* Последовательность черных и белых пикселей, полученная с помощью средней линии, анализируется для определения длин чередующихся модулей. Этот процесс позволяет преобразовать линейную последовательность бинаризованных значений пикселей в набор чисел, отражающих ширину каждого слоя «мишени». Полученные данные образуют так называемую последовательность *run-length* (длин серий), которая используется для дальнейшей проверки соответствия шаблону *bullseye*.

5. *Проверка соответствия шаблону bullseye.* Для каждой $run-length_j = \{r_1, r_2, \dots, r_n\}$ выполняется анализ семи соседних элементов $\{r_i, r_{i+1}, \dots, r_{i+6}\}$, соответствующих структуре компактного центрального шаблона (внешний – внутренний – промежуточный – центральный – промежуточный – внутренний – внешний). Проверка пропорциональности слоев осуществляется по следующему условию, которое означает, что толщина соседних пар слоев не должна отличаться более чем на 33%:

$$\frac{\max(r_i + r_{i+1})}{\min(r_i + r_{i+1})} \leq \frac{4}{3}, \quad (1)$$

где r_i, r_{i+1} – ширины двух соседних слоев из $run-length_j$ последовательности.

6. *Формирование окончательного результата.* После проверки всех сегментов последовательности пикселей помечаются те квадраты, которые удовлетворяют условиям «мишени». Эти данные используются для точной локализации кода и корректировки его ориентации в кадре, что позволяет минимизировать влияние смещений и искажений.

С целью повышения вычислительной эффективности анализ целесообразно начинать с квадратов наибольшего размера. В случае если наиболее крупный из обнаруженных концентрических квадратов удовлетворяет условиям «мишени», дальнейшая проверка вложенных квадратов может не выполняться, поскольку они заведомо принадлежат той же структуре центрального шаблона.

Следующим этапом локализации после идентификации структуры поискового шаблона является извлечение фрагмента изображения с кодом (рисунок 5). Для оценки границ Aztec-кода от центра «мишени» проводится обход пикселей вдоль нескольких лучей, исходящих из центра. На каждом луче фиксируется расстояние до последовательности достаточно длинных белых пикселей, что соответствует переходу от кода к фону. Собранные расстояния используются для вычисления медианного значения, на основании которого формируется ограничивающий прямоугольник или квадрат, включающий всю структуру кода.

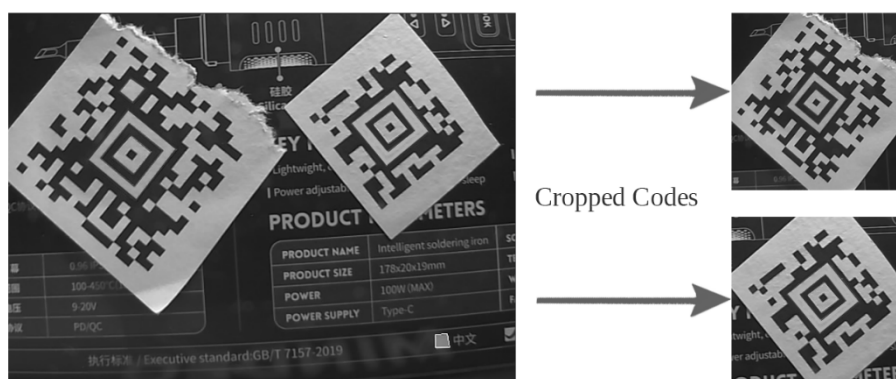


Рис. 5. Вырезанные фрагменты изображения, содержащие коды Aztec

После извлечения фрагмента изображения с целевым объектом необходимо скорректировать его ориентацию, чтобы структура кода была выровнена и обеспечивала максимально точное декодирование. Для этого определяется угол наклона μ нижнего линейного сегмента ранее помеченного квадрата по формуле:

$$\mu = \arctan \frac{y_2''' - y_1'''}{x_2''' - x_1'''} , \quad (2)$$

где x_1''' , y_1''' и x_2''' , y_2''' – координаты крайних точек нижней стороны прямоугольника.

В случае, если вычисленное значение угла наклона превышает допустимый предел в 5° , то выполняется поворот фрагмента на соответствующий угол (рисунок 6).

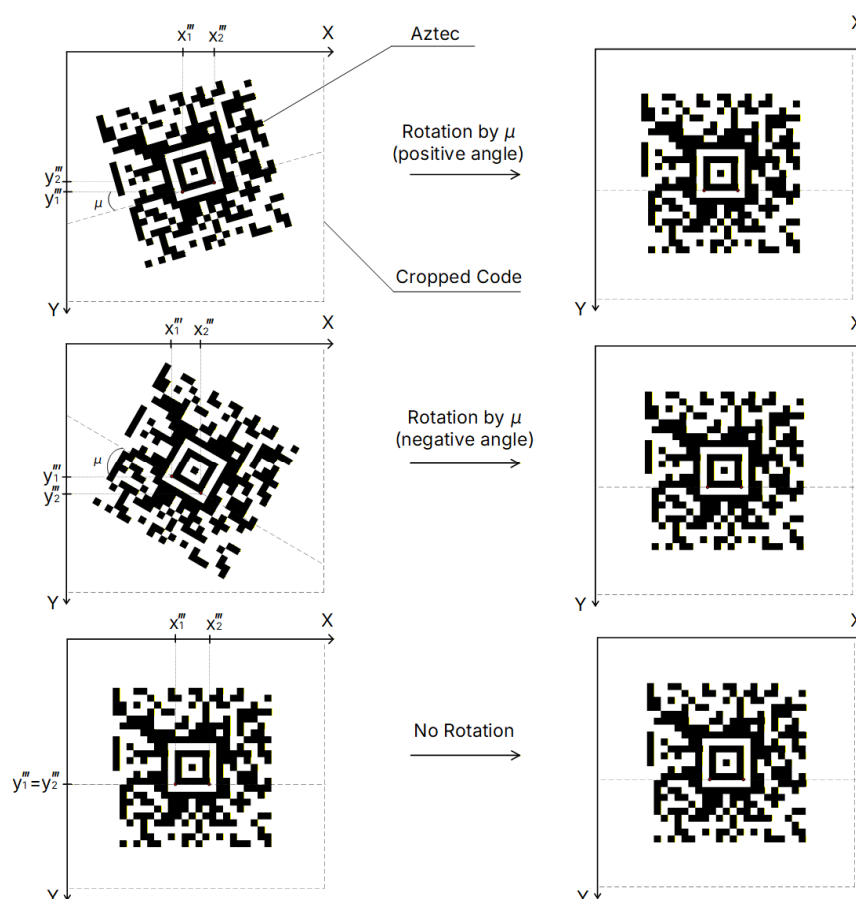


Рис. 6. Коррекция ориентации кода Aztec

После коррекции ориентации Aztec-кода полученный фрагмент изображения становится пригодным для обработки с использованием библиотек декодирования, что в большинстве случаев обеспечивает успешное считывание кода. Результат конечного этапа локализации кода представлен на рисунке 7.

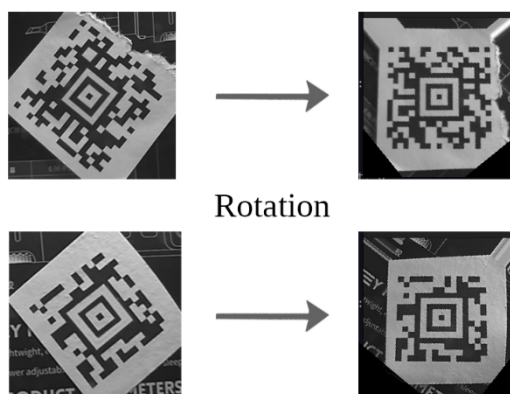


Рис. 7. Конечный результат локализации кодов Aztec

Библиографический список

1. ГОСТ Р ИСО/МЭК 24778-2010. Информационные технологии. Технологии автоматической идентификации и сбора данных. Спецификация символики штрихового кода Aztec Code = Automatic identification. Bar coding. Aztec Code symbology specification : национальный стандарт Российской Федерации : издание официальное : утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и

метрологии от 14 мая 2010 г. № 79-ст : введен впервые : дата введения 2011-07-01 / разработан обществом с ограниченной ответственностью «РИТ СЕРВИС». – Москва : Стандартинформ, 2010. – V, 45, [2] с. ; 29 см. – Текст: непосредственный.

2. LSD: a Line Segment Detector / R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, G. Randall // Image Processing On Line. – [Электронный ресурс]. – 2012. – Режим доступа: <https://www.ipol.im/pub/art/2012/gjmr-lsd>.

3. Муратов Е.Р., Ефимов А.И., Епифанов А.С. Алгоритм одновременной локализации множества кодов DataMatrix на изображении // Вестник РГРТУ. – 2024. – № 89. – С. 94-103.

4. Лапин, К. П. Способ локализации кодов datamatrix в реальном времени / К. П. Лапин, И. И. Сокол, Е. Р. Муратов // Современные технологии в науке и образовании - СТНО-2025 : Сборник трудов VIII Международного научно-технического форума. В 10-ти томах, Рязань, 04–06 марта 2025 года. – Рязань: Рязанский государственный радиотехнический университет им. В.Ф. Уткина, 2025. – С. 256-262. – EDN NPMKPL.

УДК 004.932; ГРНТИ 28.23.15

МЕТОД КОРРЕКЦИИ ИСКАЖЕНИЙ ОДНОМЕРНЫХ ШТРИХ-КОДОВ В РЕАЛЬНОМ ВРЕМЕНИ НА ОСНОВЕ АНАЛИЗА ГРАДИЕНТОВ

И.И. Сокол, Е.Р. Муратов

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, innocentbeing911@gmail.com*

Аннотация. В статье предложен метод коррекции искажений одномерных штрих-кодов, включающий глобальную коррекцию ориентации на основе взвешенной гистограммы направлений градиента, обнаружение конечных точек штриховых элементов методом направленной трассировки и локальную коррекцию остаточных искажений путем сдвига столбцов изображения в соответствии с аппроксимирующей прямой, построенной методом главных компонент. Метод обладает высоким быстродействием и предназначен для применения в промышленных системах реального времени.

Ключевые слова: одномерный штрих-код, коррекция искажений, градиент изображения, оператор Собеля, гистограмма направлений градиента, трассировка границ, метод главных компонент, аппроксимация прямой.

METHOD FOR CORRECTING DISTORTIONS OF ONE-DIMENSIONAL BARCODES IN REAL TIME BASED ON GRADIENT ANALYSIS

I.I. Sokol, E.R. Muratov

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, innocentbeing911@gmail.com*

Abstract. The article proposes a method for correcting distortions in one-dimensional barcodes, which includes global orientation correction based on a weighted histogram of gradient directions, detection of the endpoints of barcode elements using directional tracing and local correction of residual distortions by shifting image columns according to an approximating line constructed using the principal component analysis method. The method has high performance and is intended for use in industrial real-time systems.

Keywords: one-dimensional barcode, distortion correction, image gradient, Sobel operator, gradient direction histogram, edge tracing, principal component analysis, line approximation.

Основным подходом к коррекции искажений одномерных штрих-кодов является обнаружение их структурных элементов – собственно штрихов – при помощи анализа градиентов изображения и последующее выполнение геометрических преобразований в зависимости от положения и искажений данных элементов. Структура одномерного штрих-кода представляет собой последовательность чередующихся черных и белых полос различной ширины, ориентированных параллельно друг другу [1]. В идеальных условиях все штриховые элементы строго вертикальны, а их верхние и нижние границы лежат на горизонтальных параллельных прямых, что обеспечивает корректное декодирование путем анализа профиля яркости вдоль произвольной горизонтальной строки изображения, пересекающей область штрих-кода.

Поскольку штриховые элементы представляют собой прямые линии, естественным подходом к их обнаружению является применение методов детектирования прямых на изображениях. Классическое преобразование Хафа и алгоритм LSD (Line Segment Detector) являются наиболее распространенными из таких методов, однако оба универсальны и предназначены для обнаружения произвольных линейных сегментов [2]. При этом преобразование Хафа требует больших вычислительных затрат и не может быть использовано в промышленных системах. Алгоритм LSD обеспечивает более высокую производительность, однако также влечет за собой избыточные вычисления при решении узкой задачи обнаружения структурных элементов одномерного штрих-кода.

В предлагаемом методе вместо применения универсальных алгоритмов детектирования линий используется специализированный подход, непосредственно опирающийся на анализ свойств градиентного поля изображения штрих-кода. Входными данными для алгоритма коррекции служит прямоугольная область изображения, содержащая штрих-код, координаты которой определены на предшествующем этапе локализации (рисунок 1). Данная область вырезается из исходного изображения в оттенках серого и обрабатывается в качестве самостоятельного фрагмента.



Рис. 1. Вырезанная область изображения, содержащая одномерный штрих-код

Обработка каждого фрагмента начинается с вычисления градиента яркости. Как известно, градиент изображения характеризует скорость и направление изменения яркости в каждой точке и является основным инструментом для обнаружения границ объектов. В предлагаемом методе для вычисления компонент градиента применяется оператор Собеля размером 3×3 . Для каждого пикселя (x, y) внутренней области изображения (то есть при исключении граничных пикселей) горизонтальная G_x и вертикальная G_y компоненты градиента вычисляются путем свертки яркостной функции $I(x, y)$ с соответствующими ядрами:

$$G_x(x, y) = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * I(x, y), \quad (1)$$

$$G_y(x, y) = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * I(x, y), \quad (2)$$

где $I(x, y)$ – значение яркости пикселя с координатами (x, y) .

На их основе для каждого пикселя определяются магнитуа и направление градиента:

$$M(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2}, \quad (3)$$

$$\theta(x, y) = \arctan\left(\frac{G_y(x, y)}{G_x(x, y)}\right), \quad (4)$$

где $M(x, y)$ – магнитуа градиента в точке с координатами (x, y) ; $\theta(x, y)$ – направление градиента в точке с координатами (x, y) .

Магнитуа градиента, вычисленная для вырезанного фрагмента изображения со штрих-кодом, представлена на рисунке 2.



Рис. 2. Магнитуа градиента

Для определения преобладающей ориентации штриховых элементов строится взвешенная гистограмма направлений градиента. Диапазон углов $[-180^\circ; 180^\circ]$ разбивается на N равных интервалов (бинов), и каждый пиксель с магнитудой выше определенного порогового значения вносит равный его магнитуде вклад в соответствующий бин [3]. Пиксели с малой магнитудой отбрасываются, поскольку они соответствуют однородным областям, не содержащим границ штриховых элементов. Взвешивание магнитудой обеспечивает доминирование вклада пикселей с выраженными перепадами яркости, то есть пикселей, лежащих непосредственно на границах штрихов.

Направление градиента на границах штриховых элементов перпендикулярно самим штрихам, поэтому бин гистограммы с наибольшим накопленным значением соответствует направлению, ортогональному преобладающей ориентации штрихов (рисунок 3).

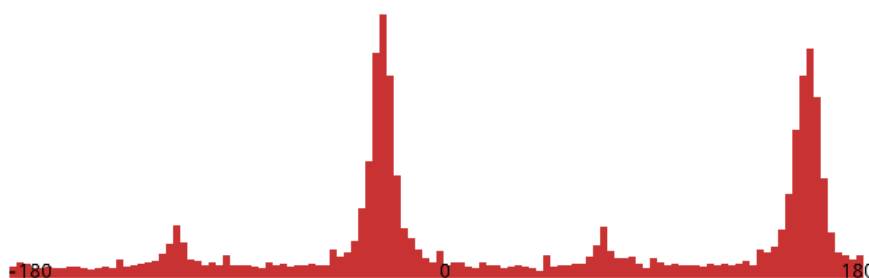


Рис. 3. Гистограмма направлений градиента

На основе направления доминирующего бина вычисляется угол поворота, необходимый для приведения штрих-кода к горизонтальной ориентации. Если вычисленный угол выходит за пределы диапазона $[-90^\circ; 90^\circ]$, он приводится к данному диапазону путем прибавления или вычитания 180° , что устраняет неоднозначность, связанную с противоположными направлениями градиента на левой и правой границах штриха. Затем выполняется поворот изображения на вычисленный угол (рисунок 4).



Рис. 4. Результат коррекции ориентации штрих-кода

Следующим этапом является обнаружение конечных точек штриховых элементов. Целью данного этапа является определение координат концов штриховых линий, совокупность которых описывает форму границы штрих-кода. В отсутствие искажений все концы лежат на одной горизонтальной прямой, а при наличии перспективных или иных искажений они образуют наклонную конфигурацию, анализ которой позволяет выполнить локальную коррекцию на последующем этапе.

Алгоритм обнаружения конечных точек включает два уровня сканирования: горизонтальное сканирование вдоль центральной строки изображения для обнаружения вертикальных границ штриховых элементов и вертикальную трассировку каждой обнаруженной границы вверх для определения ее конечной точки.

Горизонтальное сканирование начинается из центра изображения – точки (c_x, c_y) , где $c_x = \frac{W}{2}$, $c_y = \frac{H}{2}$, а W и H – ширина и высота изображения соответственно. Сканирование выполняется поочередно в двух направлениях: влево и вправо от точки (c_x, c_y) . Центральная строка выбрана в качестве начальной, поскольку она с наибольшей вероятностью пересекает все штриховые элементы.

В каждой позиции x вдоль строки $y = c_y$ проверяется наличие вертикальной границы штрихового элемента. Для этого в точке (x, c_y) анализируются компоненты градиента G_x и G_y , а также его магнитуда M . Точка классифицируется как принадлежащая вертикальной границе при одновременном выполнении двух условий:

$$M \geq M_t, \quad (5)$$

$$|G_x| \geq 1.2 * |G_y|, \quad (6)$$

где M_t – пороговое значение магнитуды градиента. Первое условие отсеивает точки с незначительными перепадами яркости. Второе условие обеспечивает выделение именно вертикальных границ, так как на границах штриховых элементов изменение яркости происходит преимущественно в горизонтальном направлении. Коэффициент 1.2 допускает незначительное отклонение границы от строго вертикального направления, что обеспечивает устойчивость к остаточным искажениям после этапа глобальной коррекции ориентации.

При обнаружении вертикальной границы в позиции (x, c_y) запускается вертикальная трассировка, заключающаяся в отслеживании данной границы от строки c_y в направлении убывания координаты y . На каждой строке y в горизонтальной окрестности $[x_b - r, x_b + r]$, где r – радиус поиска, x_b – горизонтальная координата границы на предыдущей строке, отыскивается пиксель с наибольшей магнитудой градиента, его координаты принимаются

как текущее положение границы, а счетчик последовательных пропусков обнуляется. В противном случае счетчик увеличивается на единицу, при этом горизонтальная координата поиска сохраняется от предыдущей строки. Трассировка прекращается при достижении верхнего края изображения либо при превышении счетчиком пропусков максимально допустимого количества последовательных строк без обнаружения границы m . Последнее условие соответствует ситуации, в которой на протяжении m последовательных строк не обнаружено ни одной точки, удовлетворяющей критериям вертикальной границы, что свидетельствует об окончании штрихового элемента. Координаты последней успешно обнаруженной точки фиксируются как конечная точка штриха.

Для исключения ложных обнаружений, обусловленных шумовыми откликами градиента, вводится ограничение на минимальную высоту трассировки. Конечная точка принимается только в том случае, если расстояние $c_y - y_{last} \geq h_{min}$, где y_{last} – вертикальная координата обнаруженной конечной точки; h_{min} – минимально допустимая высота штрихового элемента. После фиксации конечной точки горизонтальная позиция сканирования смещается на s пикселей в текущем направлении обхода от горизонтальной координаты конечной точки. Данный сдвиг предотвращает повторное обнаружение той же границы и обеспечивает переход к следующему штриховому элементу. Если же в текущей позиции x вертикальная граница не обнаружена, позиция сканирования сдвигается на один пиксель в текущем направлении.

Результатом выполнения указанной процедуры является множество точек, описывающих положения концов обнаруженных штриховых элементов (рисунок 5).

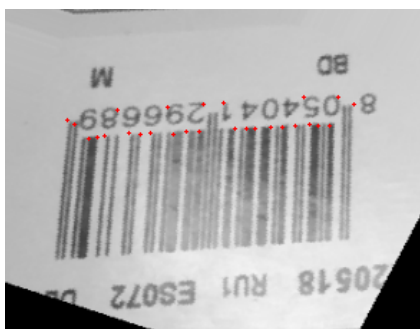


Рис. 5. Обнаруженные конечные точки штриховых элементов

По полученному множеству конечных точек выполняется аппроксимация прямой линией методом главных компонент (РСА). Для множества точек $\{(x_i, y_i)\}_{i=1}^n$, где n – количество обнаруженных конечных точек, вычисляются средние координаты

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (7)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad (8)$$

после чего формируется ковариационная матрица

$$C = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{bmatrix}, \quad (9)$$

где

$$S_{xx} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (10)$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad (11)$$

$$S_{yy} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2. \quad (12)$$

Направление аппроксимирующей прямой определяется собственным вектором ковариационной матрицы C , соответствующим ее наибольшему собственному значению [4]. Наибольшее собственное значение вычисляется по формуле:

$$\lambda_1 = \frac{S_{xx} + S_{yy}}{2} + \sqrt{\left(\frac{S_{xx} - S_{yy}}{2}\right)^2 + (S_{xy})^2}. \quad (13)$$

Соответствующий собственный вектор определяется как $(\lambda_1 - S_{yy}, S_{xy})$, нормируется к единичной длине и принимается в качестве направляющего вектора (v_x, v_y) . Аппроксимирующая прямая L задается точкой $(\bar{x}, \bar{y})$ и направляющим вектором (v_x, v_y) . На рисунке 6 представлен результат построения аппроксимирующей прямой для множества конечных точек.

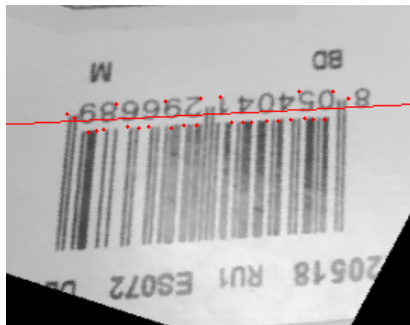


Рис. 6. Аппроксимирующая прямая для множества конечных точек

Поскольку среди обнаруженных конечных точек могут присутствовать выбросы, обусловленные градиентными откликами на элементах, не являющихся штрихами, выполняется процедура фильтрации найденных точек. Для этого каждая точка (x_i, y_i) классифицируется относительно прямой L по знаку косого произведения вектора $(x_i - \bar{x}, y_i - \bar{y})$ и направляющего вектора прямой (v_x, v_y)

$$d_i = (x_i - \bar{x}) * v_y - (y_i - \bar{y}) * v_x. \quad (14)$$

Знак величины d_i определяет, по какую сторону от прямой L расположена точка: точки с $d_i > 0$ относятся к одной стороне, точки с $d_i < 0$ – к другой. Из двух образовавшихся групп сохраняется только та, которая содержит большее количество точек, а меньшая группа исключается из рассмотрения. Данная процедура основана на предположении о том, что конечные точки штриховых элементов образуют компактную конфигурацию, расположенную преимущественно по одну сторону от аппроксимирующей прямой, тогда как выбросы, вызванные посторонними структурами, распределены менее плотно и оказываются

по противоположную сторону. На рисунке 7 представлен результат фильтрации конечных точек



Рис. 7. Результат фильтрации конечных точек

После исключения меньшей группы по оставшимся точкам повторно выполняется аппроксимация прямой методом главных компонент (рисунок 8). Повторная аппроксимация уточняет положение и наклон прямой L , поскольку исключение выбросов устраняет их влияние на ковариационную матрицу и, следовательно, на направление собственного вектора. Полученная уточненная прямая описывает границу штрих-кода и используется на следующем этапе для выполнения локальной коррекции искажений.

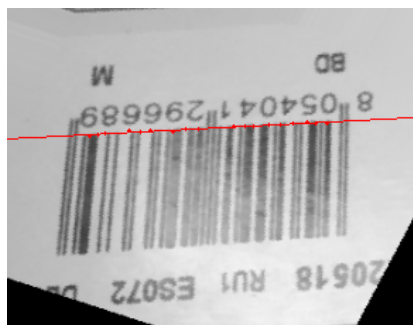


Рис. 8. Уточненная аппроксимирующая прямая после фильтрации конечных точек

Заключительным этапом коррекции искажений является выравнивание столбцов изображения на основе уточненной аппроксимирующей прямой L . Целью данного этапа является приведение границы штрих-кода, описываемой прямой L , к строго горизонтальному положению путем вертикального сдвига каждого столбца изображения на индивидуальную величину, зависящую от отклонения прямой L от горизонтали в данном столбце.

Для прямой L , заданной точкой $(\bar{x}, \bar{y})$ и направляющим вектором (v_x, v_y) , вычисляется наклон

$$k = \frac{v_y}{v_x}. \quad (15)$$

Для каждого столбца $x \in [0, W - 1]$ определяется вертикальная координата точки пересечения прямой L с данным столбцом

$$y_L(x) = \bar{y} + k * (x - \bar{x}). \quad (16)$$

Целевая горизонтальная координата y_i , к которой приводится прямая L , по умолчанию принимается равной $\lceil \bar{y} \rceil$. Величина вертикального сдвига столбца x составляет

$$\Delta y(x) = y_i - \lfloor y_L(x) \rfloor. \quad (17)$$

Каждый пиксель столбца x с координатой (x, y) переносится в позицию $(x, y + \Delta y(x))$. Пиксели, выходящие за пределы изображения после сдвига, отбрасываются, а незаполненные пиксели сохраняют значение фона, а именно белый цвет.

В результате выполнения данной процедуры прямая L , описывающая границу штрих-кода, приводится к горизонтальному положению, а вместе с ней выравниваются и штриховые элементы, концы которых соответствовали данной прямой. Результат выполнения описанного этапа представлен на рисунке 9.



Рис. 9. Результат выравнивания столбцов изображения: а – исходное изображение, б – результат коррекции

Таким образом, предложенный метод коррекции искажений одномерных штрих-кодов включает три последовательных этапа: глобальную коррекцию ориентации на основе взвешенной гистограммы направлений градиента, обнаружение конечных точек штриховых элементов методом направленной трассировки и локальную коррекцию остаточных искажений путем вертикального сдвига столбцов изображения в соответствии с аппроксимирующей прямой, построенной методом главных компонент. Совокупность перечисленных этапов обеспечивает приведение штрих-кода к стандартной горизонтальной ориентации с выровненными концами штриховых элементов, что является необходимым условием для корректного декодирования.

На рисунке 10 представлено сопоставление исходного изображения со штрих-кодом и конечного результата, полученного после последовательного выполнения всех трех этапов коррекции.



Рис. 10. Сопоставление исходного и скорректированного изображений штрих-кода: а – исходное изображение с искажениями, б – результат после выполнения всех описанных этапов коррекции

Приведенное сопоставление наглядно демонстрирует, что совокупность выполненных преобразований существенно уменьшает искажения исходного изображения, обеспечивая конфигурацию штриховых элементов, благоприятную для корректного декодирования распространенными программными средствами.

Существенным преимуществом предложенного метода является его высокое быстродействие. Это позволяет применять данный метод в промышленных системах реального времени, предъявляющих жесткие требования к времени обработки каждого кадра.

Библиографический список

1. ГОСТ Р 51294.3-99. Автоматическая идентификация. Кодирование штриховое. Термины и определения = Automatic identification. Bar coding. Terms and definitions : государственный стандарт Российской Федерации : издание официальное : принят и введен в действие Постановлением Госстандарта России от 20 августа 1999 г. № 267-ст : введен впервые : дата введения 2000-03-01 / разработан Ассоциацией автоматической идентификации ЮНИСКАН/EAN Россия/AIM Россия. – Москва : ИПК Издательство стандартов, 1999. – IV, 31, [1] с. ; 29 см. – Текст: непосредственный.
2. Лапин К.П., Сокол И.И., Муратов Е.Р. Способ локализации кодов DataMatrix в реальном времени / К.П. Лапин, И.И. Сокол, Е.Р. Муратов // Современные технологии в науке и образовании – СТНО-2025: Сборник трудов VIII Международного научно-технического форума. В 10-ти томах, Рязань, 04-06 марта 2025 года. – Рязань: Рязанский государственный радиотехнический университет им. В.Ф. Уткина, 2025. – С. 256-262.
3. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints / D.G. Lowe // International Journal of Computer Vision. – 2004. – Vol. 60. – P. 91-110. – DOI: 10.1023/B:VISI.0000029664.99615.94.
4. Jolliffe, I.T. Principal component analysis / I.T. Jolliffe. – 2nd ed. – New York: Springer, 2002. – 518 p. – ISBN 0-387-9544202.

УДК 004.65: 004.415.53; ГРНТИ 20.23.17

РАЗРАБОТКА ПРОТОКОЛА БЕНЧМАРКИНГА С УЧЕТОМ ПОКРЫТИЯ И РАБОЧЕЙ НАГРУЗКИ ДЛЯ СЦЕНАРНО-ОРИЕНТИРОВАННОГО СИНТЕЗА РЕЛЯЦИОННЫХ ТЕСТОВЫХ БАЗ ДАННЫХ

А.И. Баранчиков, М. А. М. Абди

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, maheesh@just.edu.so

Аннотация. В настоящей статье представлен протокол бенчмаркинга для сценарно-ориентированного синтеза реляционных тестовых баз данных. Протокол предназначен для генераторов реляционных данных, применяемых в задачах обеспечения качества программного обеспечения и нагрузочного тестирования, где данные должны удовлетворять явно заданным тестовым директивам, а не только воспроизводить свойства исходной базы данных. В работе предложен набор стандартизированных компонентов, обеспечивающих воспроизводимое сравнение: единица бенчмаркинга, каталог сценариев, набор рабочих нагрузок, политика выполнения, семейства метрик и пакет артефактов. Оценивание проводится по пяти взаимодополняющим измерениям: структурной валидности, сходству распределений, сохранению зависимостей, соответствию сценариям и полезности с точки зрения рабочей нагрузки. Кроме того, вводятся измеримые конструкции покрытия для семейств сценариев, классов ограничений, путей соединения и целевых граничных условий, что позволяет выполнять оценивание с учетом покрытия. Протокол также определяет контур оценивания, учитывающий аспекты оптимизации, для замкнутого цикла уточнения сценариев в условиях ограниченного бюджета.

Ключевые слова: протокол бенчмаркинга, тестирование баз данных, синтетические данные, сценарная генерация, управление тестовыми данными, оценка нагрузки.

THE DEVELOPMENT OF A COVERAGE AND WORKLOAD-AWARE BENCHMARKING PROTOCOL FOR SCENARIO-DRIVEN SYNTHESIS OF RELATIONAL TEST DATABASES

A. I. Baranchikov, M. A. M. Abdi

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, maheesh@just.edu.so

The summary. This article presents a benchmarking protocol for scenario-driven synthesis of relational test databases. The protocol is designed for relational data generators used in software quality assurance and performance testing, where data must satisfy explicit test directives rather than merely resemble a seed database. We proposed standardized components to enable reproducible comparison: a benchmark unit, a scenario catalog, a workload suite, an execution policy, metric families, and an artifact package. We evaluate five complementary dimensions: structural validity, distribution similarity, dependency preservation, scenario compliance, and workload utility. This introduces measurable coverage constructs for scenario families, constraint classes, join paths, and targeted edge conditions to enable coverage-aware assessment. The protocol also defines an optimization-aware track for closed-loop scenario refinement under budget constraints.

Keywords: benchmarking protocol, database testing, synthetic data, scenario generation, test data management (TDM), workload evaluation.

Введение

Реляционные тестовые базы данных играют важную роль в проверке корректности, устойчивости и производительности программных систем, ориентированных на работу с данными. В задачах обеспечения качества (QA) тестовые данные должны не только удовлетворять ограничениям схемы, но и обеспечивать целенаправленное воспроизведение стрессовых условий, контролируемое масштабирование, рост в режиме добавления, активацию граничных случаев и реалистичное поведение SQL-запросов в рамках заданных рабочих нагрузок. Классические подходы, включая вручную создаваемые тестовые наборы, случайную генерацию только по схеме и построение данных с учетом запросов, сохраняют практическую ценность в узких прикладных ситуациях, однако часто плохо масштабируются, недостаточно отражают межтабличные зависимости и обеспечивают ограниченную управляемость сценариями [1–4].

Последние достижения в области генерации синтетических реляционных данных позволили повысить структурную реалистичность и межтабличную согласованность за счет модульных, графовых и диффузионных моделей [5–8]. Одновременно современные исследования по бенчмаркингу уточнили подходы к оценке правдоподобия и полезности синтетических реляционных данных [9, 10]. Тем не менее потребности сценарно-ориентированного синтеза тестовых баз данных для задач обеспечения качества по-прежнему удовлетворяются лишь частично. В частности, практика бенчмаркинга остается недостаточно стандартизированной в отношении явной реализации сценариев, безопасного роста в режиме добавления, сохранения свойств рабочей нагрузки и воспроизводимости на уровне артефактов. В результате генератор может демонстрировать высокие показатели по метрикам, ориентированным на правдоподобие данных, и при этом оставаться недостаточно пригодным для целевых задач тестирования.

В сценарно-ориентированном тестировании инженерам необходимо задавать и проверять такие требования, как:

- (i) коэффициенты масштабирования для отдельных таблиц;
- (ii) инъекция пропусков и экстремальных значений;
- (iii) стресс-сценарии для глубоких цепочек внешних ключей (FK);
- (iv) рост данных в режиме *append* с сохранением целостности по первичным (PK) и внешним ключам (FK), а также реалистичного поведения соединений.

В нашей предыдущей исследовательской работе эти требования реализуются с помощью графово-диффузионного генератора реляционных данных (GDR-TDG), компилятора/контроллера сценариев и типизированного декларативного языка (ScenarioDSL), а на более позднем этапе – с помощью механизма замкнутого уточнения (CATS), предназначенного для настройки параметров с учетом тестового покрытия [11–13]. Однако, как и во многих практических конвейерах, бенчмаркинг остается фрагментированным, если не стандартизированы процедуры оценки, каталоги сценариев, наборы рабочих нагрузок и правила статистического представления результатов.

Цель данной статьи заключается в решении следующих вопросов: предложена формальная единица бенчмаркинга, совместно задающая схему, исходные данные, сценарии, набор рабочих нагрузок, спецификацию метрик и политику выполнения; разработана система метрик, включающая пять семейств: структурную валидность, сходство распределений, сохранение зависимостей, соответствие сценариям и полезность с точки зрения рабочей нагрузки; введено измеримое определение оценивания с учетом покрытия, основанное на покрытии семейств сценариев, классов ограничений, путей соединения и граничных условий; предложена стандартизированная процедура выполнения и представления результатов для синтеза в режиме замещения и в режиме добавления, включая проверки осуществимости, учет отказов и раскрытие артефактов; а определен контур оценивания, учитывающий аспекты оптимизации, для замкнутого цикла уточнения сценариев в условиях ограниченного бюджета.

Постановка задачи

Пусть реляционная схема представлена в виде графа с внешним ключом

$$G = (T, E), \quad (1)$$

где T – обозначает набор таблиц, а E – обозначает набор ребер с внешним ключом. Каждое ребро $(u, v) \in E$ указывает, что таблица u ссылается на таблицу v .

Исходная база данных над схемой G обозначается как D_0 . Пусть сценарная программа S будет скомпилирована в промежуточное представление (IR)

$$I = \langle n, O, R, \Delta_{pk}, \Delta_{fk} \rangle, \quad (2)$$

где n – вектор целевого количества строк в таблице;

O – набор нормализованных директив переопределения;

R – набор правил и бюджетов для крайних случаев;

Δ_{pk} – метаданные о смещении первичных ключей;

Δ_{fk} – метаданные о необходимом переназначении внешних ключей для синтеза в режиме добавления. При наличии исходных данных D_0 , генератора Gen и скомпилированных ограничений I синтез приводит к созданию синтетической базы данных

$$\hat{D} = Gen(G, D_0, I; \xi), \quad (3)$$

где ξ – случайное начальное значение.

Цель бенчмаркинга заключается не только в определении того, насколько $\hat{D}$ статистически соответствует исходным данным. Напротив, задача состоит в оценке пригодности $\hat{D}$ для тестирования. Это требует анализа по пяти измерениям: структурной валидности, сходства распределений, сохранения зависимостей, соответствия сценарным условиям и полезности с точки зрения рабочей нагрузки. Соответственно, результатом бенчмаркинга является не единственная скалярная оценка, а структурированный отчет по единицам бенчмаркинга, сценариям, рабочим нагрузкам и повторным запускам.

Протокол бенчмаркинга с учетом покрытия и рабочей нагрузки

Единица бенчмаркинга

Центральным объектом данного протокола является единица бенчмаркинга

$$U = \langle G, D_0, S, Q, M, \Pi \rangle, \quad (4)$$

где G – обозначает схему, D_0 – исходная или эталонная база данных, S – каталог сценариев, Q – набор рабочих нагрузок, M – спецификация метрик, а Π – политика выполнения. Политика выполнения включает версию СУБД, параметры конфигурации, допущения относительно индексации, политику кэширования, число повторений, зерна генерации, временные ограничения и описание аппаратной среды.

Исследование в рамках бенчмаркинга не может считаться воспроизводимым, если не представлены все шесть компонентов. Такая структура описания позволяет избежать пространенной проблемы при оценивании синтетических данных, а именно сравнения генераторов в условиях рабочих нагрузок или политик выполнения, заданных неполно либо недостаточно явно.

Спецификация каталога сценариев

Каждый сценарий $s \in S$ представляет собой типизированный шаблон с ограниченными параметрами, предикатами осуществимости и семантикой разрешения конфликтов. Протокол предусматривает четыре обязательных семейства сценариев:

- сценарии масштабирования – глобальное изменение числа строк или изменение числа строк в отдельных таблицах;
- сценарии граничных случаев – инъекция пропусков, усиление тяжелых хвостов распределения, размещение экстремальных значений, категориальный дисбаланс или стресс-тестирование цепочек внешних ключей;
- сценарии переопределения – детерминированное или вероятностное изменение выбранных атрибутов или строк;
- сценарии роста – генерация в режиме добавления с безопасными смещениями ключей и переназначением, сохраняющим ссылочную целостность.

Каждый шаблон сценария должен задавать области допустимых значений параметров, правила осуществимости, правила приоритета для конфликтующих директив, а также предварительное представление скомпилированной формы, необходимое для обеспечения аудируемости.

Неосуществимые сценарии не должны молча отбрасываться; они должны учитываться и отражаться в отчете как недопустимые кандидаты.

Проектирование набора рабочих нагрузок

Набор рабочих нагрузок Q должен включать группы запросов, отражающие потребности в обеспечении качества и нагрузочном тестировании: запросы OLTP-типа – точечные выборки, узкие фильтры, а также операции обновления и вставки, если они применимы; аналитические запросы – соединения, агрегатные операции и групповые фильтры; запросы для стресс-тестирования соединений – многотабличные соединения, чувствительные к перекосу распределений и кардинальности; сценарно-ориентированные запросы – SQL-запросы, специально предназначенные для активации конкретных инъецированных условий.

Каждый запрос $q \in Q$ должен быть документирован с указанием целевых таблиц и пути соединения, назначения запроса, ожидаемой чувствительности (например, к доле пропусков или селективности), а также конфигурации выполнения, включая индексацию, прогрев и политику повторных запусков.

Процедура выполнения

В этом разделе представляет протокол в виде явного псевдокода, а не неформального перечня действий.

Алгоритм 1. Выполнение, соответствующее протоколу, для одного запуска бенчмаркинга

Вход: единица бенчмаркинга $U = \langle G, D_0, S, Q, M, \Pi \rangle$, генератор Gen , случайное зерно ξ

Выход: синтетическая база данных $\hat{D}$ и пакет артефактов A_{pkg}

Инициализировать журнал выполнения $l \leftarrow \emptyset$

Для всех $s \in S$ выполнить

– скомпилировать s в промежуточное представление I_s

– вычислить предикаты осуществимости и применить правила разрешения конфликтов

– добавить предупреждения, отклонения и предварительные артефакты в l

Конец цикла

Сформировать скомпилированное представление сценариев $I = \{I_s : s \in S\}$

Синтезировать $\hat{D} = Gen(G, D_0, I; \xi)$ в режиме замещения или в режиме добавления

Проверить структурные ограничения на $\hat{D}$

Выполнить набор рабочих нагрузок Q на D_0 и $\hat{D}$ в соответствии с политикой Π

Вычислить все семейства метрик, заданные в M

Сформировать A_{pkg} на основе журналов, метаданных конфигурации и сводок по повторным запускам

Вернуть $(\hat{D}, A_{pkg})$

Такой порядок выбран намеренно: проверка осуществимости сценариев, генерация данных, структурная валидация и выполнение рабочих нагрузок разделены на аудируемые этапы, чтобы итоговые сводки не скрывали отказы, возникшие на более ранних стадиях конвейера.

Семейства метрик

Протокол использует пять взаимодополняющих семейств метрик. Прежде чем вводить отдельные уравнения, целесообразно разграничить их функции: структурная валидность определяет, является ли результат исполнимым; сходство распределений и сохранение зависимостей характеризуют реалистичность; соответствие сценариям проверяет, были ли реализованы запрошенные директивы; полезность с точки зрения рабочей нагрузки отражает поведение SQL-запросов на уровне задач тестирования.

Во-первых, структурная валидность измеряет соответствие схеме и целостность, необходимую для исполнения. Она включает долю нарушений внешних ключей

$$m_{fk} = \frac{\#FK \text{ нарушений}}{\#FK \text{ проверенных ссылок}}, \quad (5)$$

и долю коллизий первичных ключей

$$m_{pk} = \frac{\#PK \text{ коллизий}}{\#вставленных строк}, \quad (6)$$

Протокол также требует явного подтверждения наличия всех необходимых таблиц, столбцов, ключей и объявленных типов данных.

Во-вторых, сходство распределений количественно характеризует маргинальную реалистичность для репрезентативных атрибутов. Пусть A обозначает выбранное множество атрибутов, а $d_a \in [0,1]$ – нормированное расстояние для атрибута $a \in A$. Тогда

$$m_{dist} = \frac{1}{|A|} \sum_{a \in A} d_a. \quad (7)$$

Публикация результатов бенчмаркинга должна содержать как значения для отдельных атрибутов, так и агрегированные сводки. Использование только агрегированных показателей может скрывать локальные нарушения.

В-третьих, сохранение зависимостей измеряет, сохраняются ли многотабличные отношения, значимые для тестирования. Для множества J оцениваемых соединений пусть c_j и $\hat{c}_j$ обозначают кардинальности соединений в исходной и синтетической базах данных соответственно. Тогда

$$m_{join} = \frac{1}{|J|} \sum_{j \in J} \frac{|\hat{c}_j - c_j|}{\max(c_j, 1)}. \quad (8)$$

В зависимости от прикладного сценария это семейство метрик может также включать ошибку условных распределений вдоль цепочек внешних ключей и отклонение селективности для предикатов рабочей нагрузки.

В-четвертых, соответствие сценариям измеряет точность реализации запрошенных директив. Для каждого правила $r \in R$ с целевым значением t_r и реализованным значением $\hat{t}_r$ ошибка на уровне правила определяется как

$$e_r = \hat{t}_r - t_r. \quad (9)$$

Бенчмаркинг должен представлять допуски, специфичные для каждого класса, а также агрегированную среднеквадратичную ошибку

$$m_{scn} = \sqrt{\frac{1}{|R|} \sum_{r \in R} e_r^2}. \quad (10)$$

Интервалы допустимых отклонений должны задаваться отдельно для каждого класса сценариев; например, ошибки в числе строк, доле пропусков и параметрах перекося распределения не следует объединять без предварительной нормализации.

В-пятых, полезность с точки зрения рабочей нагрузки отражает, сохраняет ли синтетическая база данных поведение SQL-запросов, значимое для тестирования. Для запроса $q \in Q$ со средним временем выполнения на исходных данных L_q^{mean} , средним временем выполнения на синтетических данных $\hat{L}_q^{mean}$, p95 перцентилем времени выполнения на исходных данных L_q^{p95} , p95 перцентилем на синтетических данных $\hat{L}_q^{p95}$, селективностью на исходных данных S_q и селективностью на синтетических данных $\hat{S}_q$ протокол определяет

$$\rho_q^{mean} = \frac{\hat{L}_q^{mean}}{L_q^{mean}}, \quad (11)$$

$$\rho_q^{p95} = \frac{\hat{L}_q^{p95}}{L_q^{p95}}, \quad (12)$$

$$\delta_q^{sel} = \left| \hat{s}_q - s_q \right|, \quad (13)$$

Значения отношений, близкие к единице, и малые отклонения селективности указывают на более высокую реалистичность рабочей нагрузки. При необходимости также могут представляться доля превышений времени ожидания, частота изменения планов выполнения и рост дисперсии.

Формализация оценивания с учетом покрытия

В настоящей статье покрытие рассматривается как явная измеримая характеристика, включающая четыре компонента.

Первый компонент – покрытие семейств сценариев. Пусть F_{req} обозначает множество требуемых семейств сценариев, а F_{obs} – множество семейств, фактически представленных в единице бенчмаркинга. Тогда покрытие семейств сценариев определяется как

$$C_{fam} = \frac{|F_{obs} \cap F_{req}|}{|F_{req}|}, \quad (14)$$

Второй компонент – покрытие классов ограничений. Пусть K_{req} обозначает множество классов ограничений, считающихся значимыми для рассматриваемого приложения, например уникальность первичных ключей, корректность внешних ключей, поведение ограничений NOT NULL, границы доменов, временное упорядочение или согласованность составных ключей. Пусть K_{act} – подмножество этих классов, фактически задействованное набором сценариев и рабочей нагрузкой. Тогда

$$C_{con} = \frac{|K_{act}|}{|K_{req}|}, \quad (15)$$

Третий компонент – покрытие путей соединения. Пусть P_{target} обозначает заданное множество путей соединения в графе схемы, а P_{exec} – подмножество путей, активируемых набором рабочих нагрузок. Тогда

$$C_{join} = \frac{|P_{exec} \cap P_{target}|}{|P_{target}|}, \quad (16)$$

Эта величина важна, поскольку реалистичность рабочей нагрузки не может быть корректно оценена, если значимые структуры соединений фактически не задействуются.

Четвертый компонент – покрытие граничных условий. Пусть E_{target} обозначает целевые классы граничных условий, например всплески пропусков, выбросы, перекошенные ключи, скачки коэффициента ветвления или давление со стороны потенциальных «осиротевших» записей до этапа валидации, а E_{hit} – подмножество условий, которые были фактически активированы и проверены сценарно-ориентированными запросами. Тогда

$$C_{edge} = \frac{|E_{hit}|}{|E_{target}|}, \quad (17)$$

Наконец, последний компонент – агрегированная оценка покрытия. Протокол допускает использование взвешенной агрегированной сводной оценки покрытия

$$C(\theta) = \alpha_1 C_{fam} + \alpha_2 C_{con} - \alpha_3 C_{join} - \alpha_4 C_{edge}, \quad (18)$$

где θ обозначает вектор параметров сценария, $\alpha_i \geq 0$, а $\sum_i \alpha_i = 1$. Использование агрегированной оценки является опциональным, тогда как представление значений на уровне отдельных компонентов обязательно. В качестве нейтрального значения по умолчанию может использоваться $\alpha_i = 0.25$ для всех четырех компонентов. Если бенчмаркинг применяется в задачах обеспечения корректности, целесообразно назначать большие веса C_{con} и C_{edge} ; если же приоритетом является сохранение свойств рабочей нагрузки, обоснованным является увеличение весов C_{join} и C_{fam} . Во всех случаях вектор весов должен быть зафиксирован до начала экспериментов и сопровождаться по меньшей мере одним анализом чувствительности во избежание апостериорной подстройки.

Требования к представлению результатов и воспроизводимости

Публикация результатов бенчмаркинга, основанная на данном протоколе, не должна сводить оценивание к единственной скалярной метрике правдоподобия. Как минимум, статья или пакет артефактов должны содержать: описание схемы и исходных данных; шаблоны сценариев, области допустимых значений параметров и правила осуществимости; предварительные представления скомпилированных сценариев или их промежуточные представления; SQL-запросы рабочей нагрузки, политику параметризации и настройки выполнения; полную информацию о конфигурации СУБД и аппаратной среды; статистику по повторным запускам при различных зернах генерации; количество недопустимых сценариев, отказы генерации, нарушения структурной целостности и случаи превышения времени ожидания; скрипты вычисления метрик или их точные формулы; а сведения о версионировании артефактов и политике управления случайными состояниями.

Контур оценивания, учитывающий аспекты оптимизации

В реалистичных процессах обеспечения качества параметры сценариев нередко настраиваются итеративно с целью повышения ценности тестирования при одновременном сохранении реалистичности и структурной корректности. Это естественным образом приводит к задаче ограниченной, зашумленной оптимизации по принципу черного ящика в условиях ограниченного бюджета оценивания [14]. По этой причине протокол включает контур оценивания, учитывающий аспекты оптимизации.

Пусть θ обозначает вектор параметров сценария. Бенчмаркинг допускает оценивание с использованием скаляризованной целевой функции

$$J(\theta) = w_c C(\theta) + w_w W(\theta) - w_r R_{err}(\theta) - w_s S_{err}(\theta), \quad (19)$$

где $C(\theta)$ – полезность покрытия, определяемая согласно (18), $W(\theta)$ – полезность с точки зрения рабочей нагрузки, $R_{err}(\theta)$ – мера ошибки реалистичности, а $S_{err}(\theta)$ – мера ошибки соответствия сценарию. Протокол не зависит от конкретного метода оптимизации, однако требует представления следующих сведений: бюджет оценивания и бюджет реального времени; политика инициализации и политика выбора случайных зерен; доля отклоненных по критерию осуществимости конфигураций; траектория наилучшего на текущий момент значения целевой функции; показатели в фиксированных контрольных точках, например после

10, 25 и 50 вычислений; а сводки компромиссов в стиле Парето между покрытием, реалистичностью и полезностью с точки зрения рабочей нагрузки.

Данный контур расширяет бенчмаркинг от однократного сравнения генераторов до итеративной инженерии тестовых данных.

Заключение

В данной статье был представлен протокол бенчмаркинга, учитывающий покрытие и характеристики рабочей нагрузки, для сценарно-ориентированного синтеза реляционных тестовых баз данных. Протокол определяет единицы бенчмаркинга, каталоги сценариев, наборы рабочих нагрузок, семейства метрик, правила выполнения, требования к артефактам, а также отдельный контур оценивания, учитывающий аспекты оптимизации. Основной вклад работы состоит в смещении акцента с оценки, основанной исключительно на сравнении правдоподобия данных, к более широкой процедуре анализа, охватывающей структурную корректность, реализацию сценариев, покрытие и полезность для тестирования в условиях исполнимых рабочих нагрузок.

В ходе дальнейших исследований представляется целесообразным применить этот протокол к общедоступным единицам бенчмаркинга, расширить его применимость к публикации тестовых данных с учетом требований конфиденциальности и оценить стабильность кросс-СУБД в контексте политик контролируемого выполнения.

Библиографический список

1. Codd E. F. A Relational Model of Data for Large Shared Data Banks // Commun. ACM. 1970. Т. 13, № 6. С. 377–387.
2. Binnig C., Kossmann D., Lo E., Özsu M. T. QAGen: generating query-aware test databases // Proceedings of the 2007 ACM SIGMOD international conference on Management of data. Beijing China: ACM, 2007. С. 341–352.
3. Rigger M., Su Z. Finding bugs in database systems via query partitioning // Proc ACM Program Lang. 2020. Т. 4, № OOPSLA. С. 211:1–211:30.
4. Li Y., Zhang R., Yang X., Zhang Z., Zhou A. Touchstone: Generating Enormous {Query-Aware} Test Databases // 2018 USENIX Annual Technical Conference (USENIX ATC 18). 2018. С. 575–586.
5. Patki N., Wedge R., Veeramachaneni K. The synthetic data vault // 2016 IEEE international conference on data science and advanced analytics (DSAA). IEEE, 2016. С. 399–410.
6. Li J., Zhao Z., Abdollahzadeh M., Sikdar B., Tay Y. C. IRG: Modular Synthetic Relational Database Generation with Complex Relational Schemas. 2025.
7. Ketata M. A., Lüdke D., Schwinn L., Günnemann S. Joint Relational Database Generation via Graph-Conditional Diffusion Models: arXiv:2505.16527. arXiv, 2025.
8. Hudovernik V., Xu M., Shi J., Šubelj L., Ermon S., Štrumbelj E., Leskovec J. RelDiff: Relational Data Generative Modeling with Graph-Based Diffusion Models: arXiv:2506.00710. arXiv, 2025.
9. Hudovernik V., Jurković M., Štrumbelj E. Benchmarking the Fidelity and Utility of Synthetic Relational Data: arXiv:2410.03411. arXiv, 2024.
10. Jurkovic M., Hudovernik V., Štrumbelj E. SyntheRela: A Benchmark For Synthetic Relational Database Generation. 2025.
11. Баранчиков А.И., Абди М. А. М. Основанный на графической диффузии подход для создания синтетических реляционных тестовых баз данных // XXI век: итоги прошлого и проблемы настоящего плюс. XXI век: итоги прошлого и проблемы настоящего плюс, 2026.
12. Баранчиков А.И., Абди М. А. М. SCENARIODSL: Независимый от предметной области язык сценариев и компилятор для синтетических реляционных тестовых баз данных // XXI век: итоги прошлого и проблемы настоящего плюс. XXI век: итоги прошлого и проблемы настоящего плюс, 2026.
13. Баранчиков А. И., Абди М.а.м. Алгоритм Масштабирования С Учетом Сценариев И Применения Ограничений В Синтетических Реляционных Тестовых Базах Данных. Рязанский государственный радиотехнический университет им. В.Ф. Уткина, Рязань, 2025. С. 471–479.
14. Shahriari B., Swersky K., Wang Z., Adams R. P., De Freitas N. Taking the human out of the loop: A review of Bayesian optimization // Proc. IEEE. IEEE, 2015. Т. 104, № 1. С. 148–175.

УДК 004.942 ГРНТИ 27.39

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ОКОННЫХ ФУНКЦИЙ НА КАЧЕСТВО СПЕКТРОГРАММ

Н.А. Попов, Б.В. Костров

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, popovnikitaalexeevich@gmail.com*

Аннотация. В работе проводится исследование влияния различных оконных функций на качество спектрального анализа сигналов и построение спектрограмм. Рассматриваются основные характеристики оконных функций, включая ширину главного лепестка, уровень боковых лепестков, скорость их затухания и эквивалентную шумовую полосу. Выполняется моделирование в среде MATLAB для нескольких типов сигналов: гармонического сигнала, сигнала с нецелым числом периодов и сигнала, содержащего две близкие частоты на фоне шума. Полученные результаты демонстрируют влияние выбора оконной функции на спектральную утечку, частотное разрешение и устойчивость спектральной оценки. Сформулированы рекомендации по выбору оконных функций в задачах цифровой обработки сигналов.

Ключевые слова: спектральный анализ, оконные функции, быстрое преобразование Фурье, спектрограмма, спектральная утечка, метод Уэлча, цифровая обработка сигналов.

INVESTIGATION OF THE INFLUENCE OF WINDOW FUNCTIONS ON THE QUALITY OF SPECTROGRAMS

N.A. Popov, B.V. Kostrov

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, popovnikitaalexeevich@gmail.com*

The summary. The Article examines the influence of various window functions on the quality of spectral analysis of signals and the construction of spectrograms. The main characteristics of window functions are considered, including the width of the main lobe, the level of the side lobes, their attenuation rate and the equivalent noise band. MATLAB simulations are performed for several types of signals: a harmonic signal, a signal with a non-integer number of periods, and a signal containing two similar frequencies against a background of noise. The results obtained demonstrate the influence of the choice of the window function on spectral leakage, frequency resolution, and stability of spectral estimation. Recommendations on the choice of window functions in digital signal processing tasks are formulated.

Keywords: spectral analysis, window functions, fast Fourier transform, spectrogram, spectral leakage, Welch method, digital signal processing.

В условиях стремительного развития цифровых технологий и увеличения объемов данных особую актуальность приобретают задачи анализа и визуализации сигналов. Спектрограмма, как одно из наиболее информативных и наглядных представлений частотно-временных характеристик сигнала, является ключевым инструментом в таких областях, как аудиообработка, телекоммуникации, сейсмология, медицинская диагностика и анализ речевых данных. Однако качество спектрограммы напрямую зависит от методов обработки сигнала, среди которых важнейшую роль играет выбор оконной функции при построении кратковременного преобразования Фурье (STFT), которое подробно рассмотрено в фундаментальных трудах по цифровой обработке сигналов [1–3]. Методика STFT и её применение к анализу нестационарных сигналов подробно описаны в работе [4].

При работе в условиях смешанных аддитивных помех традиционные подходы к фильтрации часто оказываются недостаточно эффективными, что может приводить к потере полезной информации, особенно при анализе нестационарных сигналов со сложной частотной структурой. В связи с этим актуализируется задача разработки и оптимизации методов предварительной обработки, включая обоснованный выбор оконных функций, которые позволяют находить компромисс между разрешающей способностью по частоте и степенью подавления шумов. Выбор оконной функции существенно влияет на спектральную утечку, ширину главного лепестка и уровень боковых лепестков, что детально исследовано в классической работе Ф. Дж. Харриса [5].

В данной статье исследуется влияние различных оконных функций – Ханна, Хэмминга и Блэкмана – на качество формирования спектрограмм в условиях помех. Цель исследования – провести сравнительный анализ характеристик и определить оптимальные условия применения каждой из них для повышения точности и наглядности частотно-временного представления сигналов.

Основная часть

Оконная функция представляет собой весовую функцию, которая применяется к сигналу перед вычислением его спектра. С математической точки зрения оконное преобразование можно описать следующим выражением (1):

$$x_{w(n)} = x(n) \cdot w(n), \quad (1)$$

где $x(n)$ – исходный сигнал, $w(n)$ – оконная функция, а $x_{w(n)}$ – сигнал после применения окна. Применение оконной функции позволяет уменьшить резкие скачки сигнала на границах временного интервала [1, 5].

При отсутствии оконного преобразования сигнал фактически умножается на прямоугольное окно. Такое окно обладает узким главным лепестком спектра, однако характеризуется высоким уровнем боковых лепестков. В результате энергия сигнала распространяется на соседние частоты, что приводит к искажению спектральной картины [5, 7].

Использование оконных функций позволяет уменьшить уровень боковых лепестков и тем самым снизить спектральную утечку. Однако это сопровождается расширением главного лепестка, что приводит к уменьшению частотного разрешения. Таким образом, выбор оконной функции всегда связан с определённым компромиссом [2, 6].

Спектрограмма строится на основе кратковременного преобразования Фурье (STFT) (2), которое позволяет анализировать нестационарные сигналы путем разбиения на короткие сегменты и вычисления спектра для каждого сегмента:

$$STFT\{x(t)\}(\tau, f) = \int_{-\infty}^{\infty} x(t)w(t - \tau)e^{-j2\pi ft} dt, \quad (2)$$

где $x(t)$ – анализируемый сигнал, $w(t)$ – оконная функция, τ – временной сдвиг, f – частота.

Выбор оконной функции играет ключевую роль в качестве спектрограммы. Идеальное окно должно иметь узкий главный лепесток для хорошего частотного разрешения и низкий уровень боковых лепестков для минимизации утечки спектра.

Для сравнения оконных функций используются несколько ключевых параметров [1, 5, 6]:

1. Ширина главного лепестка определяет частотное разрешение системы. Чем меньше ширина главного лепестка, тем лучше система способна различать близкие частоты;
2. Уровень боковых лепестков характеризует степень спектральной утечки. Высокие боковые лепестки приводят к появлению ложных спектральных компонентов;
3. Скорость затухания боковых лепестков показывает, насколько быстро уменьшается амплитуда спектра окна при удалении от главного лепестка;
4. Эквивалентная шумовая полоса (ENBW) определяет влияние окна на уровень шумового фона при анализе случайных сигналов.

Сравнительный анализ оконных функций

Окно Ханна:

$$w(n) = 0.5 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right), \quad (3)$$

где N – ширина окна.

Окно Ханна обеспечивает хороший баланс между шириной главного лепестка и уровнем боковых лепестков. Ширина по уровню -3 дБ составляет примерно $1.5f_s/N$, первый боковой лепесток находится на 31 дБ ниже главного [9].

Окно Хэмминга:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right). \quad (4)$$

Уровень первого бокового лепестка ниже на 42 дБ, но затухание последующих лепестков медленнее. Частотное разрешение сравнимо с окном Ханна.

Окно Блэкмана (5)

$$w(n) = 0.42 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right) + 0.8 \cos\left(\frac{4\pi n}{N-1}\right). \quad (5)$$

Обеспечивает лучшее подавление боковых лепестков (первый лепесток на 58 дБ ниже), но главный лепесток шире ($1.75f_s/N$), что ухудшает частотное разрешение.

Метод спектральной оценки

Для оценки спектральной плотности мощности в работе используется метод Уэлча. Этот метод является модификацией классической периодограммы и позволяет уменьшить дисперсию спектральной оценки.

Алгоритм Уэлча включает следующие этапы: разбиение сигнала на сегменты, умножение каждого сегмента на оконную функцию, вычисление БПФ и усреднение полученных спектров.

Использование перекрывающихся сегментов и усреднения позволяет получить более устойчивую спектральную оценку, особенно при анализе шумовых сигналов.

Экспериментальное исследование

Для исследования влияния оконных функций было проведено моделирование в среде MATLAB. В рамках эксперимента рассматривались два типа сигналов.

Первый тип – гармонический сигнал с нецелым числом периодов в пределах анализируемого интервала. В этом случае наблюдается выраженная спектральная утечка [7]. На рисунке 1 представлены результаты спектрального анализа этого сигнала с использованием различных окон.

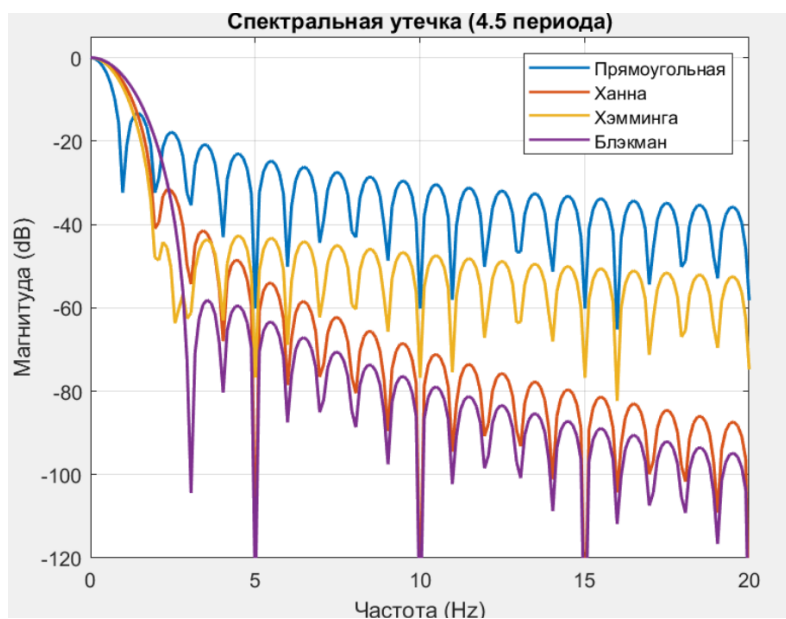


Рис. 1. Влияние оконных функций на спектральную утечку гармонического сигнала с нецелым числом периодов

Анализ графиков на рисунке 1 показывает, что прямоугольное окно (верхний график) дает самый узкий пик, но вокруг него наблюдается множество высоких боковых лепестков, которые можно ошибочно принять за реальные сигналы. Окно Ханна (средний график) эффективно подавляет боковые лепестки, незначительно расширяя основной пик. Окно Блэкмана (нижний график) практически полностью убирает боковые лепестки, однако главный пик становится заметно шире, что свидетельствует о потере частотного разрешения [9].

Второй тип сигнала содержит две близкие гармоники (50 и 55 Гц) и шумовую составляющую (рисунок 2). Данный эксперимент позволяет оценить способность различных окон различать близкие частоты в условиях помех [6].



Рис. 2. Сигнал с двумя близкими гармониками

Результаты эксперимента показали, что прямоугольное окно обеспечивает наилучшее частотное разрешение, позволяя четко разделить два пика, однако высокий уровень шума и боковых лепестков делает спектр "грязным". Окна Ханна и Хэмминга обеспечивают хороший компромисс: два пика различимы, а уровень шума между ними существенно подавлен. Окно Блэкмана, эффективно подавляя шум, "сливает" два близких пика в один из-за широкого главного лепестка, что делает его непригодным для задач высокого разрешения.

Заключение

В ходе работы было проведено исследование влияния оконных функций на качество спектрального анализа. Изучены теоретические материалы и ключевые параметры окон (ширина главного лепестка, уровень боковых лепестков). Реализовано моделирование в среде MATLAB, которое наглядно продемонстрировало теоретические выводы.

Практическая значимость исследования заключается в разработке рекомендаций по выбору оконной функции в зависимости от приоритетов обработки:

- для задач, требующих высокого разрешения близких частот, предпочтительнее использовать окно Ханна, как обеспечивающее наилучший баланс характеристик;
- в случаях, когда критично максимальное подавление шумов и помех от удаленных источников, а высокое разрешение не требуется, оптимальным выбором является окно Блэкмана;

— окно Хэмминга может быть полезно, когда необходимо сильное подавление ближайших боковых лепестков, но можно мириться с худшим затуханием дальних.

Предложенный подход может быть успешно внедрён в системы обработки сигналов в реальном времени, а также в научные и инженерные приложения, где важна точность и наглядность спектральной визуализации.

Перспективой дальнейших исследований может стать разработка адаптивных алгоритмов выбора оконной функции в зависимости от динамических характеристик сигнала, а также использование параметрических окон (Кайзера, Дольфа-Чебышева), позволяющих гибко управлять компромиссом между разрешением и подавлением [5, 8].

Библиографический список

1. Оппенгейм А. В., Шафер Р. В. Цифровая обработка сигналов. — М.: Техносфера, 2006. — 856 с.
2. Проакис Дж. Г., Манолакис Д. Г. Цифровая обработка сигналов: принципы, алгоритмы и применение. Вильямс, 2007. — 1072 с.
3. Оппенгейм А. В., Шафер Р. В., Бак Дж. Р.. Дискретные сигналы и системы. — М.: Техносфера, 2012. — 640 с.
4. Allen J. B., Rabiner L. R. A unified approach to short-time Fourier analysis and synthesis // Bell System Technical Journal. — 1977. — Vol. 56, № 8. — P. 1555–1577.
5. Фредерик Дж. Харрис. On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform // Proceedings of the IEEE. — 1978. — Vol. 66, № 1. — P. 51–83.
6. Бабичев С. А., Литвинов В. Н. Анализ влияния оконных функций на точность спектральных оценок Цифровая обработка сигналов. — 2015. — № 3. — С. 12–18.
7. Коваленко А. Н., Миронов Д. А. Исследование спектральной утечки при использовании различных оконных функций // Радиотехника. — 2017. — № 9. — С. 45–49.
8. Сафонов В. И., Егоров А. С. Применение оконных функций в задачах анализа нестационарных сигналов // Информационно-измерительные и управляющие системы. — 2018. — Т. 16, № 4. — С. 27–32.
9. Григорьев И. В., Петрова М. Ю. Сравнительный анализ методов построения спектрограмм в условиях шумовых помех // Вестник МГТУ им. Н. Э. Баумана. Сер. Приборостроение. — 2020. — № 2. — С. 56–64.

УДК 004.932 + 004.85; ГРНТИ 50.01

СРАВНЕНИЕ ТРАДИЦИОННЫХ МЕТОДОВ ОБРАБОТКИ ИЗОБРАЖЕНИЙ И АЛГОРИТМОВ ИИ

Е.П. Кузьмина, А.А. Вьюгина

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, vuyuginaa@mail.ru

Аннотация. В работе проводится сравнительный анализ традиционных методов обработки изображений и подходов на основе искусственного интеллекта. Рассматриваются основные области применения цифровой обработки изображений, такие как дистанционное зондирование, аэрофотосъемка с БПЛА и медицинская диагностика, что обуславливает актуальность поиска оптимальных методов анализа растущего объема визуальных данных. Исследование фокусируется на трех ключевых критериях: точность, скорость работы и сложность разработки. Подробно рассмотрены особенности внедрения нейросетевых методов в системах реального времени с учетом ресурсных ограничений и требований детерминированности. Дана сравнительная характеристика классических алгоритмов, основанных на математических преобразованиях, и методов глубокого обучения. Обосновываются условия, при которых замена классических подходов нейросетевыми является оправданной.

Ключевые слова: цифровая обработка изображений, глубокое обучение, нейронные сети, классические алгоритмы обработки изображений, беспилотные летательные аппараты (БПЛА), системы реального времени, сравнительный анализ.

COMPARISON OF TRADITIONAL IMAGE PROCESSING METHODS AND AI ALGORITHMS

E.P. Kuzminova, A.A. Vyugina

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, vyuginaa@mail.ru

The summary. This paper presents a comparative analysis of traditional image processing methods and approaches based on artificial intelligence. The main application areas of digital image processing are considered, such as remote sensing, UAV aerial photography, and medical diagnostics, which determines the relevance of searching for optimal methods for analyzing the growing volume of visual data. The study focuses on three key criteria: accuracy, speed, and development complexity. The specifics of implementing neural network methods in real-time systems are examined in detail, taking into account resource constraints and determinism requirements. A comparative characterization of classical algorithms based on mathematical transformations and deep learning methods is provided. The conditions under which replacing classical approaches with neural network ones is justified are substantiated.

Keywords: digital image processing, deep learning, neural networks, classical image processing algorithms, unmanned aerial vehicles (UAVs), real-time systems, comparative analysis.

Традиционные (классические) методы обработки изображений

В классическом понимании обработка изображений представляет собой совокупность математических и логических операций, применяемых к матрице пикселей. Изображение в цифровой форме описывается функцией двух переменных $f(x, y)$, а его обработка сводится к преобразованиям этой функции с использованием методов линейной алгебры, математического анализа и теории вероятностей [1]. Подробное изложение классических подходов представлено также в учебнике Шапиро и Стокмана [2].

Существует несколько базовых операций при работе с изображениями, одной из наиболее распространённых является фильтрация. При пространственной обработке изображение сворачивают с ядром (маской) заранее выбранного размера. Для подавления высокочастотных шумовых составляющих нередко задействуют фильтр Гаусса, где весовые коэффициенты убывают по нормальному закону. Для выделения границ востребованы операторы Собеля и Кэнни. Первый оценивает градиент яркости при помощи пары масок (по горизонтали и вертикали) и строит карту градиентов, а второй выполняет более сложную цепочку действий: сначала размывает картинку по Гауссу, затем вычисляет направления перепадов, подавляет не-максимумы и завершает работу двойной пороговой отсечкой, чтобы получить аккуратные и непрерывные линии контуров.

Другое важное направление – морфологическая обработка, оперирующая формами и структурами на бинарных или полутоновых изображениях. Базовые операции – эрозия и дилатация. Эрозия «сжимает» области объекта, удаляя мелкие выступы и изолированные шумовые точки, а дилатация, наоборот, расширяет границы, заполняя разрывы. Последовательное применение этих операций в виде открытия и закрытия позволяет эффективно очищать изображения от шума и выделять связные области.

Сегментация, то есть разделение изображения на смысловые фрагменты, долгое время выполнялась с помощью пороговых методов. Простейший из них – бинаризация по глобальному порогу, когда все пиксели со значением выше заданного уровня относят к объекту, а остальные – к фону. Более гибкими являются адаптивные пороговые алгоритмы, учитывающие локальные характеристики изображения.

Все эти приёмы давно и надёжно реализованы в открытых инструментах для компьютерного зрения, самым популярным из которых является библиотека OpenCV [6]. Классические алгоритмы встроены в неё в форме высокооптимизированных функций, благодаря чему их удаётся запускать в режиме реального времени даже на маломощных устройствах.

К достоинствам традиционного подхода можно отнести детерминированность: при одних и тех же входных данных и параметрах результат всегда будет одинаковым, что кри-

тически важно для систем, требующих предсказуемости. Кроме того, такие методы обычно не нуждаются в больших размеченных наборах данных и обучении; их работа прозрачна и допускает аналитическое обоснование. Вычислительная сложность заранее известна и часто позволяет достигать высокой производительности на центральных процессорах без специализированных ускорителей.

Классические методы обработки изображений имеют существенные ограничения и недостатки. Во-первых, их эффективность сильно зависит от ручного подбора параметров (размеров ядер, порогов, коэффициентов), который может быть трудоёмким и плохо обобщаться при изменении условий съёмки. Во-вторых, они слабо адаптируются к сложным текстурам, неоднородному освещению и вариативности формы объектов. Например, детектор Кэнни, как отмечается в современных обзорах, даёт качественные контуры лишь при правильно подобранных порогах, а при наличии шумов или размытых границ требует предварительной настройки, которая не всегда универсальна. Классические алгоритмы остаются инструментом, хорошо зарекомендовавшим себя в контролируемых условиях, но начинают уступать при работе с разнородными и зашумлёнными данными.

Методы на основе ИИ (глубокое обучение)

В отличие от классических алгоритмов, где каждый этап обработки жёстко задаётся разработчиком, методы на основе искусственного интеллекта строятся по иному принципу: модель самостоятельно выявляет значимые закономерности, обучаясь на размеченных данных. Как отмечается в обзоре Alzubaidi и соавторов [3], традиционные подходы опираются на признаки, сконструированные вручную, тогда как глубокое обучение позволяет извлекать признаки непосредственно из данных, выстраивая иерархические представления – от простых краёв до сложных семантических категорий.

Ключевой тип архитектур в этой области – свёрточные нейронные сети (CNN). В их основе лежат свёрточные слои, которые, подобно фильтрам Собеля или Гаусса. Но есть принципиальное различие: весовые коэффициенты этих окон не изобретаются человеком, а автоматически подгоняются в ходе обучения на огромном массиве примеров. Далее слои пулинга сокращают размерность данных, оставляя только самую существенную информацию. Среди зарекомендовавших себя архитектур можно назвать ResNet (где остаточные связи позволяют строить очень глубокие сети) [4], U-Net (де-факто стандарт для медицинской сегментации) и семейство YOLO – быстрые детекторы объектов, способные работать в реальном времени [5]. При анализе снимков с дронов и спутников всё чаще прибегают к гибридным схемам и фундаментальным моделям, хотя, по мнению ряда экспертов, в узких прикладных задачах небольшие сети, дообученные на целевых данных, зачастую оказываются надёжнее универсальных гигантов.

К достоинствам ИИ-подхода можно отнести высокую точность на сложных сценах – при вариативном освещении, частичных перекрытиях и текстурированных объектах – а также автоматическое извлечение признаков, которое освобождает разработчика от ручного подбора параметров. Кроме того, модели, обученные на разнообразных данных, демонстрируют устойчивость к шумам и искажениям, но у такого подхода есть и серьёзные недостатки. Главным из них является потребность в больших размеченных наборах данных, создание которых требует значительных временных и финансовых затрат. Глубокое обучение также отличается высокими вычислительными затратами как на этапе обучения, так и на этапе инференса, что особенно критично для оборудования с ограниченными ресурсами (например, бортовых систем БПЛА). К тому же поведение нейросетей менее предсказуемо по сравнению с детерминированными классическими алгоритмами, а риск переобучения остаётся серьёзной проблемой: модель может отлично работать на обучающей выборке, но терять качество на новых данных, если они выходят за рамки исходного распределения.

Сравнительный анализ классических методов и на основе ИИ

Для сравнения двух подходов обратимся к экспериментальным данным, полученным авторами в ходе исследования. Выбор YOLO в качестве нейросетевого инструментария объясняется его способностью быстро обрабатывать видеопоток [5]. В задаче бинарной классификации объектов на снимках, сделанных с борта беспилотника, был проведен синтез методов – применение классического алгоритма обработки изображений с детектором YOLO. Традиционная цепочка действий включала бинаризацию методом Оцу, морфологическое открытие для очистки от помех и последующий разбор геометрических параметров выделенных контуров (занимаемая площадь, степень компактности). В качестве второго подхода была задействована свёрточная сеть YOLOv5, предварительно натренированная на размеченной выборке кадров.

По результатам эксперимента, классический метод при стабильных условиях съёмки (однородный фон, постоянное освещение) обеспечивал точность около 82% и время обработки одного кадра 0,2–0,3 секунды на стандартном процессоре. Однако при изменении освещённости, появлении шумов или частичных перекрытиях точность классического алгоритма падала до 58–65%, причём каждая корректировка параметров (порогов, размера морфологического ядра) требовала повторного ручного подбора. Нейросетевой детектор, обученный на данных, включающих различные погодные условия и ракурсы, показал устойчивую точность 94–96% и работал со скоростью 8–12 миллисекунд на кадр при использовании встроенного GPU. Аналогичные результаты получены и для задачи сегментации: традиционные методы давали IoU в диапазоне 0,47–0,58, тогда как U-Net и Mask R-CNN достигали 0,85–0,90, хотя время обработки возрастало до 1,2 секунды.

Сложность разработки также оценивалась по нескольким параметрам. Классический алгоритм был написан за два дня, не требовал сбора данных, но последующая доводка под различные сценарии заняла ещё несколько недель. Нейросетевой подход, напротив, потребовал около двух недель на сбор и разметку датасета (около 5000 изображений) и ещё неделю на обучение и настройку гиперпараметров. При этом код самого детектора оказался компактнее (фактически вызов готовой архитектуры), однако полный цикл разработки оказался длиннее и ресурсоёмче.

Ниже приведена сводная таблица сравнения по трём заявленным критериям (таблица 1).

Таблица 1. Сравнительный анализ

Критерий	Традиционные методы	Методы на основе ИИ
Точность	Высокая в контролируемых условиях, но резко падает при изменении освещения, появлении шумов или нестандартных ракурсах	Стабильно высокая при обучении на разнообразных данных; устойчивы к вариативности условий
Скорость работы	Быстрая работа на CPU; время обработки предсказуемо и не зависит от сложности сцены	Зависит от аппаратуры: на CPU может уступать классике, на GPU часто превосходит; возможна работа в реальном времени
Сложность разработки	Не требует сбора данных; код реализуется быстро; основная трудоемкость – ручная настройка параметров под разные условия	Требует большого размеченного датасета и вычислительных ресурсов для обучения; код инференса прост; полный цикл разработки длиннее и дороже

Таким образом, выбор подхода определяется техническим заданием конкретного применения. Классические методы выигрывают в простоте и предсказуемости при стабильных условиях, тогда как ИИ-методы обеспечивают более высокую точность и адаптивность ценой больших затрат на данные и вычислительные ресурсы.

Проведённый сравнительный анализ показывает, что традиционные методы обработки изображений и подходы на основе искусственного интеллекта обладают взаимодополняющими достоинствами. Классические алгоритмы отличаются детерминированностью, предсказуемой скоростью работы на центральных процессорах и отсутствием потребности в больших размеченных наборах данных. Они остаются незаменимыми в условиях жёстких аппаратных ограничений, когда каждая миллисекунда обработки критична, а поведение алгоритма должно быть строго обосновано. В то же время методы глубокого обучения обеспечивают высокую точность и устойчивость к вариативности условий съёмки, автоматически извлекая сложные признаки из данных, но требуют значительных вычислительных ресурсов и трудоёмкой подготовки обучающих выборок.

Ни один из подходов не является универсальным. В связи с этим перспективным направлением становится гибридная обработка, объединяющая сильные стороны обоих методов. Например, на начальном этапе можно применить классические фильтры и морфологические операции для быстрого выделения областей интереса, после чего передать эти области в компактную нейросеть для точной классификации или детекции. Такой каскадный подход позволяет сократить объём данных, поступающих на нейросетевой модуль, и снизить вычислительную нагрузку, сохранив при этом высокую точность. Выбор конкретной архитектуры гибридного решения определяется требованиями к точности, доступными вычислительными ресурсами и условиями эксплуатации, однако сам принцип объединения детерминированных математических преобразований с обученными нейросетевыми модулями позволяет достичь оптимального баланса между качеством и эффективностью.

Библиографический список

1. Гонсалес, Р. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – 3-е изд. – М. : Техносфера, 2012. – 1104 с.
2. Шапиро, Л. Компьютерное зрение / Л. Шапиро, Дж. Стокман ; пер. с англ. А.А. Богуславского ; под ред. С.М. Соколова. – Москва : БИНОМ. Лаборатория знаний, 2006. – 752 с.
3. Alzubaidi, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions / L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-dujaili, Y. Duan, O. Al-Shamma, J.I. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan // Journal of Big Data. – 2021. – Vol. 8, No. 1. – P. 53.
4. He, K. Deep residual learning for image recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – Las Vegas, 2016. – P. 770–778.
5. Redmon, J. You Only Look Once: Unified, Real-Time Object Detection / J. Redmon, S. Divvala, R. Girshick, A. Farhadi // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 779–788.
6. OpenCV: библиотека компьютерного зрения [Электронный ресурс]. – Режим доступа: <https://opencv.org/>.

УДК 004.65; ГРНТИ 20.23.21

ОСОБЕННОСТИ ХРАНЕНИЯ ДАННЫХ В УСЛОВИЯХ РАЗНОРОДНОСТИ ИНФОРМАЦИОННЫХ ПОТОКОВ

А.А. Учуватова, Н.Н. Гринченко

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, uchuvatova_nastya@mail.ru

Аннотация. В статье рассматривается проблема организации хранения данных в условиях гетерогенности информационных потоков, характеризующихся множеством форматов представления информации — от строго структурированных реляционных таблиц до слабо структурированных журналов событий и графовых моделей социальных взаимодействий. Обосновывается ограниченность классических реляционных СУБД при обработке таких разнородных массивов, а также анализируются функциональные возможности NoSQL-решений, обеспечивающих масштабируемость и гибкость схемы данных. Особое внимание уделяется принципам полиглотного хранения, предполагающего совместное использование различных типов баз данных, каждая из которых оптимизирована для вы-

полнения специфических операций. На примере архитектуры интернет-магазина демонстрируется практическая реализация подхода, включающая одновременное применение реляционной базы данных, документоориентированного хранилища, кэш на основе ключ-значение и графовой базы данных.

Ключевые слова: разнородные данные, реляционные базы данных, NoSQL, полиглотное хранение, интеграция данных, масштабируемость, большие данные.

FEATURES OF DATA STORAGE IN THE CONTEXT OF HETEROGENEOUS INFORMATION FLOWS

A.A. Uchvatova, N.N. Grinchenko

Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, uchvatova_nastya@mail.ru

Abstract. The article addresses the problem of data storage organization under conditions of heterogeneity of information flows characterized by multiple data representation formats — from strictly structured relational tables to semi-structured event logs and graph models of social interactions. The limitations of classical relational database management systems in processing such heterogeneous data arrays are substantiated, and the functional capabilities of NoSQL solutions that ensure scalability and data schema flexibility are analyzed. Special attention is paid to the principles of polyglot persistence, which involve the joint use of different types of databases, each optimized for performing specific operations. Using the architecture of an online store as an example, the practical implementation of the approach is demonstrated, including the simultaneous use of a relational database, a document-oriented storage, a key-value cache, and a graph database.

Keywords: heterogeneous data, relational databases, NoSQL, polyglot storage, data integration, scalability, big data.

Введение

Современная информационная среда характеризуется многообразием источников и неоднородностью форматов данных [1]. Это традиционные таблицы с заказами, нагруженные вложенными структурами JSON-документы из мобильных приложений, миллионы логов от датчиков, запутанные связи между пользователями в социальных сетях. Раньше большинство задач решалось с помощью реляционных баз данных – строгих, надёжных, проверенных. Но чем разнообразнее становятся потоки информации, тем чаще становится очевидным наличие ограничений такого подхода.

Невозможно поместить в реляционную таблицу информацию об объектах, для которых заранее неизвестно, сколько и каких именно характеристик будет иметь каждый объект. Или добиться стабильной работы такой базы на сотнях серверов одновременно даже при росте нагрузки. При необходимости быстро искать связи между людьми, товарами и событиями реляционная модель так же начинает давать сбои.

Именно поэтому актуальными становятся NoSQL-системы. Они не ориентированы на универсальность, а напротив — специализированы для работы с конкретными типами данных. Одна система эффективно функционирует с документальными структурами, другая — с высокопроизводительными запросами по ключу, третья — с интенсивными потоками временных рядов, четвёртая — с графовыми моделями связей. В связи с этим обоснованным является их совместное использование, при котором каждая система решает свою частную задачу. Как показывает практика, такая интеграция не только возможна, но и целесообразна. Данный подход получил название полиглотного хранения.

1. Ограничения реляционных СУБД в условиях гетерогенности данных

Реляционные базы данных представляют собой классический инструмент, основанный на строгом математическом аппарате. Они обеспечивают целостность данных и надёжность транзакций. Это является критически важными для финансового учёта, систем бронирования и различных видов отчётности. Однако наряду с достоинствами имеются и существенные ограничения.

Главная проблема заключается в жёсткой фиксации схемы данных. Перед началом хранения необходимо точно определить состав таблиц, перечень столбцов, типы данных и связи между таблицами. В ситуациях, когда структура данных изменяется динамически, реляционная модель требует проведения миграций. Например, если приложение начинает передавать новые поля, не предусмотренные исходным проектом. Миграцией называют процесс преобразования базы данных, меняющий версию и затрагивающий схему. Она является длительной и рискованной операцией, часто требующий остановки сервиса [4]. В контексте современных гибких методологий разработки это обстоятельство становится существенным ограничением.

Вторая проблема связана с масштабированием. Реляционные базы данных эффективно масштабируются вертикально. Для этого достаточно установить более производительный сервер, увеличить объём памяти и вычислительные мощности. Но когда данных становится так много, что ни один, даже самый мощный сервер, не справляется, приходится думать о горизонтальном масштабировании. Оно заключается в распределении данных по множеству узлов. В реляционных СУБД это достигается посредством сегментирования данных, но реализовать его сложно. Кроме того, после сегментирования часто перестают корректно функционировать привычные SQL-запросы с объединением таблиц, а также транзакции между отдельными сегментами.

Наконец, реляционные базы данных демонстрируют ограниченную совместимость с неструктурированными данными. При наличии глубоко вложенного JSON-документа в реляционной модели придётся либо распределить его по множеству таблиц, либо сохранить целиком в BLOB-поле, из которого последующий поиск данных становится неэффективным. Оба варианта являются неудобными с практической точки зрения.

2. Специализированные подходы к хранению данных в NoSQL-системах

NoSQL-системы развиваются в направлении функциональной специализации. В отличие от универсальных реляционных СУБД, каждая NoSQL-система оптимизирована для решения ограниченного круга задач. Рассмотрим четыре основных типа таких систем.

Документориентированные базы (MongoDB, CouchDB). В них данные хранятся в виде документов, чаще всего в виде JSON-объектов. Каждый документ – это самостоятельная единица, и у разных документов может быть разный набор полей. Это позволяет отказаться от жёсткой схемы, принятой в реляционных моделях. Документориентированные базы данных является идеальным вариантом для каталогов товаров, профилей пользователей, контентных систем и других случаев, где у каждого объекта своя структура [5]. Так же, за счет поддержки документами вложенных массивов, можно хранить, например, список позиций заказа внутри документа заказа. Это позволяет получить результат без создания дополнительных JOIN-соединения между таблицами.

Базы «ключ-значение» (Redis, Memcached). Данный тип представляет собой наиболее простую и высокопроизводительную модель хранения. Доступ к значению осуществляется по уникальному ключу, при этом внутренняя структура значения не регламентируется. Такие базы функционируют преимущественно в оперативной памяти. За счет этого временные задержки измеряются микросекундами. Такие базы используют для кэширования, управление сессиями, организации очередей задач и в других случаях, когда особенно важна скорость обработки запросов [6].

Столбцовые хранилища (Apache Cassandra, HBase). Такие системы заточены под большие объёмы данных и работу на тысячах серверов. Данные группируются в семейства столбцов, что позволяет эффективно сжимать разреженные матрицы. Типичными задачам являются сбор логов, обработка временных рядов, генерируемых датчиками, и аналитика в реальном времени.

Графовые базы (Neo4j, ArangoDB). В базах такого типа приоритетное значение имеют не сами объекты, а связи между ними. Узлы и рёбра структурированы таким образом, что можно быстро перемещаться по графу, даже при наличии миллионов элементов. Такие

системы являются эффективным инструментом для реализации социальных сетей, рекомендательных систем и анализа мошеннических схем.

3. Интеграция разнородных баз данных в единой системе

При анализе реального прикладного программного обеспечения обнаруживается, что в нём редко присутствует только один тип данных. В интернет-магазине одновременно сосуществуют строго структурированные заказы, гибкий каталог товаров, пользовательские сессии и сложные связи для построения рекомендаций. Попытка хранить всю информацию в единой базе данных – значит выбирать «золотую середину», которая на самом деле оказывается неоптимальной ни для одной из решаемых задач.

Интеграция нескольких баз данных в одной архитектуре [2, 3] позволяет подобрать для каждого типа данных идеальное хранилище, вместо того чтобы пытаться приспособить одну базу под все задачи. Кроме того, разные части системы можно масштабировать независимо друг от друга. Так, кэш может быть усилен без вмешательства в основное хранилище заказов. Такой подход повышает общую надёжность. Например, при отказе кэша основные данные остаются неповреждёнными, а система сохраняет работоспособность, пусть и на сниженном уровне производительности.

Сделать так, чтобы несколько разных баз данных работали как единое целое, достаточно сложно. Сразу возникает вопрос согласованности, поскольку одни и те же или логически связанные данные могут храниться в разных местах, а, значит, нужно постоянно следить, чтобы они не противоречили друг другу. Решать это приходится посредством использования распределённых транзакций или асинхронной синхронизации через шину сообщений. Не менее важно обеспечить унифицированный доступ, поскольку приложениям совершенно не нужно знать, где расположено каждое хранилище и как к нему обращаться. Для этого обычно делают слой агрегации или внедряют сервисную шину, которая скрывает всю внутреннюю сложность. Кроме того у каждого типа хранилища существуют свои требования к обслуживанию. Например, для первого необходимо одно расписание резервного копирования, а для второго – особые правила архивации и удаления данных. И наконец, в распределённой системе с несколькими базами невозможно обойтись без централизованного мониторинга. Необходимо видеть, как работает каждый узел, и иметь возможность быстро реагировать на сбои.

4. Практическая реализация полиглотного подхода на примере интернет-магазина

Рассмотрение практического применения изложенных принципов проведём на примере архитектуры интернет-магазина. Для обеспечения бесперебойного функционирования такой системы необходимо решить несколько задач.

Во-первых, требуется надёжное хранение заказов и платежей, где ошибки недопустимы. Во-вторых, необходимо управление каталогом товаров, при котором различные типы товаров могут иметь разные характеристики, и все эти данные должны согласованно существовать в едином хранилище. В-третьих, следует обеспечить быстрый доступ к корзине покупателя и данным о его сессии, поскольку пользователи критичны к задержкам загрузки. И наконец, требуется формирование персонализированных рекомендаций на основе истории покупок, чтобы покупателю предлагались релевантные товары.

В данной работе предлагается архитектура решения, которая включает следующие компоненты (рис. 4):

1. Реляционная база данных PostgreSQL отвечает за хранение заказов, платежей и учётных записей пользователей. Она использует строгую схему и поддерживает работу с транзакциями и внешними ключами. Это позволяет исключить компромиссы в отношении целостности данных.

2. Документоориентированная база данных MongoDB используется для управления каталогом товаров. Каждый товар представлен в виде JSON-документа с произвольным набором полей.

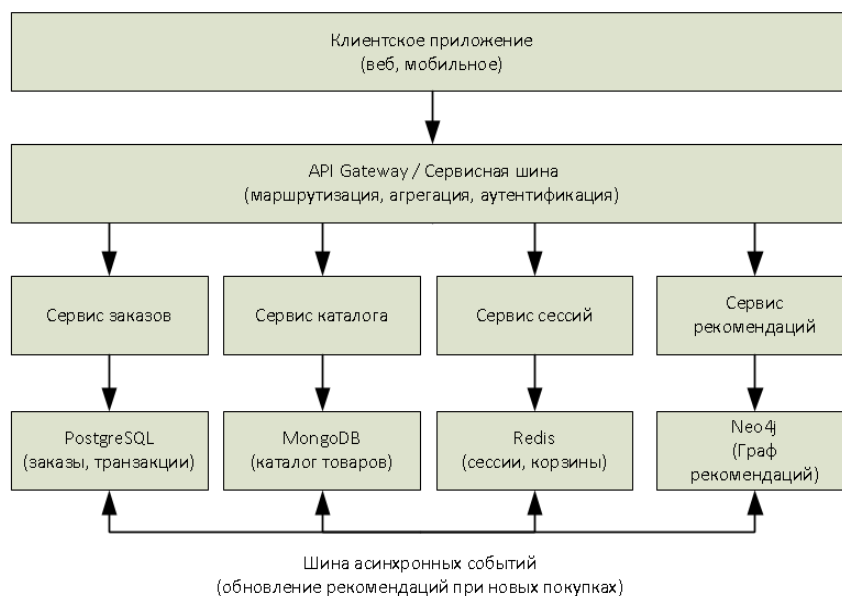


Рис. 4. Архитектура полиглотного хранения данных интернет-магазина

3. База данных типа «ключ-значение» Redis хранит в оперативной памяти пользовательские сессии и содержимое корзины. За счет этого обеспечивается мгновенная загрузка страниц даже во время максимальных нагрузок.

4. Графовая база данных Neo4j [7] обслуживает рекомендательную систему. Пользователи, товары, покупки и просмотры моделируются как узлы и рёбра. Запрос «найти товары, которые покупали пользователи, похожие на данного...» выполняется за миллисекунды. Для этого достаточно пройти по связанным узлам, не делая сложных соединений.

Взаимодействие компонентов организовано следующим образом. При оформлении заказа сервис заказов сохраняет все данные в PostgreSQL и публикует событие в шину сообщений. Сервис рекомендаций подписан на данные события. Получив уведомление о новом заказе, он добавляет в графовую базу данных ребро между пользователем и купленным товаром. При обновлении каталога изменения попадают в MongoDB. Для наиболее популярных товаров дополнительно обновляется кэш в Redis, что ускоряет их загрузку. Клиентские запросы к корзине или проверка сессии направляются непосредственно в Redis. За счет этого обеспечивается высокая скорость и не создаётся избыточная нагрузки на основные базы данных.

Таким образом, предложенная архитектура позволяет каждой подсистеме выполнять свою специализированную функцию. Заказы хранятся надёжно и в строгом соответствии с требованиями к финансовым операциям. Каталог остаётся гибким, поскольку добавлять новые категории и характеристики можно без тяжелых миграций. Высокая производительность достигается за счёт использования кэша. Рекомендации формируются релевантно и с минимальными задержками.

Заключение

Не существует универсальной базы данных, оптимальной для всех сценариев использования. В условиях гетерогенности информационных потоков целесообразно применять гибкую стратегию хранения. Для финансовых транзакций следует использовать реляционную базу данных, обеспечивающую надёжность и целостность. Для каталога товаров предпочтительным является документоориентированное хранилище. Для управления сессиями и корзинами пользователей эффективен сверхбыстрый кэш типа «ключ-значение». Для построения рекомендательных систем наиболее подходит графовая база данных.

Реализация подхода полиглотного хранения, при котором в рамках единой системы совместно используются несколько типов СУБД, требует дополнительных усилий, включая

синхронизацию данных, унификацию доступа к разнородным хранилищам и мониторинг состояния всех компонентов. Однако преимущества, обусловленные решением каждой задачи оптимальным инструментом, полностью оправдывают указанные затраты.

Библиографический список

1. Гуляева К. А., Артемьева И. Л.. Технология автоматизации процесса формирования систем понятий и баз знаний для предметных областей с объектами сложной структуры// Вестник Рязанского государственного радиотехнического университета. 2023. № 85. С. 65-82.
2. Пылькин А.Н., Тишкина В.В. Информационная система формирования управленческих решений по улучшению текущего состояния объекта управления // Вестник Рязанского государственного радиотехнического университета. 2025. № 91. С. 76-87.
3. Сермягин К. А., Пылькин А. Н. Построение эффективной архитектуры для высоконагруженных и масштабируемых сервисов на FastAPI: принципы и лучшие практики// Вестник Рязанского государственного радиотехнического университета. 2025. № 92. С. 49–57.
4. Редмонд Э., Уилсон Д.Р. Семь баз данных за семь недель. Введение в современные базы данных и идеологию NoSQL. – М.: ДМК Пресс, 2013. – 384 с.
5. Фаулер М., Садаладж П.Дж. NoSQL: новая методология разработки нереляционных баз данных. – М.: Вильямс, 2013. – 192 с.
6. Комаров В.И. Путеводитель по базам данных. – М.: ДМК-Пресс, 2024. – 520 с.
7. Отраднов К. К., Алёшкин А. С., Калинин В. Н. Преимущества использования графовых баз данных при разработке прикладных приложений // Вестник Рязанского государственного радиотехнического университета. 2023. № 86. С. 73-85.

УДК 004.89; ГРНТИ 28.23.15

УНИВЕРСАЛЬНЫЙ ВЕБ-СЕРВИС ДЛЯ ДИАГНОСТИКИ И УСТРАНЕНИЯ ПОЛОМОК С ПОМОЩЬЮ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

П.Р. Ахадов, Н.Н. Гринченко, И.Д. Громко

*Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, Eagersupchik@gmail.com*

Аннотация. Статья посвящена актуальности проблемы оперативной диагностики бытовых поломок, обзору существующих решений в сфере ремонта бытовой техники, электроники и интерьерных работ, выявлению их функциональных ограничений, а также описанию архитектуры и основных принципов работы разрабатываемого веб-сервиса «AI Repair», использующего искусственный интеллект для идентификации неисправностей.

Ключевые слова: веб-сервис, искусственный интеллект, диагностика неисправностей, идентификация поломок, устранение проблем, интеллектуальные системы, цифровая трансформация, обслуживание и ремонт.

A UNIVERSAL WEB SERVICE FOR DIAGNOSTICS AND TROUBLESHOOTING USING ARTIFICIAL INTELLIGENCE

P. R. Ahadov, N.N. Grinchenko, I.D. Gromko

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, Eagersupchik@gmail.com*

Abstract. The article is devoted to the relevance of the problem of rapid diagnosis of household breakdowns, a review of existing solutions in the field of repair of household appliances, electronics and interior work, identifying their functional limitations, as well as describing the architecture and basic principles of the developed web service «AI Repair», which uses artificial intelligence to identify faults.

Keywords: web service, artificial intelligence, fault diagnosis, fault identification, problem solving, intelligent systems, digital transformation, maintenance and repair

Введение

В повседневной жизни пользователи часто сталкиваются с проблемой оперативного выявления причин поломок бытовой техники, электроники или элементов интерьера. Самостоятельная диагностика требует значительных временных затрат и не гарантирует достоверности результатов. Обращение к профильным специалистам сопряжено с высокими финансовыми издержками, необходимостью личного присутствия и ожиданием очереди на обслуживание. В ряде случаев пользователь не может объективно оценить сложность проблемы, что приводит либо к излишним тратам, либо к затягиванию ремонта.

В качестве решения предлагается разработка веб-сервиса «AI Repair», позволяющего автоматизировать процесс первичной диагностики неисправностей с предложением пользователю возможных способов их устранения.

Цели статьи — обосновать актуальность создания веб-сервиса с использованием искусственного интеллекта для идентификации поломок, провести сравнительный анализ существующих аналогов и представить архитектуру разрабатываемого решения [1].

В качестве эффективного решения данной проблемы предлагается разработка универсального веб-сервиса «AI Repair», предназначенного для автоматизации процесса первичной диагностики неисправностей различной природы. Проблема может касаться как ремонта бытовой техники или электроники, так и решений в области интерьера или инженерных коммуникаций. Ключевой особенностью сервиса является использование технологий искусственного интеллекта (ИИ) для идентификации причин поломок и предоставления пользователю персонализированных рекомендаций по их устранению.

Целью настоящей статьи является обоснование актуальности создания веб-сервиса с применением ИИ для диагностики неисправностей, проведение сравнительного анализа существующих цифровых решений на российском рынке, а также представление архитектуры разрабатываемого продукта.

1. Современные средства диагностики поломок

На сегодняшний день российский рынок цифровых сервисов для ремонта представлен преимущественно платформами-агрегаторами услуг (например, «YouDo», «Профи») и узкоспециализированными сайтами по ремонту отдельных видов техники. Такие сервисы [2–4] решают задачу поиска исполнителя, но не предоставляют пользователю инструментов для самостоятельной предварительной диагностики. В большинстве случаев пользователь вынужден самостоятельно искать информацию о причинах неисправности, что не всегда приводит к успеху, либо сразу обращаться к платным специалистам без понимания сути проблемы.

Существующие решения не реализуют функцию интеллектуальной поддержки: отсутствует интерактивный диалог с пользователем, не используются алгоритмы машинного обучения для анализа описаний неисправностей, а рекомендации носят общий характер. Это создаёт функциональный пробел между потребностью в быстрой диагностике и возможностями доступных сервисов.

Для выявления функциональных ограничений существующих решений целесообразно провести их сравнительную оценку по ключевым критериям.

2. Сравнительный анализ существующих решений

Для выявления функциональных ограничений существующих платформ проведён сравнительный анализ по ключевым критериям. В качестве объектов исследования были выбраны наиболее популярные на российском рынке сервисы: агрегаторы услуг «YouDo» и «Профи», а также узкоспециализированные платформы, ориентированные на ремонт опре-

делённых видов техники. В качестве эталона для сравнения выступает архитектура и функционал проектируемого сервиса «AI Repair».

Результаты анализа сведены в таблицу 1, где наглядно продемонстрированы различия в подходах к диагностике и взаимодействию с пользователем.

Таблица 1. Сравнительная характеристика сервисов диагностики и ремонта

Критерий	YouDo	Профи	Специализированные сервисы ремонта	AI Repair
Диагностика поломок с помощью ИИ	Нет	Нет	Ограничено	Да
Охват категорий (техника, электроника, интерьер)	Частично	Частично	Нет	Да
Интерактивное взаимодействие с пользователем	Ограничено	Ограничено	Нет	Да
Предложение вариантов устранения	Ограничено	Ограничено	Да	Да
Вызов специалиста	Да	Да	Да	Опционально
Индивидуальный подход к проблеме	Нет	Нет	Ограничено	Да

Анализ данных таблицы 1 показывает, что существующие решения имеют ряд системных недостатков. Они ориентированы преимущественно на организацию вызова специалистов и не предоставляют пользователю инструментов для самостоятельной предварительной диагностики. Ни одна из рассмотренных платформ не сочетает в себе широкий охват типов поломок с использованием искусственного интеллекта для интерактивного анализа проблемы.

3. Анализ результатов сравнительного исследования

На основе проведённого анализа можно сделать вывод о наличии системного разрыва между существующими сервисами по подбору мастеров и реальной потребностью пользователей в быстрой, доступной и интеллектуальной первичной диагностике поломок.

Ключевыми функциональными пробелами существующих платформ являются:

- отсутствие интеллектуального помощника, способного задавать уточняющие вопросы и сужать круг возможных неисправностей на основе ответов пользователя;
- необходимость самостоятельного поиска информации или немедленного обращения к специалистам без предварительной оценки сложности поломки;
- узкая специализация сервисов, не позволяющая решать комплексные задачи в рамках одной платформы.

Отсутствие интеллектуального помощника приводит к тому, что пользователь вынужден либо тратить время на самостоятельный поиск информации в разрозненных источниках, что не гарантирует успеха, либо сразу обращаться к платным специалистам. При этом у пользователя нет объективного понимания сложности неисправности и потенциальной стоимости ремонта. Это создаёт функциональный пробел, который не закрывается существующими решениями и формирует запрос на создание универсальной интеллектуальной системы.

4. Обоснование необходимости разработки нового сервиса

На основе проведённого анализа можно сделать вывод о наличии системного разрыва между существующими сервисами по подбору мастеров и потребностью пользователей в быстрой, доступной и интеллектуальной первичной диагностике поломок. Разрабатываемый веб-сервис «AI Repair» призван заполнить этот пробел, предложив пользователю:

- единый интерфейс для работы с разными категориями неисправностей;
- интеллектуального помощника на основе ИИ для идентификации поломок;
- интерактивный сценарий взаимодействия с уточняющими вопросами;
- рекомендации по возможному устранению неисправности с опцией вызова специалиста при необходимости.

5. Архитектура разрабатываемого сервиса

Прототип веб-сервиса «AI Repair» строится на основе модульной архитектуры, включающей [1, 5]:

- фронтенд-часть с адаптивным интерфейсом для взаимодействия с пользователем;
- бэкенд-компонент, отвечающий за обработку данных и логику диагностики;
- модуль искусственного интеллекта, реализующий алгоритмы классификации поломок на основе вводимых пользователем признаков;
- базу знаний, содержащую описание типовых неисправностей, их причин и способов устранения [6].

Такой подход позволяет обеспечить масштабируемость решения и его адаптацию под различные категории поломок.

Заключение и перспективы развития платформы

Проведённый в рамках данной работы анализ существующих цифровых сервисов в сфере диагностики и организации ремонта позволил выявить системный функциональный пробел на современном рынке. Существующие решения, представленные агрегаторами услуг и узкоспециализированными платформами, не обеспечивают пользователя инструментами для интеллектуальной, быстрой и доступной первичной диагностики. Главная задача таких платформ — соединить заказчика и исполнителя. Они не помогают разобраться в причинах неисправности, поэтому пользователю приходится тратить время на поиски решения или сразу вызывать мастера, не зная точной стоимости и сложности ремонта.

Разрабатываемый веб-сервис «AI Repair» направлен на комплексное устранение выявленных ограничений. Внедрение технологий искусственного интеллекта позволяет не просто автоматизировать рутинные процессы, но и создать принципиально новый пользовательский опыт. Сервис предлагает единый интерфейс для работы с различными категориями неисправностей. Это объединение множества проблем: от ремонта бытовой техники и электроники до элементов интерьера и инженерных систем, что обеспечивает универсальность предлагаемого сервиса. Ключевым преимуществом является интеллектуальный помощник, реализующий интерактивный сценарий взаимодействия [7], когда система задаёт уточняющие вопросы, анализирует ответы и на основе базы знаний формирует гипотезы о причинах поломки, предлагая персонализированные рекомендации по их устранению.

Результаты проведённого исследования подтверждают высокую актуальность и целесообразность дальнейшей разработки и внедрения данного сервиса. Практическая значимость проекта заключается в потенциальном снижении временных и финансовых затрат пользователей, а также в повышении общей цифровой грамотности населения в вопросах бытового обслуживания [9-12].

Перспективы развития платформы «AI Repair» включают несколько ключевых направлений:

- расширение базы знаний, то есть постоянное наполнение системы новыми кейсами, типовыми неисправностями и методами их устранения для повышения точности диагностики [6];
- совершенствование алгоритмов посредством интеграции более сложных моделей машинного обучения, способных анализировать не только текстовые описания [8], но и мультимедийные данные (например, фотографии дефектов), что выведет точность идентификации поломок на новый уровень;

– интеграция с внешними экосистемами, поскольку создание API для взаимодействия с сервисами вызова специалистов, интернет-магазинами запчастей и системами «умного дома» позволит сформировать единую цифровую среду для решения бытовых проблем.

Таким образом, создание веб-сервиса «AI Repair» является своевременным ответом на вызовы современного рынка бытовых услуг и представляет собой значимый шаг в направлении цифровой трансформации сферы ремонта и технического обслуживания.

Библиографический список

1. Пылькин А.Н., Сермягин К. А. Построение эффективной архитектуры для высоконагруженных и масштабируемых сервисов на Fastapi: принципы и лучшие практики // Вестник Рязанского государственного радиотехнического университета. 2025. № 92. С. 49-56.
2. YouDo. Платформа для поиска специалистов. — Режим доступа: <https://youdo.com/>
3. Профи. Сервис по подбору специалистов. — Режим доступа: <https://profi.ru/>
4. Смирнов, В. А. Цифровые платформы в сфере услуг: тенденции и перспективы. — М.: Издательский дом НИУ ВШЭ, 2022.
5. Пылькин А.Н., Тишкина В.В. Информационная система формирования управленческих решений по улучшению текущего состояния объекта управления // Вестник Рязанского государственного радиотехнического университета. 2025. № 91. С. 76-87.
6. Гуляева К. А., Артемьева И. Л.. Технология автоматизации процесса формирования систем понятий и баз знаний для предметных областей с объектами сложной структуры// Вестник Рязанского государственного радиотехнического университета. 2023. № 85. С. 65-82.
7. Каширин И. Ю. Формирование LLM-ориентированных ресурсов знаний на основе генерации дополнительной информации // Вестник Рязанского государственного радиотехнического университета. 2025. № 94. С. 85-97.
8. Леохин Ю. Л., Фатхулин Т. Д. Разработка методов и алгоритма формализации текстового запроса к онлайн-сервисам, генерирующим изображения посредством нейросетевых технологий // Вестник Рязанского государственного радиотехнического университета. 2023. № 85. С. 82-95.
9. Как нейросети трансформируют техобслуживание и ремонт на производстве / CNews. — Текст: электронный // CNews. — URL: https://axenix.cnews.ru/articles/2024-11-22_umnyj_remont_kak_genii_menyaet_protssesy (дата обращения: 29.03.2026).
10. Ремонт ноутбуков с помощью искусственного интеллекта (ИИ) / RemAdmin. — Текст: электронный // RemAdmin: — URL: <https://remadmin.ru/?p=718> (дата обращения: 29.03.2026).
11. Искусственный интеллект в ремонте и обслуживании техники / [Блог компании]. — Текст: электронный // URL: <https://xn--e1agacsdhpmkbam.xn--p1ai/tpost/hv01tall31-iskusstvennii-intellekt-v-remonte-kak-ne> (дата обращения: 29.03.2026).
12. Новая российская технология предскажет поломку двигателя заранее. Уникальный прорыв с ИИ / 360.ru. — Текст: электронный // 360.ru: — URL: <https://360.ru/news/tehnologii/novaja-rossijskaja-tehnologija-predskazhet-polomku-dvigatelja-zarancee-unikalnyj-proryv-s-ii/> (дата обращения: 29.03.2026).

УДК 004.3; ГРНТИ 50.33.00

ПРОАКТИВНАЯ СИСТЕМА АНАЛИЗА И ПРОГНОЗИРОВАНИЯ МИКРОКЛИМАТА НА БАЗЕ МИКРОКОНТРОЛЛЕРА ESP32-S3

Н.Н. Гринченко, С.А. Кулезнёв, И.С. Лёвушкин, Р.В. Хруничев

Рязанский государственный радиотехнический университет имени В.Ф. Уткина,
Российская Федерация, Рязань, kusekuse@mail.ru

Аннотация. В данной статье представлен первый этап разработки системы проактивного климат-контроля. Описана архитектура исследовательского измерительного стенда на базе ESP32-S3, предназначенного для сбора и первичной обработки обучающих выборок для моделей машинного обучения. Архитектура стенда построена на FreeRTOS, что позволяет реализовать многозадачность, а также обеспечить неблокирующий опрос датчиков (BME280, DS3231) и запись данных на SD-карту. Для снижения дисперсии шума измерительных каналов применен алгоритм фильтрации Калмана непосредственно на микроконтроллере (Edge Computing). Представленная архитектура является фундаментом для последующей разработки предиктивных моделей временных рядов (LSTM).

Ключевые слова: микроклимат, ESP32-S3, FreeRTOS, фильтр Калмана, сбор данных, обучающая выборка, прогнозирование.

PROACTIVE MICROCLIMATE ANALYSIS AND FORECASTING SYSTEM BASED ON ESP32-S3 MICROCONTROLLER

N.N. Grinchenko, S.A. Kuleznev, I.S. Levushkin, R.V. Khrunichev

*Ryazan State Radio Engineering University named after V.F. Utkin,
Russia, Ryazan, kusekuse@mail.ru*

Abstract. This article presents the first stage of the development of a proactive climate control system. The architecture of a research measuring stand based on ESP32-S3, designed for collecting and primary processing of training samples for machine learning models, is described. The architecture of the stand is based on FreeRTOS, which allows for multitasking, providing non-blocking polling of sensors (BME280, DS3231) and data recording on an SD card. The Kalman filtering algorithm is applied directly on the microcontroller (Edge Computing) to reduce the noise dispersion of the measuring channels. The presented architecture is the foundation for the subsequent development of predictive time series models (LSTM).

Keywords: microclimate, ESP32-S3, FreeRTOS, Kalman filter, data collection, training sample, forecasting.

Введение

В современном мире для обеспечения комфортных условий жизнедеятельности человека повсеместно используются различные климатические системы: от обычных кондиционеров и увлажнителей воздуха до сложных решений, выполняющих самые разные комбинации функций и обеспечивающих, в итоге, контроль параметров помещений, связанных с отоплением, водоснабжением, кондиционированием и т.п. Все эти системы работают по реактивному принципу. Если контролируемые параметры выходят за установленные пороговые значения, то активируются соответствующие механизмы, устраняющие проблемы. Такой подход, обладающий существенной инерционностью, обладает значительными недостатками. Переход к проактивному управлению климатическими системами и комплексами, который позволяет накапливать исторические данные, и затем на их основе прогнозировать динамику среды, позволяет повысить энергоэффективность и улучшить комфортные условия.

1. Цель исследования

В рамках данной работы решается задача разработки специального оборудования – аппаратно-программного стенда для достоверного сбора временных рядов микроклиматических параметров с возможностью их первичной обработки. В дальнейшем данный стенд станет основой для анализа нелинейных процессов в локальной среде.

2. Аппаратная архитектура измерительного стенда

Основой архитектуры предложенного в работе измерительного стенда является двухъядерный микроконтроллер ESP32-S3 [1], который отличается высокой производительностью. Благодаря большому объему оперативной памяти можно реализовать многозадачность операционной системы реального времени, а также выполнять буферизацию данных. Кроме того, выбор микроконтроллера ESP32-S3 продиктован наличием аппаратной поддержки всех ключевых и классических коммуникационных интерфейсов.

Высокоточный метеодатчик BME280 [2, 3] осуществляет сбор первичных данных о параметрах среды, а именно измеряет температуру, относительную влажность и атмосферное давление.

Для анализа временных рядов критически важным является обеспечение строгой временной синхронизации записей, поэтому в устройство добавлен высокоточный модуль часов реального времени DS3231.

Ниже приведена структурная схема аппаратной части измерительного стенда (рисунок 1). Подключение компонентов BME280 и DS3231 модуля сбора данных к микрокон-

троллеру ESP32-S3 происходит через цифровую шину I2C посредством линий SDA и SCL. Поскольку оба подключаемые устройства имеют уникальные адреса, то микроконтроллер может адресовать их выборочно. Модуль управления формирует SCL и задает таким образом тактовую частоту передачи, в то время как по SDA реализуется обмен данными, командами и адресами между модулем сбора данных и модулем управления.

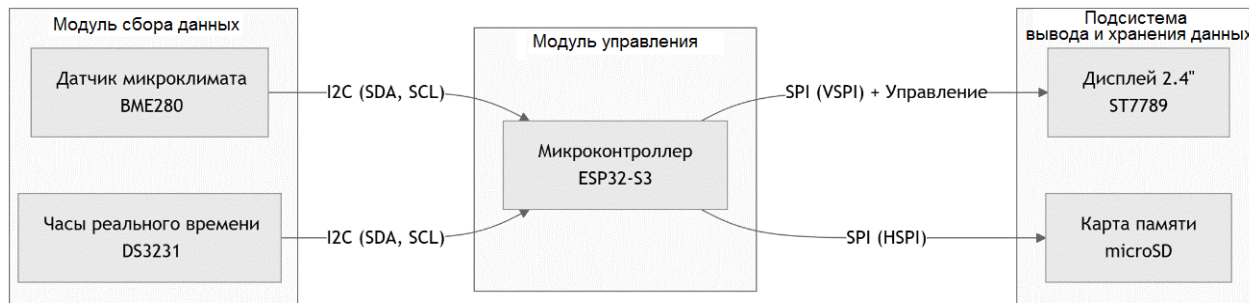


Рис. 1. Структурная схема аппаратной части измерительного стенда

Обработанные данные записываются на карту памяти microSD через интерфейс SPI. Следует отметить, что вероятность конфликтов при работе с данными будет минимальной, поскольку реализовано аппаратное разделение каналов ввода-вывода - I2C используется для считывания данных, а SPI – для записи.

В микроконтроллере ESP32-S3 доступны несколько аппаратных шин SPI. Благодаря этому возможно разделение функций системы по независимым аппаратным каналам. Чтобы повысить надежность системы и предусмотреть конфликты на шине, в предлагаемом стенде реализовано разделение функций вывода информации и ее хранения по независимым аппаратным каналам. VSPI используется для записи данных на карту памяти microSD, а HSPI обеспечивает связь с дисплеем (ST7789) и передачу сигналов управления.

3. Программная реализация с предварительной обработкой данных

При классическом подходе, где все периферийные устройства опрашиваются последовательно в цикле loop, возникает проблема блокировок. Задержка на десятки миллисекунд, которая происходит при записи блока данных на карту памяти, может привести к приостановке процесса опроса датчиков, и, как следствие, искажению формируемого временного ряда.

Для решения данной проблемы в работе предлагается использовать операционную систему реального времени FreeRTOS. Вся система декомпозирована на отдельные задачи, каждой из которых присвоен свой приоритет. Перечислим задачи, пояснив обоснование назначенного приоритета.

1. Циклический опрос модулей BME280 и DS3231 по шине I2C.

Задаче назначен высокий приоритет для исключения задержек при записи данных на карту памяти, и, как следствие, гарантии строгой регулярности сбора исходных данных.

2. Сохранение подготовленных пакетов данных на карту памяти через SPI.

Задаче назначен низкий приоритет, она функционирует в фоновом режиме и не блокирует сбор информации с датчиков по I2C.

3. В перспективе предполагается добавить задачу обработки внешних прерываний от датчиков аварийных ситуаций, например, газового датчика MQ-5. Этой задаче будет назначен наивысший приоритет, чтобы система могла мгновенно реагировать на критические события с фиксацией времени аварии.

Благодаря использованию в операционной системе реального времени FreeRTOS очередей и специальных блокировок обеспечивается синхронизация доступа к общим ресурсам (например, шинам данных) между задачами.

Для дальнейшего использования предлагаемого стенда в задачах прогнозирования динамики среды на основе накопленных исторических данных необходима подготовка данных к последующему машинному обучению. В этих целях авторами предлагается реализовать в системе алгоритм фильтрации Калмана [4, 5]. В отличие от методов скользящего среднего, которые при резких изменениях реальных величин вносят фазовое запаздывание сигнала, метод Калмана позволяет рекурсивно оценить состояние среды. На каждом шаге данный метод учитывает шум измерений (собственные ошибки датчика) и шум процесса (скачки измеряемой величины), минимизируя при этом среднеквадратичную ошибку. Таким образом, для вычисления шумов на каждом шаге фильтр Калмана использует формулу прогнозирования будущего состояния системы, основываясь на предыдущей информации о ее состоянии, и формулу корректировки прогноза, основываясь на новой информации от сенсоров [9]. Реализация данного математического аппарата прямо на микроконтроллере ESP32-S3 дает возможность получать и сохранять на карте памяти готовый для анализа набор значений, состоящий как из сырых (необработанных) данных, так и очищенных от шума. Таким образом, использование линейного фильтра позволит отсеивать измерительный шум аппаратуры, снижая общую дисперсию сигнала, и, в итоге, получать качественный датасет для последующего применения его в задачах машинного обучения. Это, в свою очередь, значительно снизит вычислительную нагрузку на последующую обработку полученного датасета.

Распределение потоков данных между двумя основными модулями показано на рисунке 2. Модуль, отвечающий за сбор и обработку данных, формирует специальную структуру данных – унифицированный пакет Systemdata. Содержимое пакета (отфильтрованные данные с привязкой к временной метке, а также служебная информация) помещается в очередь FreeRTOS и затем используется для обновления экрана. Пользовательский интерфейс реализуется на основе библиотеки LVGL. Следует отметить, что модуль сбора и обработки информации и модуль графического интерфейса работают независимо друг от друга, что обеспечивается за счет использования механизма очереди.

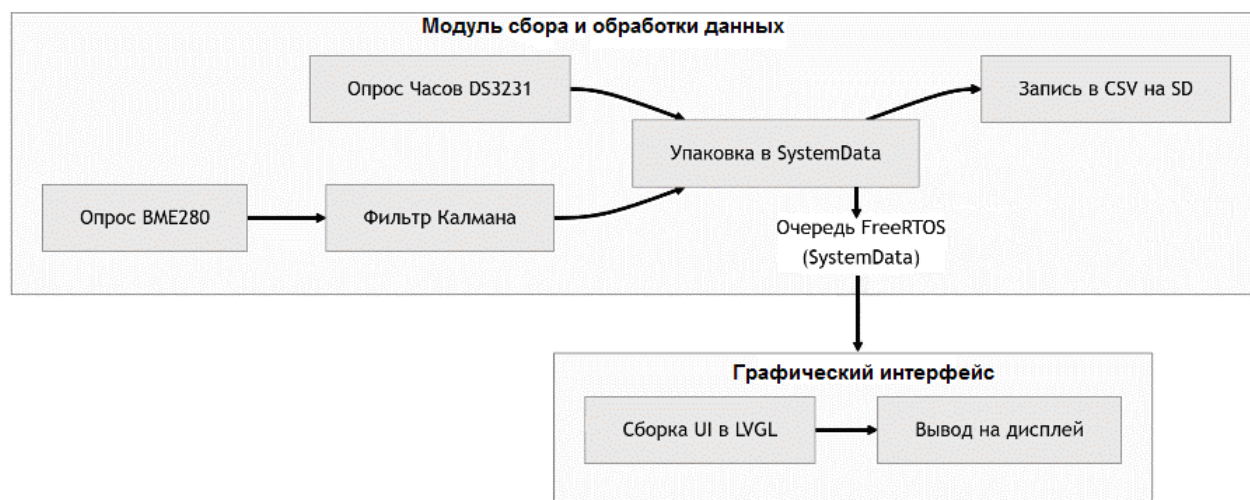


Рис. 2. Структурная схема взаимодействия программных модулей

4. Результаты эксперимента

Для подтверждения работоспособности предложенного в работе измерительного стенда была проведена серия тестов при непрерывной фиксации контролируемых климатических параметров. Итоги эксперимента подтвердили, что стенд работал стабильно без задержек и конфликтов при одновременном использовании шин I2C и SPI. При этом модуль DS3231 обеспечивал корректную генерацию временных меток. Реализация фильтра Калмана прямо на микроконтроллере ESP32-S3 была апробирована в реальных бытовых условиях

[6, 7]. Следует отметить, что при естественной вентиляции в среде наблюдался значительный разброс контролируемых измерений, с которыми фильтр Калмана успешно справился.

Пример работы реализованного измерительного стенда в реальных бытовых условиях приведен на рисунке 3. Приведенные на рисунке графики показывают изменения исходного сигнала влажности («Данные с датчика») и результат обработки сигнала с помощью фильтра Калмана («Фильтрованные данные»). Анализ полученных с помощью стенда графиков показал, что использование линейного фильтра позволило корректно и без задержек отслеживать значительные перепады контролируемых величин, а также сглаживать высокочастотные аппаратные шумы метеодатчика при стабильной влажности. Это критически важно для формирования качественного датасета.

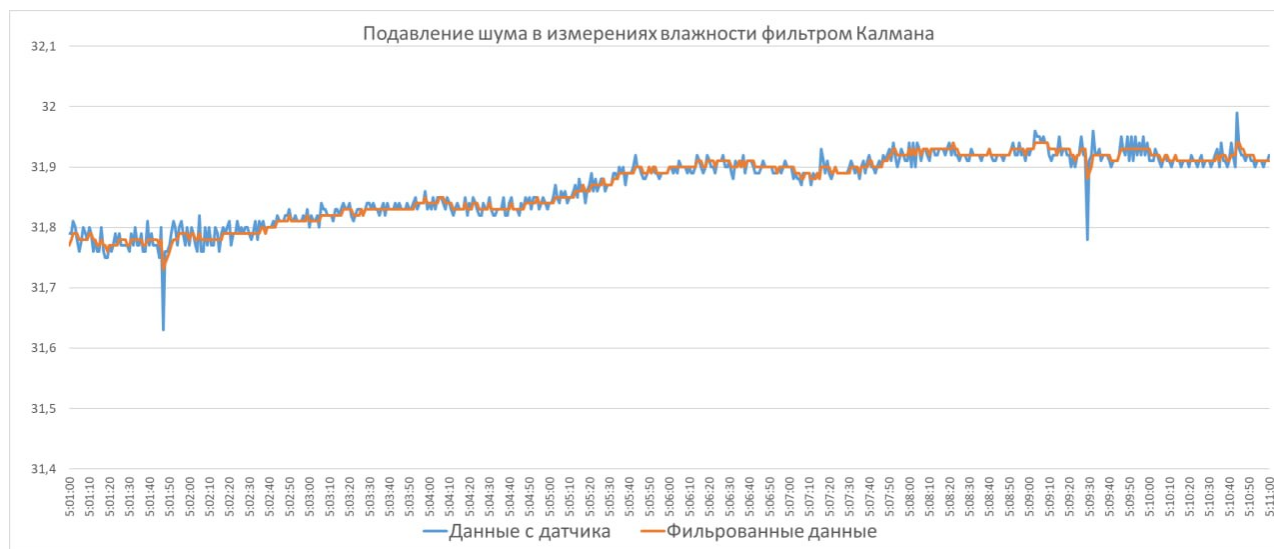


Рис. 3. Результаты эксперимента в реальных бытовых условиях

Заключение

В рамках первого этапа разработки проактивной системы управления микроклиматом был предложен и реализован вариант измерительного стенда на базе двухъядерного микроконтроллера ESP32-S3 и операционной системы реального времени FreeRTOS [8], который продемонстрировал устойчивую работу при одновременном сборе данных с нескольких устройств.

Получила экспериментальное подтверждение гипотеза о целесообразности использования математического аппарата Калмана непосредственно в прошивке микроконтроллера, в результате чего фильтрованные данные могут быть использованы в задачах машинного обучения без предварительной обработки.

Предложенный стенд легко масштабируется благодаря самой архитектуре стенда: возможностями микроконтроллера ESP32-S3 физически подключать дополнительные датчики без изменения схмотехнической структуры, многозадачностью FreeRTOS, единой структурой пакета данных и независимостью работы модулей.

Система в непрерывном режиме генерировала лог-файлы с обработанными с помощью фильтра Калмана данными, которые могут использоваться непосредственно в задачах машинного обучения, что ускоряет переход к применению сформированных пакетов при построении прогнозных моделей.

Таким образом, проект можно признать целесообразным, поскольку он протестирован в реальных условиях, продемонстрировал стабильную работу и обеспечивает получение качественных датасетов для последующего дальнейшего использования в задачах машинного обучения.

В перспективе планируется продолжение работ для получения готового решения, интегрирующего нейросетевое прогнозирование с разработанным стендом, что позволит создать полноценную проактивную систему климатического контроля, принципиально отличающуюся от климатических систем традиционного типа.

Библиографический список

1. Espressif Systems. ESP32-S3 Technical Reference Manual [Электронный ресурс] // Espressif Documentation. – URL: <https://docs.espressif.com/projects/esp-idf/en/latest/esp32s3/> (дата обращения: 20.03.2026).
2. Sensirion. SCD4x CO₂ Sensor Datasheet [Электронный ресурс] // Sensirion. – URL: <https://sensirion.com/products/co2-sensors/> (дата обращения: 20.03.2026).
3. Bosch Sensortec. BME280 Environmental Sensor Datasheet [Электронный ресурс] // Bosch Sensortec. – URL: <https://www.bosch-sensortec.com/products/environmental-sensors/humidity-sensors-bme280/> (дата обращения: 20.03.2026).
4. Welch G., Bishop G. An Introduction to the Kalman Filter // University of North Carolina at Chapel Hill. – 2006. – Vol. 95, No. 2. – P. 1–16. doi:10.1.1.117.6891(дата обращения: 20.03.2026).
5. Zaini N., Eswaran C. A Kalman Filter Framework for Indoor Air Quality Monitoring // Proceedings of the 2020 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS). – 2020. – P. 89–94. doi:10.1109/I2CACIS49202.2020.9140207(дата обращения: 20.03.2026).
6. ASHRAE. ANSI/ASHRAE Standard 62.1-2022: Ventilation for Acceptable Indoor Air Quality. – Atlanta: ASHRAE, 2022. – 104 p. doi:10.1007/978-3-030-05363-5(дата обращения: 20.03.2026).
7. Всемирная организация здравоохранения. Загрязнение воздуха внутри помещений: факты и рекомендации [Электронный ресурс] // WHO. – URL: <https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health> (дата обращения: 20.03.2026).
8. FreeRTOS Real-Time Operating System Documentation [Электронный ресурс] // FreeRTOS. – URL: <https://www.freertos.org/> (дата обращения: 20.03.2026).
9. Лаборатория микроскопов/ Фильтр Калмана: формулы и математика / [Электронный ресурс]. – Режим доступа: https://mp-lab.ru/filtr_kalmana_dlya_nachinayushchih/-regionalnoy-modeli-ved. – Дата доступа: 05.03.2026.

СОДЕРЖАНИЕ

ИНФОРМАЦИЯ О IX МЕЖДУНАРОДНОМ ФОРУМЕ «СОВРЕМЕННЫЕ ТЕХНОЛОГИИ В НАУКЕ И ОБРАЗОВАНИИ» СТНО-2026».....	3
МЕЖДУНАРОДНАЯ НАУЧНО-ТЕХНИЧЕСКАЯ КОНФЕРЕНЦИЯ «СОВРЕМЕННЫЕ ТЕХНОЛОГИИ В НАУКЕ И ОБРАЗОВАНИИ. ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА, МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ И АВТОМАТИЗИРОВАННЫЕ СИСТЕМЫ».....	6
Секция « АЛГОРИТМИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ И СЕТЕЙ».....	6
Шафаяет Х., Пруцков А.В. Обзор методов обнаружения клонов кода на основе инвертированного индекса.....	6
Каширин И.Ю. Вычисление семантического сходства интеллектуальными системами обработки естественного языка.....	9
Игнатов Е.А. Математическое и программное обеспечение информационной системы для формирования электронных интерактивных расчётных стандартов.....	14
Масленников В.В. Теоретические основы квантово-инспирированной оптимизации в задачах с вещественными переменными.....	25
Аксенов М.А., Белов В.В. Способы извлечения, обработки и структуризации разнородных абонентских данных в телекоммуникационной сфере.....	35
Буланов А.А., Пылькин А.Н. Кластеризация данных методом k-средних в n-мерном пространстве с приоритетами и адаптивными центроидами для анализа успеваемости студентов.....	42
Zhashkov A.A., Mikheev M.Yu. Comparative analysis of modern navigation methods in robotic cyber-physical systems.....	46
Ушков М.Е., Бурковский В.Л. Системный подход к визуализации больших графов в Python с использованием OpenGL: проектирование пайплайна ЦП–ГП и анализ масштабируемости.....	51
Лыкин С.В., Крошилин А.В. Современные ИТ-решения для оптимизации процессов записи и проведения приёма пациента в медицинских учреждениях.....	58
Теплов А.А. Метод фрактальных операторов и алгоритмы роста динамических фрактальных структур.....	61
Zhashkova T.V., Pepel L.N., Terekhin M.A. Development of multimodal transport systems based on the integration of unmanned technologies and digital navigation.....	69
Исаев Р.А., Игнатенко В.А. Построение нечетких когнитивных моделей группой экспертов: актуальность задачи и план исследований.....	72
Посохов М.А. Модель контроля динамической системы.....	76
Саксонов Е.А., Посохов М.А. Программный комплекс моделирования процесса контроля динамических систем.....	80
Демидов Н.А. Настройка параметров алгоритма РАСМАР в пространстве проектирования переменной размерности.....	85

Зангин Д.Д., Крошили А.В. Обзор существующих систем для распознавания дорожных знаков и указателей.....	91
Хрусталева И.Д., Крошили С.В. Сравнительный анализ и обоснование архитектуры локальной системы сбора данных в условиях избыточности облачных DWH-платформ.....	98
Богатырев И.В., Жалненков Н.В., Проказникова Е.Н. Квантовая оптимизация маршрутизации в SDN: сравнительный анализ.....	102
Анастасьев А.А., Проказникова Е.Н., Пылькин А.Н. Возможность использования разностных алгоритмов для поиска схожих фрагментов цифровых данных.....	105
Секция « ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ».....	113
Оспан Ж.Б. Применение ИИ в задачах радиомониторинга.....	113
Логинов Д.Ю., Крошили С.В. Изучение метода анализа тональности речи на основе нейронных сетей.....	118
Макеев А.Д., Уткин А.О. Применение генетического алгоритма для оптимизации гиперпараметров трансформера BERT и T5 в задачах извлечения текста.....	121
Крошили А.А. Подход к оценке зрелости процессов медицинской организации в рамках информационного инжиниринга при внедрении ИИ-агентов.....	129
Лизунов М.В., Тронь Т.А. ЭЭГ-корреляты состояния скуки и вовлеченности при работе с типовыми тренажерами СПО: поиск пороговых значений для активации ИИ-помощника.....	132
Каширин И.Ю. Технологии обработки текстов на естественном языке с использованием семантического сходства.....	137
Ветринцев И.А., Цуканова Н.И. Анализ мимики лица в системах видеоконференций на основе методов глубокого обучения.....	141
Лушков Д.М., Филатов И.Ю. Исследование методов построения музыкальных сетов на основе эмоциональной динамики.....	145
Лутиков В.А., Проказникова Е.Н. Учет временной релевантности в алгоритме персонализированной рекомендации мероприятий.....	150
Федулов М.А., Белов В.В. Возможности и ограничения использования нейронных сетей как инструмента систем поддержки принятия решений.....	152
Гудинов Н.В. Сравнительный анализ C++, GO и RUST для реализации подхода быстрого получения первых результатов.....	157
Овечкин Г.В., Прилепский И.А. Учёт внешних факторов в задачах прогнозирования спроса на товары в розничной торговле.....	161
Овечкин Г.В., Прилепский К.А. Применение методов машинного обучения для анализа кода и оценки ответов в системах технического интервьюирования.....	164
Попов А.Н., Жулева С.Ю. Исследование способов автоматизированной категоризации файлов.....	168
Антипов С.С. Нейросетевой анализ рисков в условиях строительства объектов специального назначения.....	176

Буняшин Д.А., Проказникова Е.Н. О возможностях применения систем прогнозирования карьерного роста и выбора траектории профессионального развития.....	179
Ярославцев Ю.В. Применение моделей зрительного внимания для задачи автоматической оценки качества изображений.....	183
Журавлев П.В., Дмитриева Т.А., Проказникова Е.Н. Использование механизмов внимания в трансформерных моделях для персонализированных новостных рекомендаций.....	189
Овечкин Г.В., Хлебников Н.М. Нейросетевые подходы к стилизации фотографий под стиль художника: архитектуры, параметры, сценарии применения.....	192
Петров С.А., Филатов И.Ю. Проектирование рекомендательной системы планирования свободного времени.....	197
Мацуганова М.Д., Казаков В.Е. User-based и item-based подходы коллаборативной фильтрации в задачах анализа и управления ассортиментом.....	202
Коробов А.А., Савинова К.С. Применение градиентного нейросетевого управления для замещения компонентов биотехнических систем.....	206
Овечкин Г.В., Успенский Д.И. Применение архитектуры Mixture of Experts в последовательных рекомендательных системах.....	210
Логунова Е.А. Алгоритмы машинного обучения для управления группой БПЛА.....	215
Кирсанов А.П., Крошили А.В. Интеллектуальные методы формирования персонализированных рекомендаций при выборе одежды с учётом метеорологических условий.....	224
Попов П.А., Цуканова Н.И. Применение графовых нейронных сетей с механизмом внимания в рекомендательных системах.....	229
Анохин В.А. Метод корректировки целевых меток при обучении нейронных сетей оценки качества изображений с учётом взаимодействия дефектов.....	233
Долгов Я.В., Судаков Д.Е. Перспективы применения глубоких нейронных сетей в управлении бионическими протезами кисти.....	237
Табakov Ф.А., Петрова Н.К. Оптимизация алгоритма Minimax в параметризованных шахматах: исследование оценочных функций.....	241
Секция «ЭВМ И СИСТЕМЫ».....	246
Хруничев Р.В. Основные концепции распределённых вычислений: архитектуры, модели согласованности и управление ресурсами.....	246
Хруничев Р.В. Отличия традиционных методов обработки данных от подходов анализа больших данных.....	251
Васильев К.А. Исследование алгоритмов маскирования изображений с использованием ядер размытия сложной формы.....	257
Хруничев Р.В. Облачные решения и инфраструктурные сервисы для анализа данных (Amazon Web Services, Google Cloud Platform, Yandex Cloud)	262

Бибко К.В., Осокин Д.А., Осокин А.В., Костров Б.В. Применение алгоритма муравьиной колонии в системах технического документирования.....	266
Лалин К.П., Муратов Е.Р. Способ локализации кодов AZTEC в реальном времени.....	273
Сокол И.И., Муратов Е.Р. Метод коррекции искажений одномерных штрих-кодов в реальном времени на основе анализа градиентов.....	279
Баранчиков А.И., Абди М. А. М. Разработка протокола бенчмаркинга с учетом покрытия и рабочей нагрузки для сценарно-ориентированного синтеза реляционных тестовых баз данных.....	287
Попов Н.А., Костров Б.В. Исследование влияния оконных функций на качество спектрограмм.....	296
Кузьмина Е.П., Вьюгина А.А. Сравнение традиционных методов обработки изображений и алгоритмов ИИ.....	300
Учуватова А.А., Гринченко Н.Н. Особенности хранения данных в условиях разнородности информационных потоков.....	304
Ахадов П.Р., Гринченко Н.Н., Громко И.Д. Универсальный веб-сервис для диагностики и устранения поломок с помощью искусственного интеллекта.....	309
Гринченко Н.Н., Кулезнёв С.А., Лёвушкин И.С., Хруничев Р.В. Проактивная система анализа и прогнозирования микроклимата на базе микроконтроллера ESP32-S3.....	313

**СОВРЕМЕННЫЕ ТЕХНОЛОГИИ В НАУКЕ И ОБРАЗОВАНИИ –
– СТНО-2026**

Научное издание

В 11 томах

Том 4

Под общей редакцией О.В. Миловзорова.

Подписано в печать 15.06.26. Формат 60х84 1/8.

Бумага офсетная. Печать офсетная.

Гарнитура «Times New Roman».

Усл. печ. л..

Тираж 100 экз. Заказ №.

Рязанский государственный радиотехнический университет,
Редакционно-издательский центр РГРТУ,
390005, г. Рязань, ул. Гагарина, д. 59/1.

ISBN 978-5-7722-0447-4



9 785772 204474 >