Международный журнал информационных технологий и энергоэффективности

Том 11 Номер 7 (69)

2026

Международный журнал информационных технологий и энергоэффективности
Т.11 №7(69) часть 1 2026 г.

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

СОДЕРЖАНИЕ / CONTENT

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Безверхий О.А., Гродзенский Я.С. Количественная оценка скрытых кибервоздействий в замкнутых контурах АСУ ТП термобарической установки на основе цифрового двойника // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7 (69) Ч. 1 с. 4–20

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.056.5:681.518

КОЛИЧЕСТВЕННАЯ ОЦЕНКА СКРЫТЫХ КИБЕРВОЗДЕЙСТВИЙ В ЗАМКНУТЫХ КОНТУРАХ АСУ ТП ТЕРМОБАРИЧЕСКОЙ УСТАНОВКИ НА ОСНОВЕ ЦИФРОВОГО ДВОЙНИКА

Безверхий О.А., Гродзенский Я.С.
ФГАОУ ВО "НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ "МИФИ", Москва, Россия (115409, город Москва, Каширское ш., д.31), e-mail: nik12-2001@yandex.ru

В статье предложена и апробирована методика выявления и количественной оценки области скрытых кибервоздействий в замкнутых контурах автоматизированных систем управления технологическими процессами (АСУ ТП) термобарической установки. В основе методики - цифровой двойник исследовательского назначения автоклавной установки SCHOLZ, реализованный в среде Matlab/Simulink. Введено формальное определение скрытой кибератаки, объединяющее три признака: отсутствие срабатывания промышленной безопасности, необнаружимость оператором и измеримая деградация технологического процесса. Разработаны интегральная метрика деградации, модель наблюдателя (оператора), а также показатели доли скрытых атак, среднего ущерба и интегрального риска. Для пяти типов кибервоздействий построена область скрытых параметров $\Omega_{stealth}$. Эффективность методики продемонстрирована на примере оценки киберустойчивости системы с независимым контуром контроля технологических данных. Предложенная методика формирует основу количественного анализа скрытых кибервоздействий на АСУ ТП.

Ключевые слова: Кибервоздействие, скрытое кибервоздействие, область скрытых кибервоздействий, метрика деградации технологического процесса, киберустойчивость, автоклав, АСУ ТП.

QUANTITATIVE ASSESSMENT OF STEALTH CYBERATTACKS IN CLOSED-LOOP INDUSTRIAL CONTROL SYSTEMS BASED ON DIGITAL TWIN

Bezverkhy O.A., Grodzensky Ya.S.
NATIONAL RESEARCH NUCLEAR UNIVERSITY MEPHI, Moscow, Russia (115409, Moscow, Kashirskoye sh., 31), e-mail: nik12-2001@yandex.ru

This paper proposes and validates a methodology for identifying and quantitatively assessing the domain of stealth cyberattacks in closed-loop industrial control systems of a thermobaric processing unit. The methodology is based on a research-oriented digital twin of a SCHOLZ autoclave system implemented in the Matlab/Simulink environment. A formal definition of a stealth cyberattack is introduced, combining three criteria: the absence of safety-system activation, non-detectability by the operator, and measurable degradation of the technological process. An integral degradation metric, an operator-observer model, and indicators characterizing the proportion of stealth attacks, the average damage, and the overall risk are developed. For five types of cyberattacks, the domain of stealth attack parameters, $\Omega_{stealth}$, is determined. The effectiveness of the proposed methodology is demonstrated through the assessment of cyber resilience in a system equipped with an independent technological data monitoring loop. The proposed approach provides a foundation for engineering-oriented quantitative assessment of stealth cyberattacks against industrial control systems.

Keywords: Cyberattack, stealth cyberattack, stealth cyberattack domain, process degradation metric, cyber resilience, autoclave, Industrial Control System (ICS).

1. Актуальность проблемы скрытых кибервоздействий на промышленные АСУ ТП

В последние годы наблюдается устойчивый рост числа кибератак, направленных на объекты промышленной инфраструктуры. При этом их отличительной особенностью все чаще становится нацеленность не только на нарушение доступности информационных систем, но и на изменение физических параметров технологических процессов.

Особую опасность представляют так называемые скрытые кибервоздействия (stealthy attacks). В отличие от классических определений stealthy attacks [1], в настоящей работе под скрытым кибервоздействием понимаются кибервоздействия, удовлетворяющие одновременно трем условиям: отсутствие срабатывания систем промышленной безопасности, необнаружимость оператором, наличие деградации технологического процесса.

В отличие от обычных атак, приводящих к немедленной аварии или остановке производства, скрытые воздействия реализуются таким образом, чтобы длительное время оставаться незамеченными для оператора и штатных средств мониторинга. При этом технологический процесс постепенно деградирует: ухудшается качество выпускаемой продукции, возникают отклонения параметров работы оборудования.

Важным свойством системы АСУ ТП в этих условиях выступает киберустойчивость, т.е. способность предвидеть, противостоять неблагоприятным условиям, стрессам, атакам или компрометации систем, использующих киберресурсы или поддерживаемых ими, восстанавливаться после них и адаптироваться к ним [2].

Целью исследования является разработка и апробация методики выявления и количественной оценки области скрытых кибервоздействий и киберустойчивости в замкнутых контурах автоматизированной системы управления технологическим процессом (АСУ ТП) термобарической установки. Ключевым инструментом реализации методики выступает цифровой двойник исследовательского назначения, далее для краткости именуемый «цифровой двойник».

В качестве объекта выбрана автоклавная установка SCHOLZ с системой управления на базе контроллера Siemens S7-300. Данное оборудование предназначено для термобарической обработки многослойных композитных материалов. Технологический процесс включает этапы нагрева, выдержки при заданных значениях температуры и давления, а также последующее охлаждение и сброс давления.

Выбор автоклавной установки в качестве объекта обусловлен следующими факторами:

- наличием чётко выраженных замкнутых контуров управления, уязвимых для кибервоздействий на измерительные и управляющие каналы;
- высокой критичностью отклонений параметров;
- типовой архитектурой АСУ ТП, характерной для широкого класса термобарических и химико-технологических установок.

2. Цифровой двойник как инструмент анализа киберустойчивости

В соответствии с ГОСТ Р 57700.37-2021[3], цифровой двойник изделия предполагает наличие двусторонних информационных связей с физическим объектом. В настоящем исследовании такая связь с реальным автоклавом отсутствует, поэтому используемую модель

корректно именовать «цифровым двойником исследовательского назначения», а термин «цифровой двойник» использовать только для краткости.

Цифровой двойник реализован в среде Matlab/Simulink и включает шесть основных подсистем, каждая из которых отвечает за определённый аспект функционирования автоклавной установки:

1. Подсистема промышленного компьютера (загрузки конфигурации) - хранит технологические рецепты и передаёт уставки (сигналы P_set, T_set) в программируемый логический контроллер (ПЛК).

2. Подсистема заготовки (TModel_Subsys) - описывает нагрев обрабатываемого материала под воздействием температуры и давления в рабочей камере и описывается нелинейным дифференциальным уравнением первого порядка:

$$\frac{dT}{dt} = \alpha \cdot (T_{air} - T) \cdot P^{\beta} \cdot (1 + \gamma \mid T_{air} - T \mid), \quad (1)$$

где T – температура заготовки, T_{air} – температура в автоклаве, P – абсолютное давление. Параметры α, β, γ идентифицируются по экспериментальным логам температуры.

3. Подсистема автоклава или технологического объекта (Plant). Автоклав рассматривается как замкнутая емкость постоянного объема, заполненная идеальным газом и описывается тремя уравнениями:

$$\frac{dT}{dt} = \frac{Q - k_{loss}(T - T_{env})}{C_{th}} \quad (2), \qquad \frac{dn}{dt} = w_{in} - w_{out} - w_{relief} \quad (3), \qquad P = \frac{nRT}{V} \quad (4)$$

где Q – тепловая мощность, подводимая к газу; k_{loss} – коэффициент теплопотерь; C_{th} – эффективная теплоёмкость системы; n – количество вещества газа; $w_{in}, w_{out}, w_{relief}$ – массовые расходы через впускной, выпускной и предохранительный клапаны соответственно; R – универсальная газовая постоянная; V – объём автоклава.

4. Подсистема управления (Control) - моделирует функции ПЛК, реализующего алгоритмы автоматического регулирования температуры и давления в автоклавной установке.

5. Подсистема безопасности (Safety) - описывает систему промышленной безопасности, включает модель предохранительного клапана и логику аварийной остановки по вакуумируемым контейнерам.

6. Подсистема вакуумируемых контейнеров (Containers) - моделирует группу из пяти контейнеров, в которых происходит непосредственное формование изделий.

Все подсистемы соединены в замкнутую структуру, отражающую информационные и управляющие потоки реальной АСУ ТП.

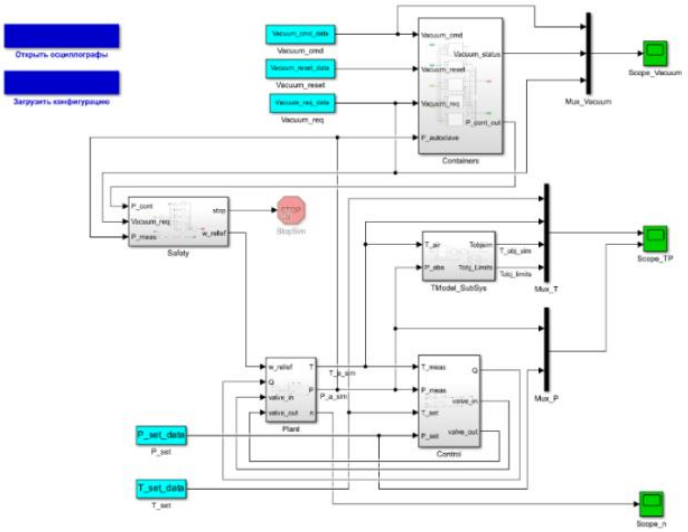


Рисунок 1 - Модель автоклавной установки

Для оценки соответствия полученной модели реальному объекту была выполнена её валидация на экспериментальных логах температуры заготовки. Использованы данные логов шести производственных циклов. Рассчитанная среднеквадратичная ошибка между модельной и экспериментальной температурами составила от 2,63 до 3,28 °C. Такой диапазон признан приемлемым для исследовательских задач, в частности для сравнительного анализа сценариев кибервоздействий. Дополнительно, эмпирически была проведена оценка логики работы основных блоков системы.

Однако проведённая валидация имела несколько ограничений:

1. Сравнение проводилось только по температуре заготовки.
2. Валидация по данным о давлении и динамике вакуумирования не производилась из-за отсутствия экспериментальных данных.
3. Валидация основных подсистем проведена в большей степени на логическом и имитационном уровне, а не по натурным данным.
4. Валидация модели задержек на натурных данных не проводилась, т.е. результаты эксперимента с моделированием задержек сигналов следует интерпретировать как исследовательские, требующие подтверждения на реальном оборудовании.
5. Все расчёты относились к штатным режимам без искажений.

Указанные ограничения не препятствуют использованию модели для сравнительного сценарного анализа киберфизических эффектов в заданной области применения, но ограничивают прямой перенос количественных значений на реальный автоклав без дополнительной валидации

Верификация программной реализации математической модели в среде Simulink выполнялась с помощью встроенных средств этой системы.

Применимость цифрового двойника ограничена следующими условиями:

1. Модель не учитывает адаптивные атаки, изменяющие параметры в ответ на действия оператора или системы контроля.
2. Редкие отказы оборудования (не связанные с целенаправленными кибервоздействиями) также не воспроизводятся.

3. Точность снижается в нестационарных режимах, где наблюдаются большие градиенты температуры и давления (то есть за пределами калибровочных данных).

Кроме того, необходимо учитывать, что в модели отсутствует двусторонняя связь с реальным объектом, поэтому непосредственное управление оборудованием на основе модели невозможно.

Тем не менее для ряда задач разработанный двойник применим. К ним относятся параметрический анализ влияния кибервоздействий (смещения сигналов, задержки, изменение уставок, модификация коэффициентов регуляторов) на динамику температуры и давления; построение областей скрытых кибервоздействий в пространстве параметров атак; количественная оценка показателей скрытой киберустойчивости; а также сравнение эффективности защитных механизмов в имитационных экспериментах.

В рамках методики цифровой двойник выполняет несколько ключевых функций:

1. Для генерации данных для анализа: для каждого набора параметров кибервоздействия θ выполняется симуляция модели, после чего вычисляются метрики деградации, скрытности и срабатывания систем безопасности).

2. Обеспечение воспроизводимости: все эксперименты проводятся в идентичных начальных и граничных условиях, что позволяет корректно сравнивать различные типы атак и конфигурации системы.

3. Безопасная замена натурных испытаний: моделирование даёт возможность исследовать сценарии, реализация, которых на реальном оборудовании либо опасна, либо затратна.

Следовательно, цифровой двойник служит центральным инструментом, определяющим практическую реализацию предложенной в статье методики.

3.Теоретические основы методики: формальные критерии и метрики

Рассматривается автоматизированная система управления технологическим процессом автоклавной установки, функционирующая как замкнутая киберфизическая система. Физический объект (автоклав) взаимодействует с контуром управления через измерительные и исполнительные каналы.

Кибервоздействие моделируется как вектор параметров $\theta = (\theta_1, \dots, \theta_k) \quad \theta \in \Theta$, изменяющих информационные потоки в системе. Задача состоит в выявлении множества значений θ, при которых одновременно выполняются три условия:

1. Технологический процесс остаётся в пределах, допускаемых системой промышленной безопасности (аварийная защита не срабатывает);

2. Наблюдатель (оператор) не обнаруживает аномалии по доступным ему данным;

3. Происходит измеримая деградация технологического процесса, т.е. фактическое отклонение параметров от эталонной траектории превышает допустимый уровень.

Такое множество обозначается Ω_{stealth} и называется *областью скрытых кибервоздействий*. Построение Ω_{stealth} и количественная оценка связанных с ней показателей и составляют цель рассматриваемой методики.

В настоящем исследовании вектор θ включает следующие компоненты:

- S_T, S_P – аддитивные смещения уставок температуры и давления (set);
- b_T, b_P – смещения сигналов датчиков температуры и давления (bias);

- τ_T, τ_P – временные задержки в каналах измерений (delay);
- $\Delta K_{p,T}, \Delta K_{i,T}, \Delta K_{p,P}, \Delta K_{i,P}$ – отклонения коэффициентов ПИ-регуляторов от номиналов (ctrl);
- X_{HMI} – искажение значений, отображаемых оператору (HMI).

Каждый параметр изменяется в некотором диапазоне Θ_i, а полное пространство атак Θ определяется как декартово произведение допустимых диапазонов отдельных компонент:

$$\Theta = \Theta_{set} \times \Theta_{\text{bias}} \times \Theta_{\text{delay}} \times \Theta_{\text{ctrl}} \times \Theta_{\text{HMI}} \quad (5)$$

Дискретизация Θ позволяет провести вычислительные эксперименты в конечном числе точек.

3.1. Метрика деградации технологического процесса

Введём эталонные (при $\theta = 0$) траектории температуры $T_{\text{nom}}(t)$ и давления $P_{\text{nom}}(t)$, соответствующие штатной работе без атак. При кибервоздействии с параметрами θ система формирует траектории $T_{\text{atk}}(t, \theta)$, $P_{\text{atk}}(t, \theta)$. Для оценки отклонения используется квадратичная мера, проинтегрированная по времени:

$$J(\theta) = \int_0^{t_f} [(T_{\text{atk}}(t, \theta) - T_{\text{nom}}(t))^2 + \alpha(P_{\text{atk}}(t, \theta) - P_{\text{nom}}(t))^2]\, dt, \quad (6)$$

где безразмерный коэффициент $\alpha = (\Delta T_{\text{lim}}/\Delta P_{\text{lim}})^2$ обеспечивает соизмеримость вкладов температуры и давления. Значения $\Delta T_{\text{lim}}, \Delta P_{\text{lim}}$ заданы технологическим регламентом как предельно допустимые отклонения, не приводящие к браку.

Показатель деградации представляет собой интегральную квадратичную меру отклонения технологического процесса от номинального режима, нормированную по допустимым отклонениям температуры и давления и учитывающую длительность существования нарушения, его абсолютные значения используются только для сравнительного анализа сценариев и не имеют самостоятельной физической интерпретации.

В реальных системах АСУ ТП интегральный критерий деградации не используют, оперируют эксплуатационными критериями: уставка, порог, допустимый интервал, таймер выдержки и в этом ключе интегральная метрика деградации имеет малоинформативную интерпретацию для технолога. Поэтому в промышленности используются пороговые критерии, временные окна и логические правила. Кроме того, с точки зрения моделирования, порогово-временной критерий является менее чувствительным к небольшим шумам, медленным колебаниям и погрешностям модели.

Математическое описание порогово-временного критерия:

$$degrad_{time}(\theta) = \left[\int_0^T 1_{|T_{atk}(t) - T_{nom}(t)| > \Delta T_{lim}} dt > \tau_T\right] \bigvee \left[\int_0^T 1_{|P_{atk}(t) - P_{nom}(t)| > \Delta P_{lim}} dt > \tau_P\right] \quad (7)$$

Порогово-временной критерий деградации считается выполненным $degrad_{time}(\theta) = 1$, если суммарное время, в течение которого абсолютное отклонение температуры превышает $\Delta T_{\text{lim}} = 10$, больше $\tau_T = 7200$ с (120 мин), или суммарное время, в течение

которого абсолютное отклонение давления превышает $\Delta P_{\text{lim}} = 0.5$ бар, больше $\tau_P = 3600$ с (60 мин). Отклонения измеряются относительно эталонной траектории (номинального режима). Именно такой критерий и был использован в цифровом двойнике.

Для проверки согласованности этих критериев были проведены численные эксперименты и выполнен ROC-анализ. В качестве положительного класса использовались режимы, классифицированные как деградация согласно порогово-временному критерию. В качестве классификатора использовалось значение интегральной деградации J. Была получена следующая ROC-кривая, площадь под которой составляет AUC= 0.998. Это говорит о том, что критерий интегральной деградации является информативным индикатором ухудшения технологического режима и хорошо согласуется с порогово-временной трактовкой деградации.

3.2. Модель обнаружения наблюдателем (оператором) и критерий скрытности

Обнаружение аномалии оператором моделируется с помощью интегральной функции наблюдаемости $f(t)$. В отличие от классических наблюдателей Люенбергера или фильтров Калмана, предлагаемая функция не требует точной математической модели объекта и менее чувствительна к шумам.

Сначала по эталонному режиму (без атак) вычислялись статистические характеристики:

- σ_T, σ_P – стандартные отклонения сигналов температуры и давления;
- σ_{dT}, σ_{dP} – стандартные отклонения их производных.

Затем для текущего режима определяются две компоненты:

$$f_1(t) = \sqrt{\left(\frac{e_T(t)}{\sigma_T}\right)^2 + \left(\frac{e_P(t)}{\sigma_P}\right)^2} \qquad (8)$$

$$e_T(t) = T_{\text{meas}}(t) - T_{\text{set}}(t) \qquad (9)$$

$$e_P(t) = P_{\text{meas}}(t) - P_{\text{set}}(t) \qquad (10)$$

$$f_2(t) = \left| \frac{dT_{\text{meas}}/dt}{\sigma_{dT}} \right| + d \left| \frac{dP_{\text{meas}}/dt}{\sigma_{dP}} \right|, d = \frac{\sigma_{dT}}{\sigma_{dP}}. \qquad (11)$$

Весовые коэффициенты $w_1 = 1/\sigma_{f1}$, $w_2 = 1/\sigma_{f2}$ (где σ_{fi} – стандартное отклонение соответствующей компоненты на нормальных данных) обеспечивают нормировку. Итоговая функция наблюдаемости:

$$f(t) = w_1 f_1(t) + w_2 f_2(t). \qquad (12)$$

Порог обнаружения ε_{op} выбирается как 99,5%-й квантиль распределения $f(t)$ на эталонном интервале. Выбор 99,5%-го квантиля обусловлен требованием не более 0,5% ложных тревог на эталонном интервале. Решение об обнаружении принимается, если $f(t) > \varepsilon_{\text{op}}$ в течение не менее чем трёх последовательных циклов наблюдения ($N = 3$, длительность цикла 30 с, такое сочетание параметров было получено по результатам нескольких тестовых симуляций и формирует необходимое сочетание чувствительности и отсутствия «ложных срабатываний»). Такое правило имитирует инерционность оператора и снижает вероятность ложных срабатываний. Кибервоздействие считается скрытым (не обнаруженным), если для всех t выполняется условие $f(t) \leq \varepsilon_{\text{op}}$.

Перед основными экспериментами выполнен прогон цифрового двойника в нормальном режиме ($\theta = 0$). По полученным данным вычислены статистические характеристики (значения получены на всём временном интервале эталонного режима, включая переходные процессы): $\sigma_T = 53.53$ K, $\sigma_P = 3.17$ бар, $\sigma_{dT} = 0.0085$ K/с, $\sigma_{dP} = 0.0052$ бар/с. Коэффициент

масштабирования $d = \sigma_{dT}/\sigma_{dP} \approx 1.63$. Порог обнаружения $\varepsilon_{\text{op}} = 7.7723$ (99,5%-й квантиль) обеспечил низкий уровень ложных срабатываний.

3.3 Критерий срабатывания системы безопасности

Система промышленной безопасности Safety переводит установку в аварийный режим при любом из следующих событий:

- давление в автоклаве превышает предельно допустимое (срабатывание предохранительного клапана);
- температура превышает предельно допустимое значение;
- давление в любом из вакуумных контейнеров не достигает порогового значения при активном требовании вакуума.

3.4. Определение области скрытых кибервоздействий

С учётом введённых критериев область Ω_{stealth} определяется следующим образом:

$$\Omega_{\text{stealth}} = \{\theta \in \Theta \mid \text{safety}(\theta) = 0, \text{detected}(\theta) = 0, degrad_{time}(\theta) = 1\} \quad (13)$$

Иными словами, это множество параметров атаки, при которых:

- система не переходит в аварийный режим;
- оператор не фиксирует аномалию;
- интегральная деградация превышает критический порог.

3.5. Показатели киберустойчивости к скрытым кибервоздействиям

На основе Ω_{stealth} вводятся следующие количественные характеристики:

Доля скрытых атак – геометрическая мера уязвимости:

$$\rho_{\text{stealth}} = \frac{|\ \Omega_{\text{stealth}}\ |}{|\ \Theta\ |} \quad (14)$$

где $|\cdot|$ обозначает количество дискретных точек после дискретизации пространства параметров.

Средний нормированный ущерб - рассчитывается только по точкам, принадлежащим Ω_{stealth}:

$$\bar{J}_{\text{stealth}} = \frac{1}{|\ \Omega_{\text{stealth}}\ |} \sum_{\theta \in \Omega_{\text{stealth}}} \min\left(\frac{J(\theta)}{J_{\text{ref}}},\ 1\right) \quad (15)$$

где $J_{\text{ref}} = Q_{0.95}(J)$ – 95-й процентиль распределения $J(\theta)$ по всем точкам Θ (включая обнаруживаемые атаки). Нормировка к процентилю, а не к максимуму, уменьшает влияние единичных выбросов.

Интегральный риск скрытых атак определяется как произведение доли и среднего ущерба:

$$\tilde{\rho} = \rho_{\text{stealth}} \cdot \bar{J}_{\text{stealth}} \quad (16)$$

Киберустойчивость к скрытым кибервоздействиям – дополнительная величина:

$$C_{\text{stealth}} = 1 - R_{\text{stealth}} \quad (17)$$

Чем ближе C_{stealth} к единице, тем менее опасны скрытые воздействия.

Запас киберустойчивости к скрытым воздействиям для двухпараметрического случая (например, смещения уставок S_T, S_P) вычисляется как минимальное нормированное расстояние от нулевой точки (отсутствие атаки) до границы Ω_{stealth}:

$$S_{\text{stealth}} = \min_{(S_T,S_P)\in\partial\Omega_{\text{stealth}}} \sqrt{\left(\frac{S_T}{\Delta T_{\text{lim}}}\right)^2 + \left(\frac{S_P}{\Delta P_{\text{lim}}}\right)^2} \quad (18)$$

Малые значения S_{stealth} указывают на то, что даже незначительное воздействие может перевести систему в режим скрытой деградации.

3.6. Алгоритм построения Ω_{stealth}

Построение области выполняется в следующей последовательности:

1. Задание пространства параметров Θ и его дискретизация (формирование сетки или выборки).
2. Для каждой точки θ_i запуск симуляции цифрового двойника, регистрация временных рядов $T_{\text{atk}}(t), P_{\text{atk}}(t)$.
3. Вычисление интегральной деградации $J(\theta_i)$ по формуле (5).
4. Проверка срабатывания систем безопасности (флаг $\text{safety}(\theta_i)$).
5. Расчёт функции наблюдаемости $f(t)$ и принятие решения об обнаружении.
6. Классификация точки по трём признакам (безопасность, обнаружение, деградация).
7. Формирование множества Ω_{stealth} как объединения точек, удовлетворяющих условиям.
8. Вычисление показателей $\rho_{\text{stealth}}, \bar{J}_{\text{stealth}}, R_{\text{stealth}}, C_{\text{stealth}}, S_{\text{stealth}}$.

Алгоритм был реализован в среде Matlab с использованием скриптов, автоматизирующих циклический перебор параметров и сбор результатов.

4. Экспериментальное исследование и построение Ω_{stealth}

4.1. Классы рассмотренных кибервоздействий

Для оценки применимости методики были рассмотрены следующие классы кибервоздействий:

Таблица 1 - Виды кибервоздействий и результаты экспериментов

Кибервоздействие (атака)	Количество точек (симуляций), ед.	Максимальная деградация, J_{max}, ед.	Доля $\Omega_{stealth}$ $\rho_{stealth}$	Средний нормированный ущерб, $\bar{J}_{\text{stealth}}$	Интегральный риск скрытых атак, $\tilde{\rho}$	Кибер устойчивость к скрытым кибервоздействиям	Запас к киберустойчивости к скрытым воздействиям, $S_{stealth}$
Атака на управляющие сигналы $\theta = (S_T, S_P, 0,0,0,0,0,0,0)$	1 425	8.8×10^7	0.1518	0.2276	0.0346	0.9654	1.0000
Атака на сигналы измерений (смещение)	5 868	8.8×10^6	0.0125	0.0334	0.0004	0.9996	25.428

Кибервоздействие (атака)	Количество точек (симуляций), ед.	Максимальная деградация, J_{max}, ед.	Доля $\Omega_{stealth}$ $\rho_{stealth}$	Средний нормированный ущерб, $\bar{J}_{\text{stealth}}$	Интегральный риск скрытых атак, $\tilde{\rho}$	Киберустойчивость к скрытым кибервоздействиям	Запас киберустойчивости к скрытым воздействиям, $S_{stealth}$
$\theta = (0{,}0, b_T, b_P, 0{,}0{,}0{,}0{,}0)$							
Атака на сигналы измерений (задержка) $\theta = (0{,}0{,}0{,}0, \tau_T, \tau_P, 0{,}0{,}0)$	1 476	-	-	-	-	-	-
Атака на регулятор $\theta = (0{,}0{,}0{,}0{,}0{,}0, \Delta K_T, \Delta K_P, 0$	14 641	3.3×10^{7}	0.0408	0.01	0.00041	0.9996	0.0004
Комбинированные сценарии (Управляющие сигналы + HMI) $\theta = (S_T, S_P, 0{,}0{,}0{,}0{,}0{,}0{,}0, X_{\text{HM}}$	1425	2.19×10^{8}	0.7733	0.4348	0.3362	0.6638	1.0000

4.2. Поверхности деградации для различных видов кибервоздействий

Атака на управляющие сигналы

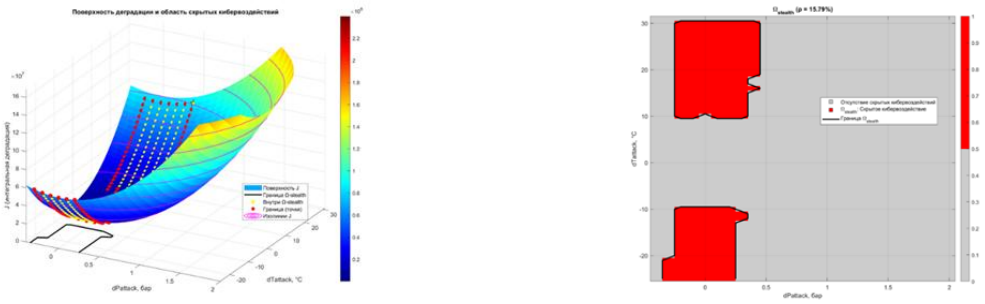


Рисунок 2 - Поверхность деградации и область киберустойчивости к скрытым кибервоздействиям для атаки на управляющие сигналы

Атака на сигналы измерений (смещение)

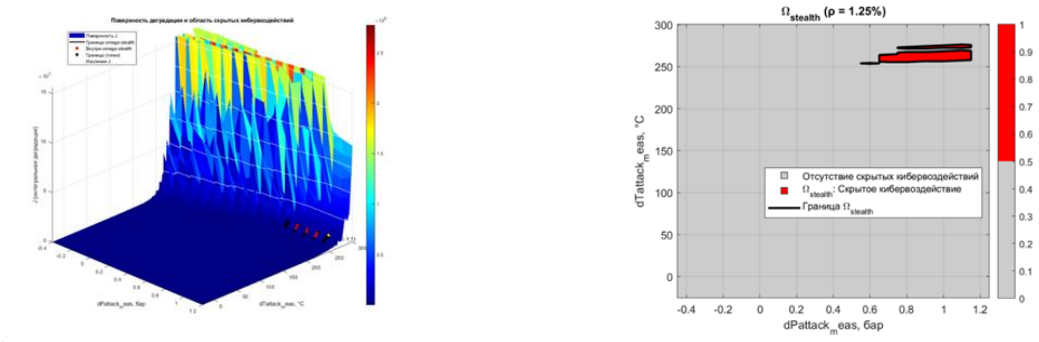


Рисунок 3 - Поверхность деградации и область киберустойчивости к скрытым кибервоздействиям для атаки на сигналы измерений

Комбинированные сценарии (Управляющие сигналы + HMI)

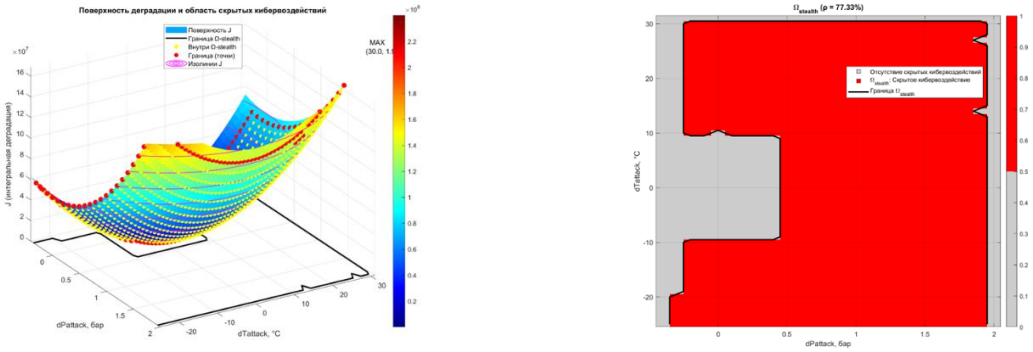


Рисунок 4 - Поверхность деградации и область киберустойчивости к скрытым кибервоздействиям для комбинированного сценария атаки

Реализация кибервоздействия в виде атаки на датчики, приводящей к задержке информационного сигнала, показала, что такие воздействия в рамках рассматриваемой модели приводят к некорректной и неустойчивой работе ПИ-регуляторов и срабатыванию системы промышленной безопасности Safety. Возможность его реализации при атаках на реальном автоклаве требует отдельного изучения. Ниже приведены графики колебаний температуры и давления при реализации задержек, красным показаны значения, превышающие допустимый порог и вызывающие срабатывания системы промышленной безопасности:

Температура (a) Давление (b)

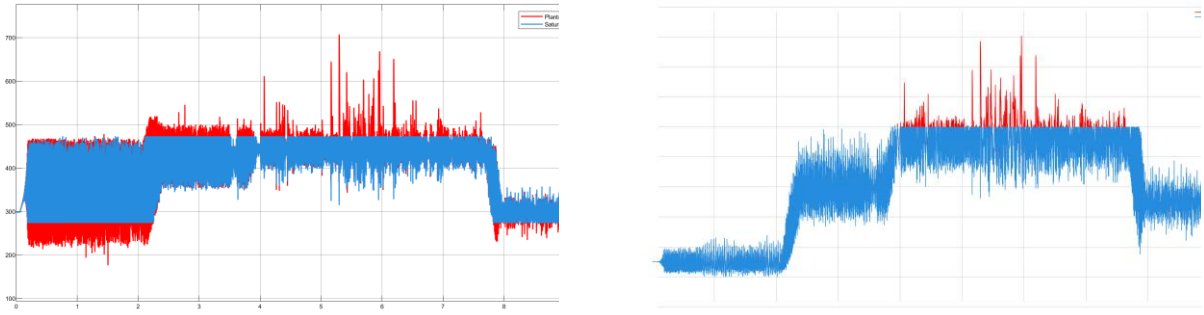


Рисунок 5 - Температура (a) и давления (b) в измерительном канале при реализации кибервоздействия «задержка сигнала в измерительном канале»

Кибервоздействие в случае атаки на ПИ регулятор представляет собой следующий вектор:

$$\theta = (0, 0, 0, 0, 0, 0, dK_p, dK_t, 0)$$

В свою очередь каждый из коэффициентов регуляторов представляет собой совокупность двух коэффициентов пропорционального и интегрального, для температуры это соответственно (Kp_T, Ki_T), а для давления (Kp_P, Ki_P). Набор комбинаций кибервоздействий на коэффициенты (Kp_T, Ki_T, Kp_P, Ki_P) вызывающих деградацию и не детектируемых оператором образуют пространство скрытых кибервоздействий. В результате поиска и визуализации найденных параметров коэффициентов получены следующие наборы значений кибервоздействий (рис. 6): каждая линия – одна симуляция; красными линиями соединены значения коэффициентов соответствующие скрытым кибервоздействиям; серые линии – симуляция выявила срабатывание защиты или были детектированы кибервоздействия.

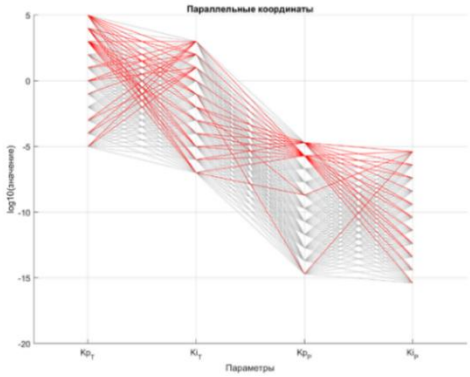


Рисунок 6 - Пространство параметров ПИ-регуляторов, соответствующих скрытым кибервоздействиям

4.3. Оценка эффективности защитной меры

В качестве примера применения методики для выбора защитных мер модельно интерпретировано требование контроля вводимых данных (ОЦЛ.4 Приказа ФСТЭК России №31) [6], а именно в цифровой двойник добавлен независимый контур «Инспектор», который сравнивает текущие параметры (давление и температура) с эталонной траекторией, хранящейся в защищённой базе.

Эксперимент проведён для наиболее опасного сценария – комбинированной атаки:

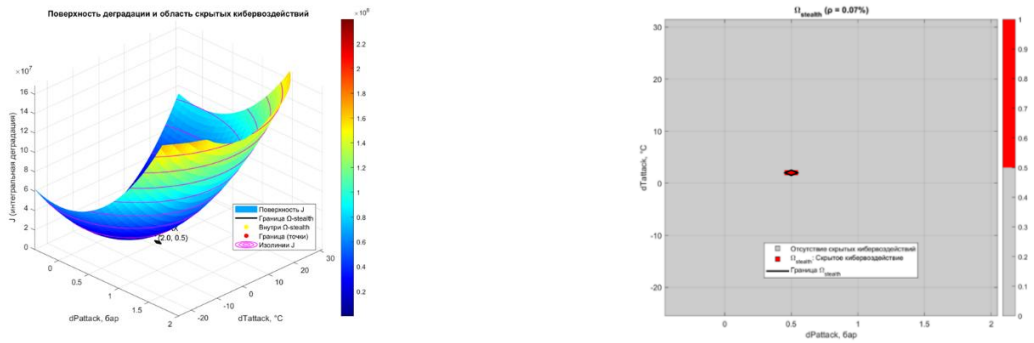


Рисунок 7 - Поверхность деградации и область киберустойчивости к скрытым кибервоздействиям для комбинированного сценария атаки при применении защитной меры

Результаты сравнивались с базовым вариантом (без инспектора). Область Ω_{stealth} сократилась с 1 102 точек до 1 точки ($\rho_{\text{stealth}} = 0.0007$) – или на 99,93%. $\bar{J}_{\text{stealth}}$ снизился до 0.0578. Максимальная деградация снизилась с 2.19×10^{8} до 9.74×10^{6} (уменьшение на 95,6%). Интегральный риск $\tilde{\rho}$ в рамках рассматриваемой модели существенно снизился (стал близо к 0), скрытая киберустойчивость повысилась до $C_{\text{stealth}} = 0{,}99996$.

Данные количественные оценки показывают, что независимый контур контроля технологических данных (даже в простейшей реализации) значительно повышает киберустойчивость системы к скрытым комбинированным атакам. При этом сам контур не влияет на работу основного управления и может быть реализован на отдельном аппаратном обеспечении.

4.4. Интерпретация результатов

Вычислительные эксперименты подтвердили существование области Ω_{stealth} для ряда типов кибервоздействий. Предложенные показатели $\rho_{\text{stealth}}, \bar{J}_{\text{stealth}}, \tilde{\rho}, C_{\text{stealth}}, S_{\text{stealth}}$ позволяют количественно сравнивать уязвимость системы к разным классам атак и оценивать эффективность защитных мер. Использование цифрового двойника делает такой анализ возможным без риска для реального оборудования и с полной воспроизводимостью условий эксперимента.

Полученные результаты легли в основу методики, которая может быть применена для других термобарических установок с замкнутыми контурами регулирования. Для этого достаточно адаптировать параметры модели и пороговые значения под конкретное оборудование и технологический регламент.

5.Обсуждение и ограничения предложенного метода

Предложенная методика построения области скрытых кибервоздействий и оценки показателей киберустойчивости была апробирована на модели автоклавной установки с двумя замкнутыми контурами регулирования (температура и давление). Архитектура АСУ ТП: промышленный компьютер, передающий уставки по Profibus на ПЛК Siemens S7-300, обратные связи от датчиков к регулятору, дискретные исполнительные механизмы – является типовой для многих термобарических, химико-технологических и энергетических объектов. Методика может быть распространена на системы с иным числом контуров (расход, уровень, концентрация) и другими типами регуляторов (ПИД, каскадные схемы) при условии адаптации математической модели и критериев деградации.

Вместе с тем существуют ограничения, связанные как с принятыми допущениями при построении цифрового двойника, так и с объёмом проведённых вычислительных экспериментов:

- Уравнение состояния идеального газа. В рабочем диапазоне давлений автоклавной установки отклонения от идеальности не превышают 2–3%, что для исследовательских задач приемлемо. Однако для более высоких давлений или при конденсации пара модель потребовала бы уточнения;
- Линейная аппроксимация теплопотерь. Коэффициент k_{loss} принят постоянным, хотя в реальности теплопередача зависит от температуры и режима конвекции;

- Как указано в разделе 2, прямая валидация выполнена только для температуры заготовки (RMSE 2,6–3,3°C). Следующие компоненты модели не сопоставлялись с натурными данными (динамика давления в камере, расходы через клапаны и инерционность пневмосистемы, температурные поля внутри автоклава)
- Модель оператора в виде функции $f(t)$ и порога ε_{op}, определённого как 99,5%-й квантиль, не учитывает некоторые аспекты реального поведения человека-оператора: утомляемость и снижение внимания при длительной работе, возможность переключения между несколькими технологическими установками, влияние визуального интерфейса (цветовая индикация, тренды, анимация) на скорость обнаружения. Предложенная модель является детерминированной и инерционной (накопление трёх циклов). Она даёт консервативную оценку скрытности, т.е. в реальности оператор может обнаружить аномалию чуть позже или чуть раньше в зависимости от индивидуальных особенностей и эргономики рабочего места. Для целей сравнительного анализа различных типов атак и защитных мер такая степень детализации достаточна;
- Во всех экспериментах параметры θ (смещения, задержки, изменения коэффициентов, HMI) оставались постоянными во времени. В реальных условиях атакующий может адаптировать своё воздействие – например, постепенно наращивать смещение датчика, чтобы избежать превышения порога обнаружения, или временно отключать атаку при появлении обслуживающего персонала.
- Для двумерных пространств параметров (например, S_T, S_P) использовалась равномерная сетка с шагом, обеспечивающим приемлемую детализацию. Однако граница Ω_{stealth} может иметь сложную форму, и при недостаточно мелком шаге возможно пропускание узких «коридоров» скрытых атак;
- При оценке эффективности защитной меры (глава 4.4) предполагалось, что независимый контур контроля «Инспектор» сохраняет свою целостность и не может быть скомпрометирован атакующим. Это сильное допущение, используемое для демонстрации потенциала методики, а не для утверждения о практической неуязвимости такого контура.

Заключение

Основные результаты исследования, имеющие научную новизну:

- Введено формальное определение скрытого кибервоздействия для технологических систем с замкнутыми контурами управления через выполнения трёх условий: отсутствие срабатывания промышленной безопасности, необнаружимость оператором и измеримая деградация технологического процесса;
- Предложена интегральная метрика деградации $J(\theta) = \int [(\Delta T)^2 + \alpha(\Delta P)^2]dt$, которая нормирована по допустимым отклонениям температуры и давления. Рассмотрено и показано согласие этого критерия с порогово-временным критерием деградации, используемом в реальном АСУ ТП при значении $J_{\text{crit}}^{\text{opt}} = 13\,505\,650$, (AUC = 0.998);

- Предложена модель наблюдателя (оператора) на основе комбинации регуляторной ошибки и аномалии производных, с порогом обнаружения, задаваемым через 99,5%-й квантиль эталонного распределения. Модель не требует точного математического описания объекта и менее чувствительна к шумам;
- Описана методика построения области Ω_{stealth} как объекта количественного анализа;
- Введены количественные показатели к киберустойчивости: доля скрытых атак ρ_{stealth}, средний нормированный ущерб $\bar{J}_{\text{stealth}}$, интегральный риск $\tilde{\rho}$, запас устойчивости S_{stealth}. Эти показатели позволяют сравнивать уязвимость системы к различным классам атак и оценивать эффективность защитных мер;
- В качестве демонстрации предложенной методики построена область Ω_{stealth} для пяти типов кибервоздействий. Показано, что наибольшую опасность представляют комбинированные атаки, одновременно подменяющие управляющие сигналы и данные на HMI;
- На примере независимого контура контроля технологических данных (модельная интерпретация меры ОЦЛ.4 Приказа ФСТЭК №31) продемонстрирована возможность применения методики для оценки защитных мер. Проведена оценка эффективности защитной меры.

Предложенная методика формирует подход к количественным методам анализа киберустойчивости, пригодным для инженерного применения. Использование цифрового двойника в качестве основного инструмента позволяет преодолеть ключевое препятствие – невозможность проведения натурных экспериментов с кибервоздействиями на реальном оборудовании. Предложенные метрики и алгоритм построения Ω_{stealth} создают основу для перехода к стандартизованной оценке киберустойчивости АСУ ТП.

Список литературы

1. Кляйнман А., Вул А. Скрытые атаки на SCADA-системы // Материалы семинара ACM по безопасности киберфизических систем. — 2017. //arXiv: 1706.09303v1 [cs.CR] 28 июня 2017 г.
2. NIST SP 800-160, том 2. 2: Разработка киберустойчивых систем (2021)
3. ГОСТ Р 57700.37-2021. Цифровые двойники изделий. Общие положения.
4. Все, что вам нужно знать о стандарте ISA/IEC 62443. Payatu Consulting Pvt. Ltd. 2021
5. IEC 62443-2-1 Промышленные коммуникационные сети – Безопасность сетей и систем – Часть 2-1: Создание программы безопасности промышленной автоматизации и систем управления Редакция 1.0 2010-11
6. Приказ ФСТЭК № 31 от 14 марта 2014г. «Об утверждении требований к обеспечению защиты информации в автоматизированных системах управления производственными и технологическими процессами на критически важных объектах, потенциально опасных объектах, а также на объектах, представляющих повышенную опасность для жизни и здоровья людей и для окружающей природной среды» (с изменениями и дополнениями от 23 марта 2017 г., 9 августа 2018 г., 15 марта 2021 г.)

7. Кравчик М., Шабтай А. Обнаружение кибератак в промышленных системах управления с использованием сверточных нейронных сетей // Материалы семинара ACM 2018 по безопасности киберфизических систем.
8. Ограничение воздействия скрытых атак на систему промышленного контроля, Урбиа Д., Хиральдо Дж., Карденас А.А. Типпенхауэр, 2016 https://csrc.nist.gov/pubs/conference/2016/10/24 / ограничение воздействия скрытых атак на информационные системы/окончательный вариант
9. Арнстром Д., Тейшейра А., Управляемая данными и скрытая деактивация фильтров безопасности //arXiv ID: 2412.013461v1 [eess.SY] 2 декабря 2024
10. Арнстром Д., Тейшейра А., Тайное отключение защитных фильтров; //arXiv: 2403.17861v1 [eess.SY] 26 марта 2024 г.
11. Наута Т., Сандберг Х., Маджио М. Скрытые атаки с задержкой вычислений на системы управления // ICCPS '25: Материалы 16-й Международной конференции ACM/IEEE по киберфизическим системам (совместно с CPS-IoT Week 2025), // URL: https://doi.org/10.1145/3716550.3722013

Reference

1. 1. Kleinmann A., Wool A. Stealthy deception attacks against SCADA systems // Proceedings of the ACM Workshop on Cyber-Physical Systems Security. — 2017. //arXiv:1706.09303v1 [cs.CR] 28 Jun 2017
2. NIST SP 800-160 Vol. 2: Developing Cyber-Resilient Systems (2021)
3. ГОСТ Р 57700.37-2021. Цифровые двойники изделий. Общие положения.
4. All you need to know about ISA/IEC 62443 standard. Payatu Consulting Pvt. Ltd. 2021
5. IEC 62443-2-1 Industrial communication networks – Network and system security – Part 2-1: Establishing an industrial automation and control system security program Edition 1.0 2010-11
6. Приказ ФСТЭК № 31 от 14 марта 2014г. «Об утверждении требований к обеспечению защиты информации в автоматизированных системах управления производственными и технологическими процессами на критически важных объектах, потенциально опасных объектах, а также на объектах, представляющих повышенную опасность для жизни и здоровья людей и для окружающей природной среды» (с изменениями и дополнениями от 23 марта 2017 г., 9 августа 2018 г., 15 марта 2021 г.)
7. Kravchik M., Shabtai A. Detecting Cyber Attacks in Industrial Control Systems Using Convolutional Neural Networks // Proceedings of the 2018 ACM Workshop on Cyber-Physical Systems Security.
8. Limiting the Impact of Stealthy Attacks on Industrial Control System, Urbia D., Giraldo J., Cardenas A.A. Tippenhauer, 2016 https://csrc.nist.gov/pubs/conference/2016/10/24/ limiting-the-impact-of-stealthy-attacks-on-ics/final
9. Arnstrom D., Teixeira A., Data-Driven and Stealthy Deactivation of Safety Filters //arXiv ID:2412.013461v1 [eess.SY] 2 Dec 2024
10. Arnstrom D., Teixeira A., Stealthy Deactivation of Safety Filters; //arXiv:2403.17861v1 [eess.SY] 26 Mar 2024

11. Nauta T, Sandberg H, Maggio M. Stealthy Computational Delay Attacks on Control Systems // ICCPS '25: Proceedings of the ACM/IEEE 16th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2025), // URL: https://doi.org/10.1145/3716550.3722013

Денисюк А.А. Сравнительный анализ эффективности реализаций алгоритмов обработки изображений на CPU и GPU// Международный журнал информационных технологий и энергоэффективности. – 2026. –Т. 11 № 7(69) Ч.1 с. 21–29

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.932

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ РЕАЛИЗАЦИЙ АЛГОРИТМОВ ОБРАБОТКИ ИЗОБРАЖЕНИЙ НА CPU И GPU

Денисюк А.А.
ФГАОУ ВО "НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ "МОСКОВСКИЙ ИНСТИТУТ ЭЛЕКТРОННОЙ ТЕХНИКИ", Москва, Россия, (124498, город Москва, город Зеленоград, пл. Шокина, д. 1), e-mail: sanya.denisyuk.26@mail.ru

В статье рассматривается задача повышения производительности классических алгоритмов цифровой обработки изображений за счёт использования различных вычислительных подходов. Выполнено сравнение последовательных CPU реализаций, многопоточных OpenMP реализаций и реализаций на графическом процессоре с использованием платформы CUDA. В качестве исследуемых алгоритмов выбраны размытие по Гауссу, медианный фильтр, оператор Собеля и быстрое преобразование Фурье. Показано, что эффективность переноса вычислений на GPU определяется не только степенью параллелизма алгоритма, но и вычислительной сложностью одной операции, размером изображения, организацией доступа к памяти и затратами на передачу данных между оперативной и видеопамятью. Установлено, что CUDA наиболее целесообразна для ресурсоёмких операций и изображений большого размера, OpenMP является эффективным компромиссным решением для многоядерных CPU, а для стандартных преобразований Фурье предпочтительно применение специализированной библиотеки cuFFT.

Ключевые слова: Цифровая обработка изображений, CPU, GPU, размытие по Гауссу, медианный фильтр, оператор Собеля, быстрое преобразование Фурье.

COMPARATIVE ANALYSIS OF IMAGE PROCESSING ALGORITHM IMPLEMENTATIONS ON CPU AND GPU

Denisyuk A.A.
"NATIONAL RESEARCH UNIVERSITY "MOSCOW INSTITUTE OF ELECTRONIC TECHNOLOGY", Moscow, Russia, (124498, Moscow, Zelenograd, Shokina Square, 1), e-mail: sanya.denisyuk.26@mail.ru

The article discusses the problem of improving the performance of classical digital image processing algorithms through the use of various computational approaches. A comparison of sequential CPU implementations, multithreaded OpenMP implementations, and GPU implementations using the CUDA platform has been performed. Gaussian blur, median filter, Sobel operator, and fast Fourier transform are selected as the studied algorithms. It is shown that the efficiency of transferring calculations to the GPU is determined not only by the degree of parallelism of the algorithm, but also by the computational complexity of a single operation, the size of the image, the organization of memory access, and the cost of data transfer between RAM and video memory. It has been found that CUDA is most suitable for resource-intensive operations and large-size images, OpenMP is an effective compromise solution for multi-core CPUs, and the use of a specialized cuFFT library is preferable for standard Fourier transforms.

Keywords: Digital image processing, CPU, GPU, Gaussian blur, median filter, Sobel operator, fast Fourier transform..

Современные системы цифровой обработки изображений используются в техническом зрении, медицине, промышленном контроле, спутниковом мониторинге и других областях, где требуется обработка больших объёмов визуальной информации. Рост разрешения

изображений приводит к увеличению количества пикселей, а следовательно, и числа однотипных операций, которые необходимо выполнить для фильтрации, выделения границ или частотного анализа. В этих условиях последовательная реализация даже классических алгоритмов может становиться ограничивающим фактором при создании прикладных систем реального времени [1].

Актуальным направлением повышения производительности является использование параллельных вычислений. Центральные процессоры обеспечивают высокую производительность одного потока и хорошо подходят для универсальных задач, однако имеют ограниченное число вычислительных ядер. Графические процессоры, напротив, ориентированы на массовое параллельное выполнение однотипных операций и обладают высокой пропускной способностью, что делает их перспективными для задач обработки изображений. При этом простой перенос алгоритма на GPU не гарантирует максимального ускорения: итоговый результат зависит от структуры вычислений, регулярности доступа к данным, объёма обмена между host и device, а также от возможности объединения нескольких этапов обработки в единый GPU конвейер [2, 3].

Целью исследования является сравнительный анализ эффективности последовательных, многопоточных и GPU реализаций классических алгоритмов обработки изображений. Для экспериментов были выбраны четыре алгоритма, представляющие различные классы вычислительных задач: размытие по Гауссу, медианный фильтр, оператор Собеля и быстрое преобразование Фурье. Для каждого алгоритма были разработаны реализации на C++ для CPU, многопоточные варианты с использованием OpenMP и CUDA версии для выполнения на GPU. Для БПФ дополнительно рассматривалась библиотечная реализация cuFFT.

Материалы и методы исследования

Размытие по Гауссу относится к линейным фильтрам и применяется для сглаживания изображения, подавления высокочастотных шумов и предварительной подготовки данных к дальнейшему анализу. Благодаря сепарабельности гауссовой функции двумерная свёртка может быть заменена двумя одномерными проходами по строкам и столбцам, что существенно снижает вычислительную сложность алгоритма. Такая структура хорошо подходит как для OpenMP, так и для CUDA, поскольку обработка пикселей может быть распределена между независимыми потоками [4].

Медианный фильтр является нелинейным методом фильтрации, эффективным при подавлении импульсного шума типа «соль и перец». В отличие от линейной свёртки, он требует выбора медианного значения в локальном окне, что делает операцию более трудоёмкой. При увеличении размера окна возрастает объём локальной сортировки или сравнения элементов, поэтому данный алгоритм особенно показателен при оценке эффективности GPU для вычислительно насыщенных операций [5].

Оператор Собеля используется для выделения границ и основан на вычислении приближённых производных яркости изображения по горизонтальному и вертикальному направлениям. Алгоритм имеет локальный характер и хорошо распараллеливается, однако вычислительная сложность на один пиксель сравнительно невелика. По этой причине при оценке CUDA реализации особенно важно учитывать не только время выполнения ядра, но и полное время с передачей данных между CPU и GPU [6].

Быстрое преобразование Фурье представляет собой алгоритм эффективного вычисления дискретного преобразования Фурье. Для двумерного изображения БПФ выполняется последовательно по строкам и столбцам. В рамках исследования была реализована собственная CUDA версия БПФ, а также проведено сравнение с cuFFT, поскольку специализированные библиотеки NVIDIA используют глубоко оптимизированные схемы планирования вычислений и доступа к памяти [7,8].

Время выполнения последовательных и OpenMP реализаций измерялось средствами стандартной библиотеки C++, а именно модулем std::chrono. Для GPU реализаций отдельно рассматривалось время выполнения вычислительного ядра и полное время, включающее копирование данных между оперативной памятью и памятью GPU, для измерения времени использовались CUDA events. Такой подход позволяет оценить как чистую вычислительную эффективность графического процессора, так и практическую применимость реализации в реальной программе.

Вычислительные эксперименты проводились на ноутбуке, оснащённом центральным процессором Intel Core i7-13700HX и графическим процессором NVIDIA GeForce RTX 4050 Laptop GPU. В качестве тестовых данных использовался набор изображений различных размеров: 256×256, 512×512, 1024×1024, 2048×2048, 4096×4096 и 8192×8192. Для исключения влияния содержательной структуры изображения на результаты сравнения использовалось одно и то же полутоновое изображение Lenna, масштабированное до требуемых размеров.

Результаты экспериментов

На Рисунке 1 представлены результаты измерения времени выполнения размытия по Гауссу при σ = 4.

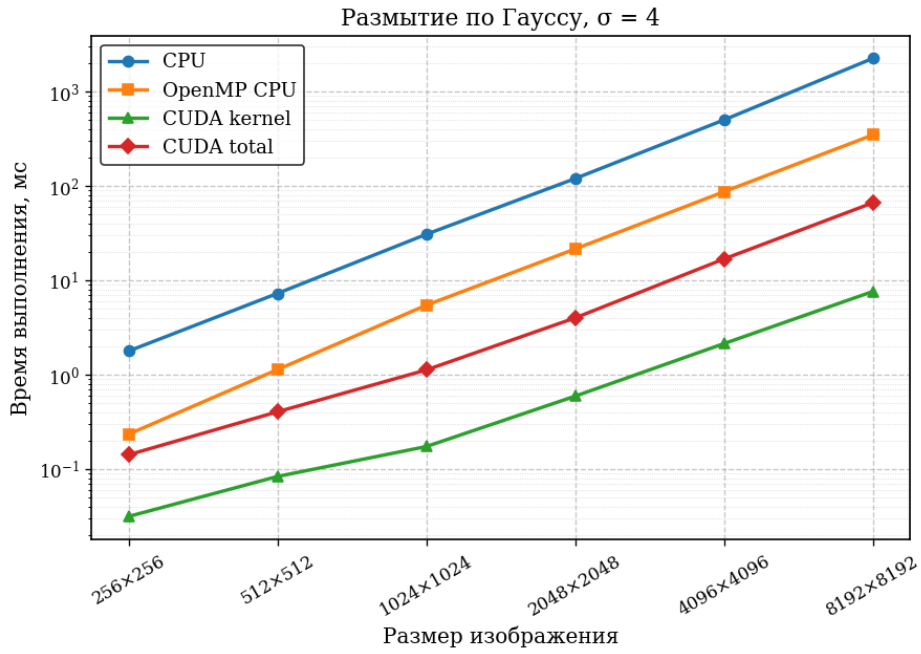


Рисунок 1 – Время выполнения размытия по Гауссу при σ = 2

Для всех реализаций наблюдается закономерный рост времени при увеличении разрешения изображения. Последовательная CPU реализация характеризуется наибольшим временем выполнения, поскольку вся обработка выполняется одним потоком. Использование

OpenMP даёт устойчивое ускорение благодаря распределению вычислений между ядрами процессора. CUDA реализация показывает минимальное время вычислительного ядра и сохраняет значительное преимущество даже при учёте передачи данных между host и device.

Полученные результаты показывают, что для размытия по Гауссу применение GPU наиболее оправдано при обработке изображений среднего и большого размера. При малых изображениях накладные расходы на передачу данных снижают итоговое преимущество CUDA, однако по мере роста разрешения вычислительная нагрузка увеличивается, и возможности массового параллелизма GPU используются эффективнее. При увеличении σ преимущество CUDA становится ещё более выраженным, так как растёт количество операций, приходящихся на каждый пиксель.

На Рисунке 2 представлены результаты измерения времени выполнения медианного фильтра при размере окна 3×3.

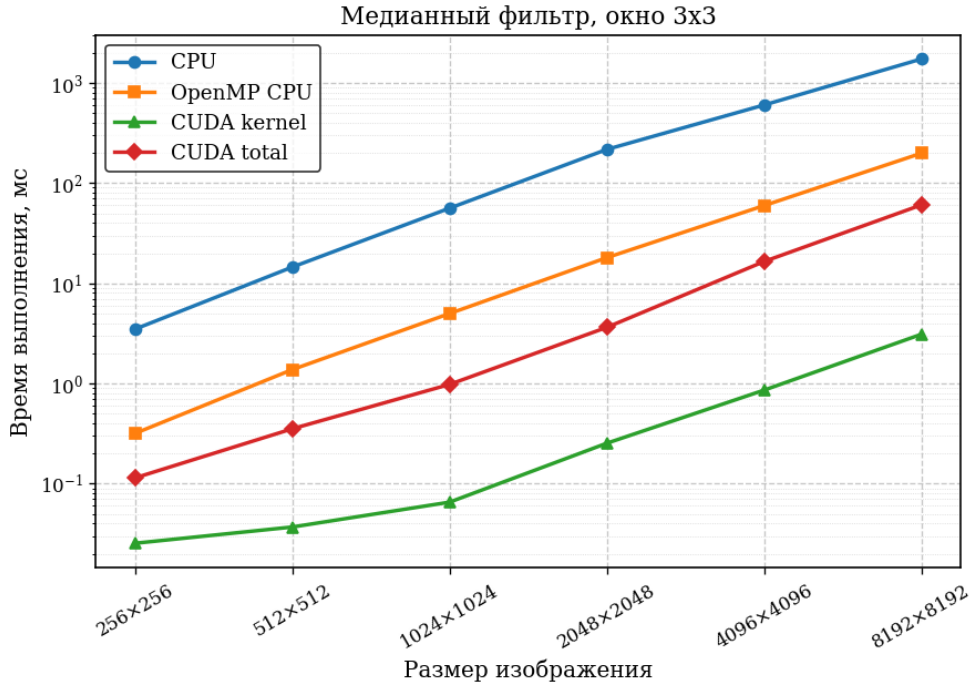


Рисунок 2 – Время выполнения медианного фильтра при размере окна 3×3

Для медианного фильтра характер масштабирования отличается от линейных фильтров. Операция выбора медианы в локальном окне требует большего числа сравнений, поэтому последовательная CPU реализация, в сравнении с размытием по Гауссу, показывает значительно большее отставание по времени выполнения от остальных реализаций. OpenMP обеспечивает стабильное ускорение, однако CUDA реализация демонстрирует наилучшие результаты, особенно для окна 5×5. Это объясняется тем, что большая вычислительная нагрузка на пиксель лучше компенсирует затраты на запуск ядра и передачу данных.

На Рисунке 3 представлены результаты измерения времени выполнения оператора Собеля.

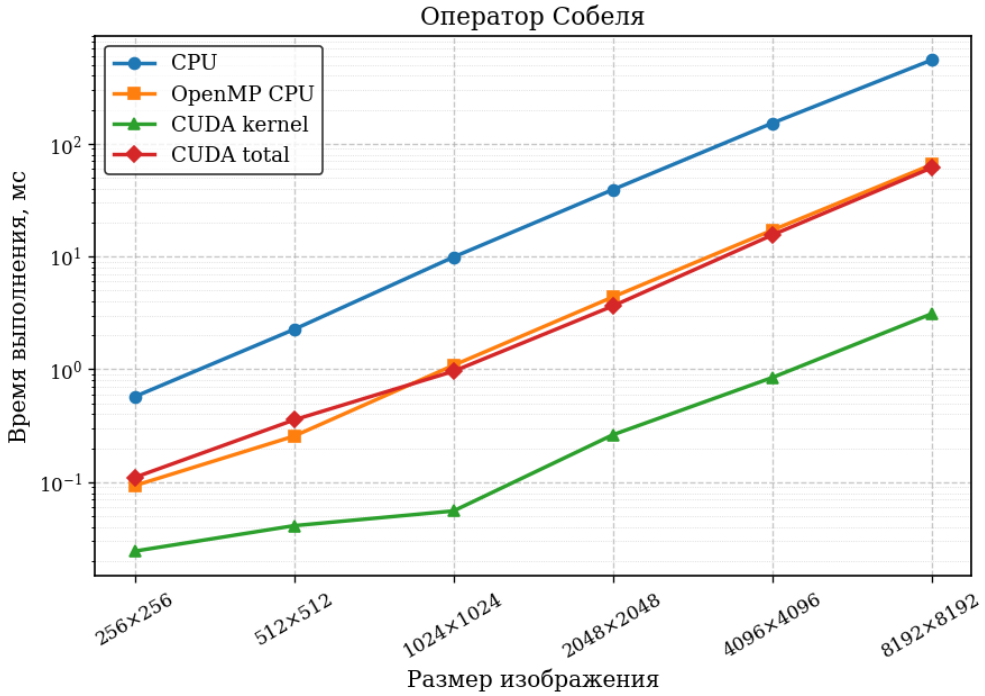


Рисунок 3 – Время выполнения оператора Собеля

Для оператора Собеля наблюдается иная ситуация. Время выполнения CUDA ядра значительно меньше времени последовательной и OpenMP реализаций, но полное CUDA время оказывается сопоставимым с OpenMP из-за малой вычислительной сложности операции на один пиксель. Использование OpenMP является более рациональным решением, поскольку не требует копирования данных в память GPU. Использование CUDA для оператора Собеля становится оправданным только при обработке в составе длинного GPU конвейера, где изображение уже находится в видеопамяти.

На Рисунке 4 представлены результаты измерения времени выполнения быстрого преобразования Фурье.

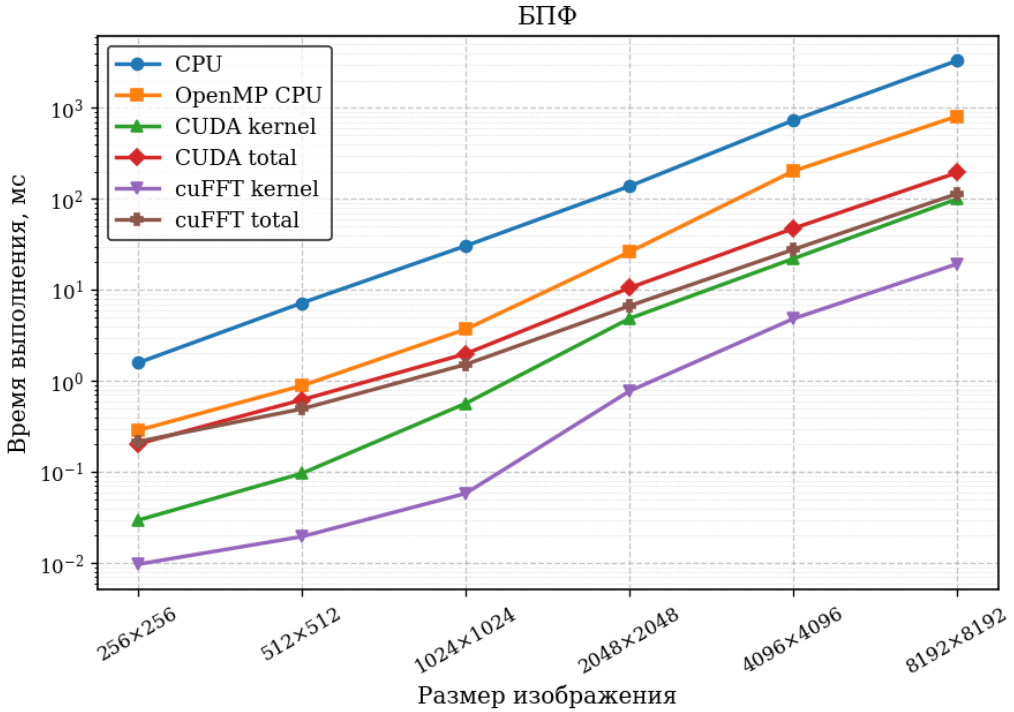


Рисунок 4 – Время выполнения быстрого преобразования Фурье

Последовательная CPU реализация показывает наибольшее время во всём диапазоне размеров изображений. OpenMP существенно ускоряет вычисления, однако при росте размера изображения преимущество GPU становится более заметным. Собственная CUDA реализация быстрее последовательного и OpenMP вариантов даже с учётом передачи данных, но наилучшие результаты показывает cuFFT, что подтверждает эффективность специализированных библиотек для стандартных вычислительных операций. Библиотечное решение имеет существенное преимущество за счёт более глубокой оптимизации. При этом собственная CUDA реализация сохраняет практическую ценность в тех случаях, когда требуется модификация алгоритма, интеграция нестандартных этапов обработки или контроль над внутренней структурой вычислений.

Обобщение рациональных областей применения различных вычислительных подходов приведено в Таблице 1.

Таблица 1 – Рациональные области применения вычислительных подходов

Алгоритм	Рациональный подход	Обоснование
Размытие по Гауссу	CUDA; OpenMP может использоваться для малых изображений	Алгоритм хорошо распараллеливается, а при росте σ увеличивается объём вычислений на пиксель.
Медианный фильтр	CUDA	Высокая вычислительная сложность локальной обработки позволяет эффективно использовать массовый параллелизм GPU.
Оператор Собеля	OpenMP; CUDA только в составе GPU конвейера	Низкая сложность операции делает передачу данных существенным фактором полного времени выполнения.
Быстрое преобразование Фурье	cuFFT; собственная CUDA при необходимости модификации	Специализированная библиотека использует глубокие оптимизации архитектуры GPU.

Обсуждение результатов

Анализ результатов позволяет сделать вывод, что выбор вычислительной архитектуры должен выполняться с учётом характера конкретного алгоритма. Для алгоритмов с регулярной структурой и высокой вычислительной нагрузкой на пиксель перенос на GPU даёт значительный эффект. К таким задачам относятся медианная фильтрация, размытие по Гауссу при достаточно большом σ и быстрое преобразование Фурье. В этих случаях затраты на передачу данных компенсируются большим объёмом параллельно выполняемых операций.

Для простых локальных операций ситуация менее однозначна. Оператор Собеля обладает высокой степенью параллелизма, однако каждая операция требует сравнительно небольшого числа вычислений. Поэтому полное время CUDA реализации оказывается ограничено не столько скоростью вычислительного ядра, сколько обменом данными между CPU и GPU. В таких сценариях OpenMP часто является более сбалансированным решением, так как позволяет получить значительное ускорение без усложнения архитектуры программы.

Последовательные CPU реализации сохраняют значение как базовый вариант для проверки корректности, отладки и обработки небольших изображений. OpenMP является технологически простым способом ускорения существующего C++ кода, поскольку требует минимальных изменений структуры программы. CUDA целесообразна в случаях, когда необходимо обрабатывать большие изображения, выполнять ресурсоёмкие операции или строить конвейер обработки, в котором несколько этапов последовательно выполняются на GPU без возврата данных в оперативную память после каждого шага.

Заключение

В статье рассмотрены результаты сравнительного анализа реализаций классических алгоритмов цифровой обработки изображений на CPU и GPU. Показано, что эффективность параллельных вычислений существенно зависит от структуры алгоритма и соотношения между вычислительной нагрузкой и накладными расходами.

Полученные результаты могут быть использованы при разработке систем цифровой обработки изображений, где требуется обоснованный выбор между последовательной CPU реализацией, многопоточной OpenMP версией и реализацией на GPU. Практическая значимость работы заключается в формировании рекомендаций по выбору вычислительного подхода в зависимости от размера входных данных, сложности алгоритма и требований к скорости обработки.

Список литературы

1. Кузнецов А. В., Шишкина, Э. Л. Цифровая обработка изображений / А. В. Кузнецов., Э. Л. Шишкина. – Издательский дом ВГУ, 2023. – 160 с.
2. Hennessy J. L., Patterson D. A. Computer Organization and Design: The Hardware/Software Interface / J. L. Hennessy, D. A. Patterson. – 5th ed., Morgan Kaufmann, 2014. – 793 p.
3. Казеннов А. М. Основы технологии CUDA // компьютерные исследования и моделирование. – 2010.
4. Мыцких-Коробанов А. Ю. Алгоритм размытия по гауссу // Математика и ее приложения в современной науке и практике : Сборник научных статей VIII Международной научно-практической конференции. – 2018.
5. Бардин Б. В. Быстрый алгоритм медианной фильтрации // Научное приборостроение. – 2011.
6. Fang J., Zhang C. Edge Detection Based on Improved Sobel Operator // 2016 International Conference on Computer Engineering and Information Systems. – 2016.
7. Мырадовна А. Л, Баймухаммедович Р. Ш. Быстрые алгоритмы дискретного преобразования Фурье // Наука и мировоззрение. – 2024.
8. cuFFT API Reference. [Электронный ресурс] : Документация, версия от 09.03.2026. – URL: https://docs.nvidia.com/cuda/cufft/.

References

1. Kuznetsov A. V., Shishkina E. L. Digital Image Processing / A. V. Kuznetsov, E. L. Shishkina. – Voronezh State University Publishing House, 2023. – 160 p.
2. Hennessy J. L., Patterson D. A. Computer Organization and Design: The Hardware/Software Interface / J. L. Hennessy, D. A. Patterson. – 5th ed. – Morgan Kaufmann, 2014. – 793 p.
3. Kazennov A. M. Fundamentals of CUDA Technology // Computer Research and Modeling. – 2010.
4. Mytskikh-Korobanov A. Yu. Gaussian Blur Algorithm // Mathematics and Its Applications in Modern Science and Practice: Proceedings of the VIII International Scientific and Practical Conference. – 2018.
5. Bardin B. V. Fast Median Filtering Algorithm // Scientific Instrumentation. – 2011.
6. Fang J., Zhang C. Edge Detection Based on Improved Sobel Operator // 2016 International Conference on Computer Engineering and Information Systems. – 2016.

7. Myradovna A. L., Baimukhammedovich R. Sh. Fast Algorithms of Discrete Fourier Transform // Science and Worldview. – 2024.
8. cuFFT API Reference [Electronic resource]: Documentation, version dated 09.03.2026. – URL: https://docs.nvidia.com/cuda/cufft/.

Назарбаев Ф.Т., Орозбек уулу Дастан Применение искусственного интеллекта в процессе кодирования и разработки обобщенных задач по предмету геометрии 7-го класса// Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7 (69) Ч. 1 с. 30–36

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.42, 519

ПРИМЕНЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ПРОЦЕССЕ КОДИРОВАНИЯ И РАЗРАБОТКИ ОБОБЩЕННЫХ ЗАДАЧ ПО ПРЕДМЕТУ ГЕОМЕТРИИ 7-ГО КЛАССА

[1]**Назарбаев Ф.Т., Орозбек уулу Дастан**
КЫРГЫЗСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМЕНИ ЖУСУПА БАЛАСАГЫНА , Бишкек, Кыргызская Республика (720033, Бишкек, ул. Михаила Фрунзе, 547.), e-mail: [1]nazarbaevfarkhat@gmail.com

В работе рассматривается использование искусственного интеллекта (ИИ) для генерации обобщённых задач по геометрии 7-го класса с применением языка программирования Python. В процессе эксперимента выявлено, что ИИ способен создавать только простые шаблонные задания, но испытывает трудности при формировании уникальных задач с точными координатами и геометрическими зависимостями. Отмечено, что алгоритмы ИИ не всегда корректно интерпретируют пространственные отношения, что приводит к ошибкам в построениях. Вместе с тем, при правильно поставленных условиях ИИ демонстрирует понимание геометрических ограничений, что подтверждает его потенциал. Сделан вывод, что интеграция ИИ и Python может стать эффективным инструментом в образовательной практике при дальнейшем совершенствовании технологий.

Ключевые слова: Искусственный интеллект, Python, генерация задач, геометрия, цифровизация образования, алгоритмизация.

THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN CODING AND IN THE DEVELOPMENT OF GENERALIZED PROBLEMS FOR 7TH-GRADE GEOMETRY

[1]**Nazarbayev F.T., Orozbek uulu Dastan**
NATIONAL RESEARCH NUCLEAR UNIVERSITY MEPHI, Moscow, Russia (115409, Moscow, Kashirskoye sh., 31), e-mail:

This paper examines the use of artificial intelligence (AI) to generate generalized 7th-grade geometry problems using the Python programming language. The experiment revealed that the AI is capable of creating only simple template problems but has difficulty generating unique problems with precise coordinates and geometric dependencies. It is noted that the AI algorithms do not always correctly interpret spatial relationships, leading to construction errors. However, under the right conditions, the AI demonstrates an understanding of geometric constraints, confirming its potential. It is concluded that the integration of AI and Python could become an effective tool in educational practice with further technological advancement.

Keywords: Artificial intelligence, Python, problem generation, geometry, digitalization of education, algorithmization.

Современное образование переживает процесс глубокой цифровой трансформации. Искусственный интеллект (ИИ) постепенно становится частью повседневной жизни, и его интеграция в учебный процесс открывает широкие новые возможности для автоматизации, индивидуализации и повышения эффективности обучения. Особенно актуальной тема становится в контексте школьного курса геометрии, где большинство задач имеют строго алгоритмическую структуру и могут быть формализованы.

Базовые и теоретическое определение обобщенных задач введены и исследованы в работах Панкова П.С. и его учеников [1], [2], а практическое применение показано в работе [3]. Кроме того в работе [4] показано использование современных технологий проверки умений и пониманий графиков функций с использованием языка программирования Delphi, но есть ограничения совместного использования данного приложения из-за вводимых ограничений, т.е. платформенно зависимости, а в работе [5] показано разработка приложения для генерации графически обобщенных задач, для компьютерного контроля знаний удовлетворяющего предъявляемым требованиям система контроля знаний, в работе [6] показано как можно использовать современные технологии в разработке задач с возможностью генерации задач с генерацией диаграмм на основе псевдослучайных чисел, которые на самом деле могут повлиять на само решение задачи (но к сожалению рассмотрено его применение только в случае математической статистики). Но в этой работе показано как можно использовать современные ИИ системы и язык программирования python для генерации в основном графически обобщенных задач, с возможностью его дальнейшего применения в различных платформах, т.е. гарантирующая кроссплатформенность полученных методов разработки обобщенных задач.

Язык программирования Python, включённый в школьную программу, отличается простотой синтаксиса и доступностью для понимания учащимися. Это делает его идеальной платформой для разработки инструментов, способных автоматически генерировать учебные задания, визуализировать решения и помогать учителю в создании обобщённых задач. Однако, как показывает практика, процесс автоматизации генерации задач при помощи ИИ имеет не только преимущества, но и серьёзные методические и технические ограничения.

Создание обобщённых задач [1-3] по геометрии требует строгого соблюдения математических закономерностей и геометрических аксиом. Искусственный интеллект способен формировать текстовые условия и даже генерировать код построений, однако он не обладает пониманием геометрических фигур геометрии. В результате задачи, создаваемые ИИ, нередко содержат неточности, противоречия или математически невозможные условия.

Искусственный интеллект, обученный на больших текстовых массивах, способен генерировать текст условий, решений и даже элементы графического оформления в формате LaTeX и TikZ. Однако геометрия требует строгой точности и соблюдения аксиом, что делает задачу особенно сложной для языковых моделей, не обладающих встроенной математической логикой.

ИИ успешно справляется с шаблонными заданиями, где структура повторяется: например, «постройте две параллельные прямые» или «найдите угол между секущей и параллельными прямыми». При этом формулировки задач корректны, а построения визуально верны. Но при попытке сгенерировать уникальные или комбинированные задачи возникает проблема — модель не всегда способна корректно задать координаты точек и геометрические зависимости, что делает построение некорректным.

Для исследования была реализована серия функций на языке Python, каждая из которых формирует задачу по конкретной теме курса геометрии 7-го класса. Ниже приведён пример функции для темы «Параллельные прямые»:

```
def geo_0005():
    """Тема 3 — Параллельные прямые."""
    A, B = pick_letters(2)
    t = f"""
\\textbf{{Условие задачи7.}}
Постройте две параллельные прямые $ {A}{B} $ и $ {A}'{B}' $.

\\textbf{{Рисунок.}}
\\begin{{tikzpicture}}[scale=0.8]
\\draw (-1,0)--(4,0);
\\draw (-1,2)--(4,2);
\\node[left] at (0,0){{$ {A} $}};
\\node[right] at (3,0){{$ {B} $}};
\\node[left] at (0,2){{$ {A}' $}};
\\node[right] at (3,2){{$ {B}' $}};
\\end{{tikzpicture}}

\\textbf{{Решение.}}
Параллельные прямые не пересекаются.
"""
    return t
```

Рисунок 1 - Функция обобщенной задачи с генерацией графика.

Такие функции генерируют не только текст задачи, но и код для автоматического построения иллюстрации с помощью LaTeX. Этот подход позволяет создавать целые серии упражнений с минимальными изменениями в коде.

На Рисунке 2 ниже приведён пример, иллюстрирующий типичные ошибки при генерации задач ИИ:

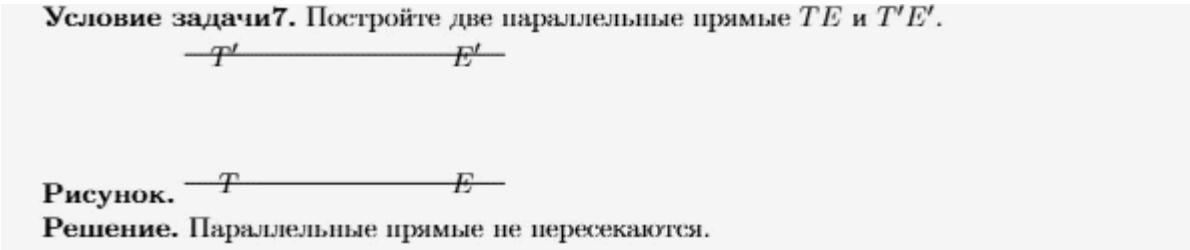


Рисунок 2 - Ошибки визуализации при построении параллельных прямых

На изображении заметно, что буквенные обозначения T и E расположены неровно, линии построения имеют дополнительные черточки и перекосы, а надписи частично пересекаются с графическими элементами. Подобные артефакты возникают из-за того, что ИИ не умеет точно рассчитывать координаты фигур и не контролирует правила оформления геометрических чертежей.

В ходе второго эксперимента при запросе «создать функцию для треугольника со сторонами 4, 5 и 11» искусственный интеллект сформировал корректный по синтаксису код на Python, однако не выполнил проверку геометрической реализуемости фигуры.

Несмотря на то, что такой треугольник физически невозможно построить, ИИ не сообщил об ошибке, а вместо этого сгенерировал другую фигуру, внешне напоминающую треугольник.

Таким образом, искусственный интеллект не только не выполнил поставленную задачу, но и создал иллюзию правильного результата, что типично для большинства современных языковых моделей.

```python
def geo_1():
    """Тема: Треугольник со сторонами 4, 5 и 11."""
    a, b, c = 4, 5, 11
    t = f"""
\\textbf{{Условие задачи.}}
Постройте треугольник со сторонами $a={a}$, $b={b}$, $c={c}$.

\\textbf{{Рисунок.}}
\\begin{{tikzpicture}}[scale=0.6]
\\coordinate (A) at (0,0);
\\coordinate (B) at ({c},0);
\\coordinate (C) at (2,3); % <-- координаты даны без проверки корректности
\\draw (A)--(B)--(C)--cycle;
\\node[below left] at (A){{$A$}};
\\node[below right] at (B){{$B$}};
\\node[above] at (C){{$C$}};
\\end{{tikzpicture}}

\\textbf{{Решение.}}
Построим треугольник $ABC$ со сторонами $AB={c}$, $BC={a}$, $CA={b}$.
Найдём его высоту и площадь по формуле Герона.
"""
    return t
```

Рисунок 3 - Функция написанный на Python

Подобное поведение объясняется принципом функционирования ИИ — модель стремится угодить пользователю и выдать ответ, соответствующий его ожиданиям, даже если он противоречит логике или законам математики.

Вместо того чтобы сообщить: «такой треугольник невозможен, поскольку 4 + 5 = 9, а это меньше 11», ИИ «вежливо» строит фигуру, которая визуально кажется правильной, но в действительности не подчиняется правилам геометрии.

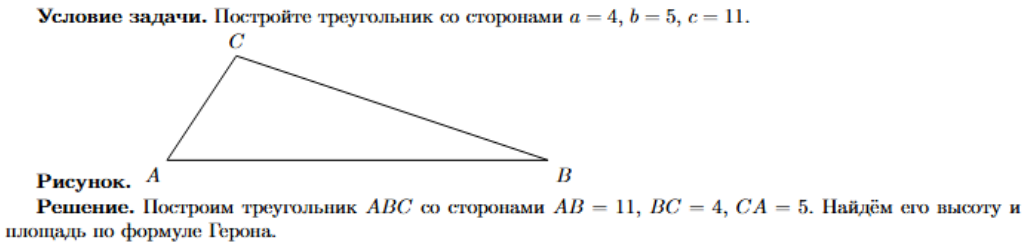


Рисунок 4 - Скриншот задачи и его решения сгенерированного на Python.

На визуализации (Рисунок 4) отчётливо видно, что здесь абсолюно другая фигура отличающаяся от того что мы запросили.

Такой результат подчёркивает одну из главных проблем использования искусственного интеллекта в образовательных целях — отсутствие критического анализа и смысловой проверки создаваемого контента. ИИ не способен оценивать истинность или физическую реализуемость решения, он лишь воспроизводит наиболее вероятные шаблоны, которые

«выглядят убедительно».В условиях обучения это может привести к тому, что учащиеся примут визуально корректное, но логически ложное построение за верное.

Следовательно, использование ИИ при генерации геометрических задач требует обязательного участия человека — преподавателя или методиста, который способен выявить подобные несоответствия и объяснить учащимся, почему фигура, созданная искусственным интеллектом, невозможна в реальности. Этот случай демонстрирует не только технические ограничения ИИ, но и важность сохранения человеческого контроля над образовательным процессом. Именно педагог остаётся тем звеном, которое соединяет цифровые технологии с истинным пониманием законов природы и математики.

В ходе этого экспериментов были выявлены следующие ключевые трудности:

1. Ограниченная уникальность. Генерация задач опирается на шаблоны, поэтому ИИ часто повторяет однотипные конструкции, не создавая новых ситуаций.
2. Ошибки координат и масштабов. При попытках построить более сложные фигуры нарушаются пропорции и взаимное расположение элементов.
3. Отсутствие геометрического контекста. Модели ИИ не обладают «пониманием» геометрических свойств и не способны проверить существование фигуры без явной формулы.
4. Проблема текстовой избыточности. В сгенерированных задачах часто встречаются лишние элементы описания, не влияющие на смысл, но затрудняющие использование в учебных целях.

Несмотря на эти трудности, ИИ продемонстрировал способность обрабатывать абстрактные понятия, распознавать ключевые термины и создавать правильные рассуждения в простых случаях. Это открывает перспективу для разработки гибридных систем, где человек-методист контролирует структуру, а ИИ автоматизирует вариативность и визуализацию.

Использование Python в сочетании с ИИ позволяет объединить программирование, визуализацию и автоматизацию задач. Такая интеграция способствует развитию у школьников алгоритмического мышления, межпредметных связей между математикой и информатикой, а также навыков работы с цифровыми инструментами.

В ходе экспериментов использовались несколько популярных моделей ИИ:

- ChatGPT (OpenAI) — способен создавать шаблонные тексты задач, иногда корректно генерирует LaTeX-код, но не всегда проверяет логическую совместимость условий (например, треугольник со сторонами 4, 5 и 11).
- Google Gemini — демонстрирует более точные формулировки, но практически не справляется с построением графических элементов.
- GitHub Copilot — хорошо интегрируется с Python и создаёт синтаксически правильные функции, но не понимает геометрический смысл построений.

Результаты показывают, что ни одна из систем не обеспечивает полноценного цикла генерации корректной геометрической задачи, включающей условие, рисунок и решение. Таким образом, ИИ можно рассматривать лишь как вспомогательный инструмент, а не как автономного разработчика учебного материала.

В перспективе возможно создание интеллектуальных систем, которые будут:

- формировать задания различной сложности в зависимости от уровня ученика;
- автоматически проверять корректность решений;

- визуализировать построения в интерактивной форме;
- подбирать задачи с учётом типичных ошибок учащихся.

Подобные системы могут стать частью цифровой образовательной среды, где ИИ выступает не заменой учителя, а интеллектуальным помощником в разработке учебных материалов.

В результате исследования установлено, что современные инструменты искусственного интеллекта способны эффективно генерировать простые и шаблонные задачи по геометрии, особенно при использовании Python как платформы для визуализации и автоматизации. Однако создание уникальных, содержательно глубоких задач остаётся за пределами текущих возможностей ИИ.

Таким образом, ИИ и Python могут рассматриваться как дополнительный инструмент в методике преподавания геометрии — средство ускорения рутинных процессов и расширения дидактического потенциала, но не как полноценная альтернатива человеческому мышлению и педагогическому опыту.

Список литературы

1. Панков П.С. Обучающая и контролирующая программа по словоизменению в кыргызском языке на ПЭВМ. - Бишкек: Мектеп, 1992. - 20 с.
2. Панков П.С., Джаналиева Ж.Р. Проектирование и развитие программных экзаменационных комплексов по математике и физике // Образование в XXI веке: ценности и перспективы: Материалы Международной научно-практической конференции. Часть 2. - Бишкек: Кыргызский институт образования, 2001. - С. 281-284.
3. Назарбаев Ф.Т. Разработка экзаменационных систем по контролю качества образования по математике и информатике / Назарбаев Ф.Т. // Известия КГТУ им. И.Раззакова - Бишкек, 2017. №42 - С. 196-202
4. Nazarbaev F.T., Usupova N.M., Zhanybekova K.Zh. Development of a system for monitoring practical knowledge and skills in charting // Herald of Institute Mathematics of the National Academy of Sciences of the Kyrgyz Republic. 2024. № 1. С. 206-213.
5. Назарбаев Ф.Т., Ниязбаева Н.М., Создание графически обобщённых задач в экзаменационных системах // Alatoo Academic Studies, №1 2019, ISSN: 1694-5263 eISSN: 1694-7916, С. 72-80
6. Назарбаев Ф.Т., Бейшебаева Ж.К., Реализация комплексной экзаменационной системы по дисциплине "теория вероятностей и математическая статистика" и его особенности // Вестник Кыргызского Национального Университета имени Жусупа Баласагына, №S1 2019, ISSN: 1694-8033eISSN: 1694-8025, С. 212-218

Reference

1. Pankov P.S. Teaching and controlling program for inflection in the Kyrgyz language on PCs. - Bishkek: Mektep, 1992. - 20 p
2. Pankov P.S., Dzhanalieva Zh.R. Design and Development of Program Examination Complexes in Mathematics and Physics // Education in the XXI Century: Values and Prospects: Materials of the International Scientific and Practical Conference. Part 2. - Bishkek: Kyrgyz Institute of Education, 2001. - pp. 281-284.

3. Nazarbaev F.T. Development of Examination Systems for Quality Control of Education in Mathematics and Informatics / Nazarbayev F.T. // Izvestiya KSTU named after I. Razzakov - Bishkek, 2017. №42 - pp. 196-202
4. Nazarbaev F.T., Usupova N.M., Zhanybekova K.Zh. Development of a system for monitoring practical knowledge and skills in charting // Herald of Institute Mathematics of the National Academy of Sciences of the Kyrgyz Republic. 2024. № 1. pp. 206-213.
5. Nazarbayev F.T., Niyazbaeva N.M., Creating graphically generalized problems in examination systems // Alatoo Academic Studies, No1 2019, ISSN: 1694-5263 eISSN: 1694-7916, pp. 72-80
6. Nazarbayev F.T., Beishebaeva Zh.K., Implementation of the complex examination system in the discipline "probability theory and mathematical statistics" and its features // Bulletin of the Kyrgyz National University named after Zhusup Balasagyn, NoS1 2019, ISSN: 1694-8033eISSN: 1694-8025, pp. 212-218

Вагапов Р.Р. Эволюция базы знаний в крупных ИТ-проектах: от централизации до интеллектуализации// Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7(69) Ч.1 с. 37–42

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.41:005.336.2:004.89

ЭВОЛЮЦИЯ БАЗЫ ЗНАНИЙ В КРУПНЫХ ИТ-ПРОЕКТАХ: ОТ ЦЕНТРАЛИЗАЦИИ ДО ИНТЕЛЛЕКТУАЛИЗАЦИИ

Вагапов Р.Р.
ФГБОУ ВО "УФИМСКИЙ ГОСУДАРСТВЕННЫЙ НЕФТЯНОЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ" , Уфа, Россия (450064, Республика Башкортостан, город Уфа, ул. Космонавтов, д. 1), e-mail: ram.vaga@yandex.ru

В крупных ИТ-проектах ведение документации осложняется отсутствием единого подхода к ее ведению, что приводит к использованию множества разрозненных систем ведения и хранения информации. В следствии чего артефакты команды зачастую дублируются и могут терять актуальность. Данная статья предлагает рассмотрение базы знаний не только как статичное хранилище, а как эволюционирующую среду. Описаны три стадии зрелости базы знаний: централизация (единый репозиторий и формат), стандартизация (роли, шаблоны, регламент ведения) и интеллектуализация (семантический поиск на основе RAG). Каждая стадия должна проходить поэтапно, создавая фундамент для следующей. Методика является универсальной для всех крупных проектов с большим объемом документации независимо от отрасли.

Ключевые слова: База знаний, эволюция, стадии зрелости, Docs-ac-Code, централизация, стандартизация, интеллектуализация, RAG.

THE EVOLUTION OF THE KNOWLEDGE BASE IN LARGE IT PROJECTS: FROM CENTRALIZATION TO INTELLECTUALIZATION

Vagapov R.R.
UFA STATE PETROLEUM TECHNOLOGICAL UNIVERSITY , Ufa, Russia (450064, Republic of Bashkortostan, Ufa, Kosmonavtov str., 1), e-mail: ram.vaga@yandex.ru

In large IT projects, documentation management is complicated by the lack of a unified approach to its management, which leads to the use of many disparate information management and storage systems. As a result, team artifacts are often duplicated and may lose relevance. This article suggests considering the knowledge base not only as a static repository, but as an evolving environment. Three stages of knowledge base maturity are described: centralization (single repository and format), standardization (roles, templates, rules of conduct) and intellectualization (semantic search based on RAG). Each stage should take place in stages, creating the foundation for the next one. The methodology is universal for all large projects with a large volume of documentation, regardless of the industry.

Keywords: Knowledge base, evolution, maturity stages, Docs-ac-Code, centralization, standardization, intellectualization, RAG.

Крупный ИТ-проект – это десятки людей в команде, сотни артефактов и длительный жизненный цикл. При отсутствии структурированного подхода к ведению и поддержанию документации неизбежно будет наступать хаос в следствии ведения ее в различных источниках. При таком подходе документация может терять свою актуальность, сотрудники

будут затрачивать большое количество времени на ее поддержание, также проблематичен становится онбординг новых сотрудников, они вынуждены затрачивать недели на поиск и сбор целостной картины по проекту. Методология Docs-as-Code не дает полного решения проблемы хаоса с документацией, она дает техническую базу для реализации структуры и порядка за счет ведения репозиториев в Git, использования единого языка разметки и автоматической сборки через механизм Pull Request. Нужно понимать, что база знаний – это не продукт, который можно быстро внедрить и он будет работать. Это эволюционный путь команды, который требует от каждого участника понимания ее ценности и готовности поддержания ее актуальности в зависимости от роли. В статье описан путь из трех стадий становления базы знаний. Каждая стадия решает свою определенную задачу и является фундаментов для следующей.[1]

Первая стадия – централизация. На данном этапе главной целью является собрать всю документацию в едином месте и обеспечить доступ для каждого участника. [2] Единой точкой входа является Git-репозиторий, в который переносится вся существующая документация на языке разметки AsciiDoc. AsciiDoc является предпочтительнее Markdown за счет поддержки более сложных структур и расширенных возможностей форматирования, что лучше подходит для технической документации. Настраиваются CI/CD пайплайны и Antora, статический генератор, который преобразует .adoc-файлы из репозитория в удобный HTML-сайт с навигацией. [3] При каждом обновлении в репозитории через механизм Pull Request осуществляется автоматическая сборка и публикация файлов. В итоге на данном этапе команда получает единую точку входа информации, без необходимости ведения в различных источниках, появляется прослеживаемость истории изменений. Но без правил ведения централизованная документация может быстро превратиться в свалку.[4]

Вторая стадия – стандартизация. На данном этапе подразумевается переход от просто единого хранилища к управляемому процессу ведения и поддержания документации с доступом для всех участников команды. Вводятся три основных элемента: роли, шаблоны, регламент ведения. За каждой ролью закрепляется ответственность за ведение документации согласно ее профилю. Аналитик отвечает за технические спецификации и диаграммы. Разработчик за архитектуру и API. Тестировщик за тест-кейсы и чек-листы. Технический писатель за пользовательскую документацию. DevOps настраивает инфраструктуру работы базы знаний. Команда реализует шаблоны для документации в AsciiDoc. Для технических спецификаций это может быть user story, use case, описание дизайна, фронта и бэка, описание API и иных диаграмм, а также критерии приемки. Для тест-кейсов это предусловия, шаги выполнения и ожидаемый результат с ссылками на спецификации. [5] Для инструкции пользователя согласно требованиям возможно описание ролевой модели, описание функционала, пошаговые сценарии по функционалу, описание возможных ошибок в системе. Для ведения физической модели данных универсальным будет формат DBML. Шаблоны позволят экономить время и поддерживать единую структуру всей базы знаний. Предложенный регламент по ведению подразумевает под собой следующее. Любое изменение начинается с создания отдельной ветки в Git. Автор создает новый файл или вносит изменения в действующие по шаблонам. После внесения изменений оформляется Pull Request и назначаются рецензенты, которые отсматривают внесенные изменения на корректность, проверяя соответствие шаблонам и полноту описания. После получения согласования от всех рецензентов Pull Request вливается в основную ветку main, и настроенные CI/CD пайплайны

автоматически пересобирают Antora. Данный процесс гарантирует, что в базу знаний будет попадать только проверенная документация. На данном этапе исключается проблема определения ответственных за обновление документации и команда постепенно привыкает дисциплинировано вести документацию, согласно своей занимаемой роли в проекте.[6]

Третья стадия – интеллектуализация. Когда база знаний уже централизована и стандартизирована, то есть прошла вторую стадию, появляется возможность внедрения RAG (Retrieval-Augment Generation) системы. На данном этапе документация перестает быть статичной, теперь появляется возможность для реализации вопросно-ответной системы по всей базе знаний. Пользователь пишет запрос на естественном языке, и система ищет релевантные фрагменты, передает их в большую языковую модель и возвращает сгенерированный ответ со ссылками на источники. [7] Сервис запросов (RAG-движок) работает следующим образом. Пользователь вводит запрос через кастомный виджет, интегрированный в строку поиска Antora. Сервис векторизует вопрос с помощью модели эмбеддингов, затем выполняется гибридный поиск в векторной базе данных. Наиболее релевантные найденные чанки вместе с исходным вопросом подгружаются в шаблонный промт. Большая языковая модель генерирует ответ и возвращает в RAG-сервис для последующего отображения пользователю. Архитектурная схема компонентов отображена на Рисунке 1. На Рисунках 2-3 представлены диаграммы последовательности процесса индексирования и поиска

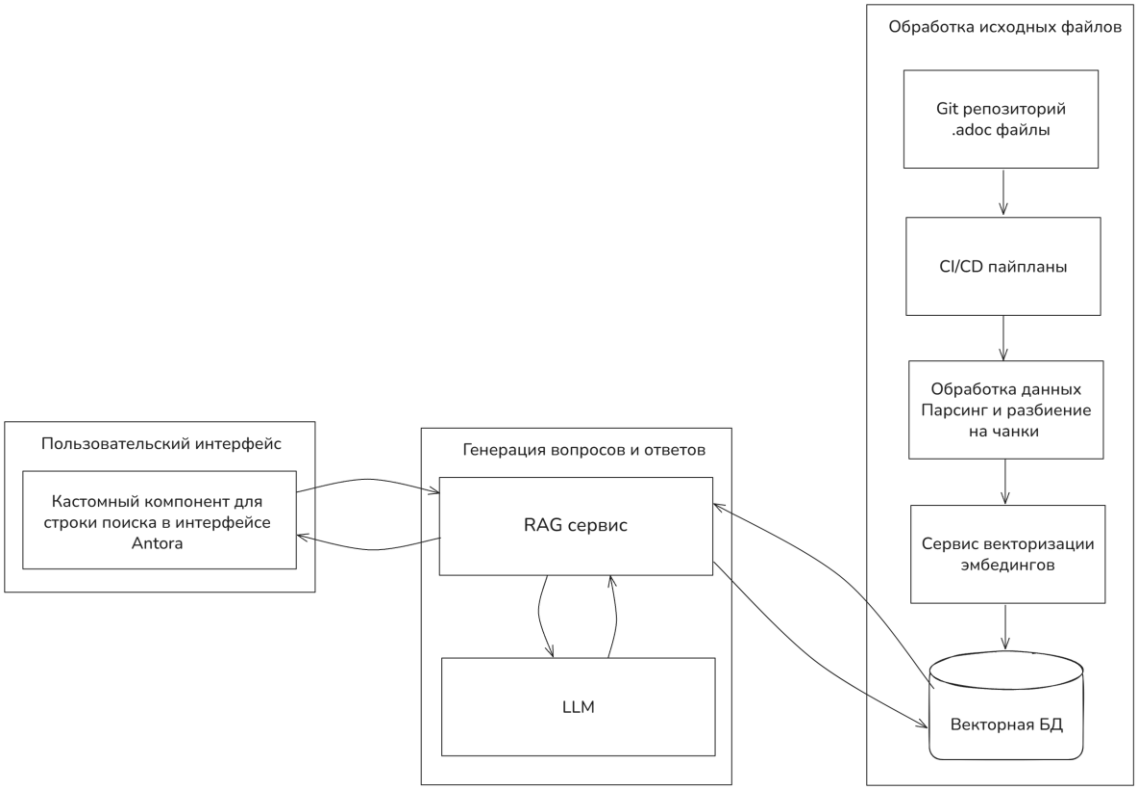


Рисунок 1 - Архитектурная схема компонентов

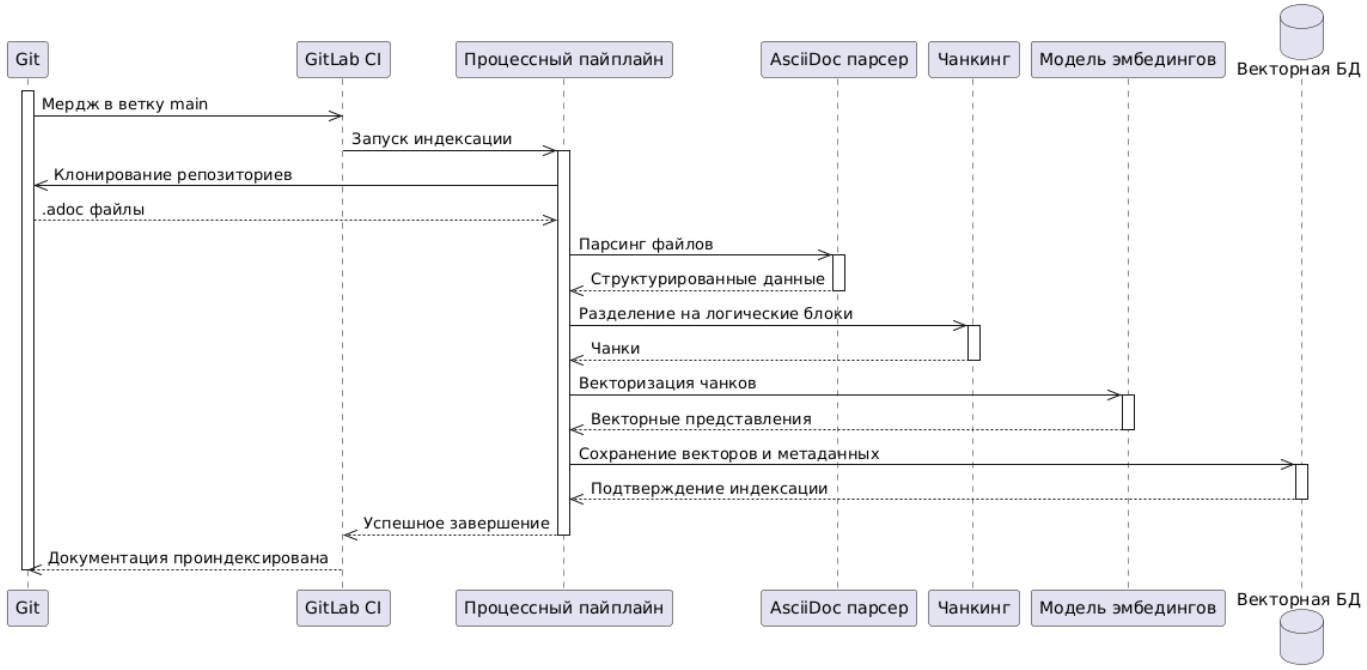


Рисунок 2 - Диаграмма последовательности процесса индексирования

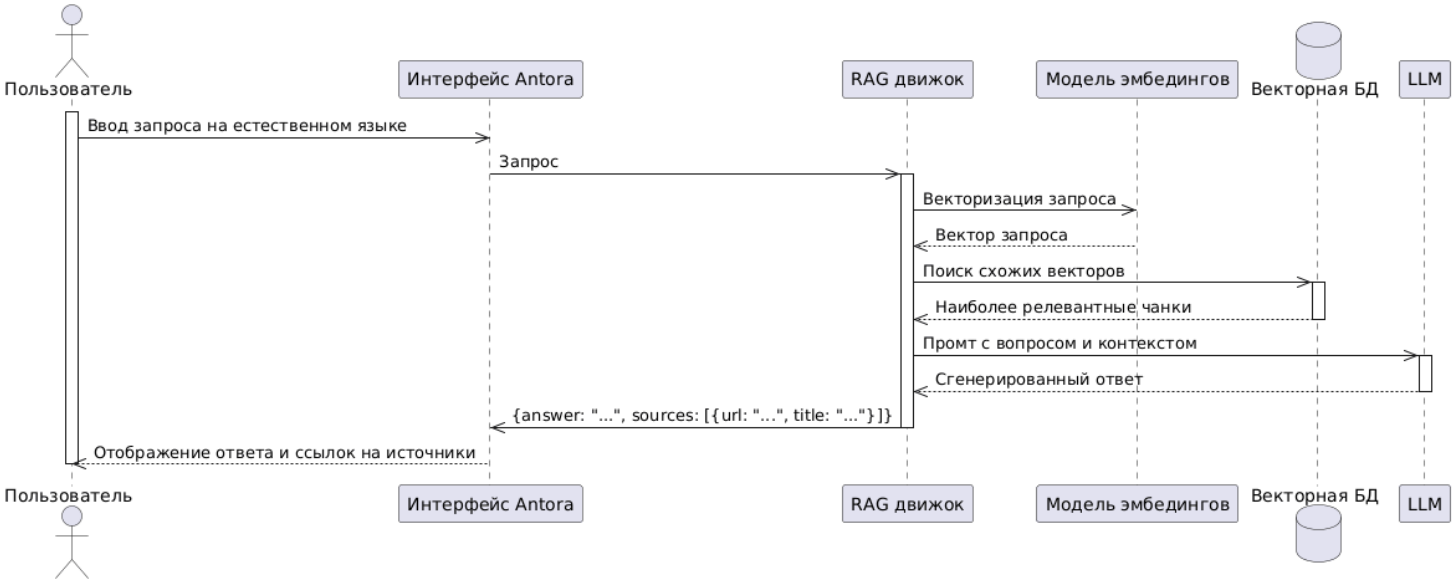


Рисунок 3 - Диаграмма последовательности процесса поиска

Главное преимущество внедрения RAG-системы перед стандартными поисками является гарантия безопасность данных за счёт полного on-premise развёртывания, обеспечение качественного семантический поиска и встраивание в CI/CD-процессы, поддерживая регулярную индексацию и постоянную актуальность данных. Однако RAG работает только при условии наличия чистой и структурированной документации. Если на вход будет попадать мусор, он будет и на выходе. [8] Не имеет смысла индексировать неактуальные спецификации, а также те, которые ведутся не структурировано. Именно поэтому внедрение RAG-системы для действительно качественной работы возможно только на третьей стадии, когда вся документация ведется по единым шаблонам, прошла множество

ревью и регулярно поддерживается в актуальном состоянии. [9] Только в таком случае RAG позволит получать точные ответы и предоставит возможность значительно экономить время всей команды, а также будет особенно полезен при онбординге новых сотрудников.

Таким образом предложенный концепт для базы знаний в крупных ИТ-проектах позволит реализовать не просто статичное хранилище, а эволюционирующую среду. Полноценное развитие возможно только через последовательные стадии, где каждая опирается на результаты предыдущей. Прыжок через этап невозможен, так как без порядка, структуры и стандартов ведения интеллектуальный поиск не даст высокого качества. Предложенный концепт универсален и применим в любом крупном проекте с большими объемами документации в независимости от отрасли. [10]

Список литературы

1. ИТ-отрасль: ключевые показатели развития за 2019-2024 гг. URL: https://issek.hse.ru/news/1047754209.html
2. Docs as Code: введение в предмет. URL: https://habr.com/ru/companies/plesk/articles/555110/
3. Five fast facts about docs as code at GitLab. URL: https://about.gitlab.com/blog/five-fast-facts-about-docs-as-code-at-gitlab/
4. Документация Antora. URL: https://docs.antora.org/antora/latest/
5. Набор инструментов для работы с документацией. URL: https://gitlab.com/antora
6. Биненда, А. Д. Практическое применение внедрения подхода Doc as Code в IT компании с применением инструмента Antora / А. Д. Биненда // Актуальные исследования. 2024. № 19-1(201). С. 17-20. EDN LYEWUU.
7. Вагапов Р.Р. Разработка концепции единой базы знаний для управления артефактами команды разработки. Редакция сборника XIV Всероссийской научно-практической конференции «АКТУАЛЬНЫЕ ВОПРОСЫ ЭКОНОМИКИ, УПРАВЛЕНИЯ и ЦИФРОВОЙ ТРАНСФОРМАЦИИ В НЕФТЕГАЗОВОМ БИЗНЕСЕ и СТРОИТЕЛЬСТВЕ» - Уфа 2025 - С. 54-56;
8. Retrieval-Augmented Generation (RAG): глубокий технический обзор. URL: https://habr.com/ru/articles/931396/
9. Полный Обзор Архитектуры RAG с Векторными Базами Данных: От Фундаментальных Принципов до Продвинутой Оптимизации URL: https://external.software/archives/74725
10. Вагапов Р.Р. От документации в виде кода до интеллектуального поиска с помощью RAG International Conference on Current Problems of Engineering and Applied Sciences (ICCPEAS 2025)

References

1. IT industry: key development indicators for 2019-2024 URL: https://issek.hse.ru/news/1047754209.html
2. Docs as Code: an introduction to the subject. URL: https://habr.com/ru/companies/plesk/articles/555110/
3. Five fast facts about docs as code at GitLab. URL: https://about.gitlab.com/blog/five-fast-facts-about-docs-as-code-at-gitlab/
4. Antora documentation. URL: https://docs.antora.org/antora/latest/

5. A set of tools for working with documentation. URL: https://gitlab.com/antora
6. Binenda, A.D. Practical application of the implementation of the Doc as Code approach in an IT company using the Antora tool / A.D. Binenda // Current research. 2024. No. 19-1(201). pp. 17-20. EDN LYEWUU.
7. Vagapov R.R. Development of the concept of a unified knowledge base for artifact management of the development team. Editorial board of the collection of the XIV All-Russian Scientific and Practical Conference "TOPICAL ISSUES OF ECONOMICS, MANAGEMENT and DIGITAL TRANSFORMATION IN THE OIL and GAS BUSINESS and CONSTRUCTION" - Ufa 2025 - pp. 54-56;
8. Retrieval-Augmented Generation (RAG): in-depth technical overview. URL: https://habr.com/ru/articles/931396/
9. A Complete Overview of the RAG Architecture with Vector Databases: From Fundamental Principles to Advanced URL Optimization: https://external.software/archives/74725
10. Vagapov R.R. From documentation in the form of code to intelligent search using the RAG International Conference on Current Problems of Engineering and Applied Sciences (ICCPEAS 2025)

Филиппов М.И. Сравнительный анализ алгоритмов WAVE FUNCTION COLLAPSE и регулярной сеточной генерации для процедурного создания городской среды // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7(69) Ч.1 с. 43–51

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.02

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ WAVE FUNCTION COLLAPSE И РЕГУЛЯРНОЙ СЕТОЧНОЙ ГЕНЕРАЦИИ ДЛЯ ПРОЦЕДУРНОГО СОЗДАНИЯ ГОРОДСКОЙ СРЕДЫ

Филиппов М.И.
ФГАОУ ВО "ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ", Волгоград, Россия (400062, Волгоградская область, город Волгоград, Университетский пр-кт, д. 100), e-mail: filippov.mikhail.2002@gmail.com

В статье рассматривается сравнительный анализ алгоритма Wave Function Collapse и алгоритма регулярной сеточной генерации, применяемых для процедурного построения городской дорожной сети в игровых приложениях. Описаны принципы работы исследуемых алгоритмов, особенности их реализации в среде Unity и используемые ограничения совместимости дорожных сегментов. Проведена серия вычислительных экспериментов, направленных на оценку устойчивости алгоритмов к изменению набора сегментов, размера генерируемой сетки и дополнительных граничных условий. Сравнение выполнено по таким показателям, как процент успешно заполненных ячеек, среднее количество откатов и время генерации. Полученные результаты позволяют определить преимущества и недостатки каждого подхода и обосновать выбор алгоритма Wave Function Collapse для задач процедурной генерации городской среды.

Ключевые слова: Процедурная генерация, Wave Function Collapse, городская среда, дорожная сеть, игровая карта, Unity, генерация контента, алгоритмы генерации, регулярная сеточная генерация.

A COMPARATIVE ANALYSIS OF WAVE FUNCTION COLLAPSE AND REGULAR GRID GENERATION ALGORITHMS FOR PROCEDURAL URBAN ENVIRONMENT GENERATION

Filippov M.I.
"VOLGOGRAD STATE UNIVERSITY", Volgograd, Russia (400062, Volgograd region, Volgograd city, Universitetsky prospekt, 100), e-mail: filippov.mikhail.2002@gmail.com

This article presents a comparative analysis of the Wave Function Collapse algorithm and the regular grid generation algorithm used for procedural urban road network generation in gaming applications. The operating principles of the studied algorithms, their implementation in the Unity environment, and the road segment compatibility constraints used are described. A series of computational experiments were conducted to assess the algorithms' robustness to changes in the segment set, the generated grid size, and additional boundary conditions. The comparison was based on metrics such as the percentage of successfully filled cells, the average number of rollbacks, and generation time. The results allow us to identify the advantages and disadvantages of each approach and justify the choice of the Wave Function Collapse algorithm for procedural urban environment generation tasks.

Keywords: Procedural generation, Wave Function Collapse, urban environment, road network, game map, Unity, content generation, generation algorithms, regular grid generation.

Введение

Процедурная генерация контента является одним из широко используемых подходов в современной разработке компьютерных игр [1, 2]. Ее применение позволяет автоматически

создавать игровые объекты и окружение без необходимости ручного проектирования каждого элемента. Особенно актуальным данное направление является при создании городских локаций [3], содержащих большое количество дорог, перекрестков и других элементов инфраструктуры.

Одной из задач процедурной генерации является построение дорожной сети города с соблюдением правил соединения отдельных сегментов. При формировании городской среды необходимо учитывать ряд ограничений. Во-первых, соседние дорожные сегменты должны корректно стыковаться между собой, обеспечивая непрерывность полос движения. Во-вторых, дороги не должны выходить за границы генерируемой области, поэтому на внешних границах города запрещается размещение сегментов, имеющих выход дорожного полотна за пределы сетки. Соблюдение указанных условий позволяет получать логически корректные и реалистичные конфигурации городской дорожной сети.

В данной работе рассматриваются два подхода к генерации городской дорожной сети: алгоритм регулярной сеточной генерации и алгоритм Wave Function Collapse (WFC). Первый подход основан на последовательном заполнении ячеек сетки с учетом уже размещенных соседних элементов. Второй использует механизм распространения ограничений между ячейками и позволяет учитывать совместимость сегментов на всех этапах построения структуры.

Для проведения экспериментов был сформирован набор дорожных сегментов, используемых при построении городской среды. Каждый сегмент представляет собой отдельный элемент дорожной сети и содержит информацию о типах соединений на четырех сторонах: северной, южной, западной и восточной. Представление карты в виде набора совместимых элементов является распространенным подходом в задачах процедурного моделирования [4].

В работе использовались прямые участки дорог, повороты, трехсторонние и четырехсторонние перекрестки, а также пустые сегменты. Для повышения вариативности каждый элемент мог использоваться в четырех ориентациях, кратных 90 градусам. В результате было подготовлено 17 уникальных дорожных элементов, что с учетом поворотов позволило получить 68 возможных вариантов размещения.

Для управления частотой появления отдельных элементов использовались весовые коэффициенты. Более простые сегменты, например, прямые дороги, имели больший вес по сравнению с перекрестками, что позволяло формировать более реалистичную структуру городской сети. На рисунке 1 приведены примеры дорожных сегментов.

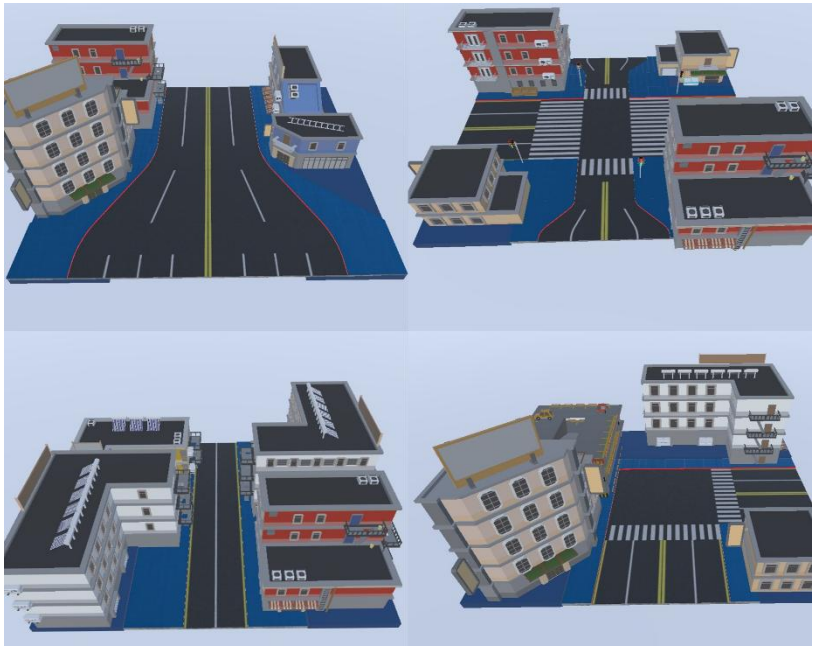


Рисунок 1 - Примеры дорожных сегментов

Алгоритм регулярной сеточной генерации

Алгоритм регулярной сеточной генерации основан на последовательном обходе двумерной сетки. Заполнение выполняется построчно, начиная с верхней левой ячейки.

Для каждой новой позиции учитываются только уже размещенные соседи: верхний и левый. На основании их параметров формируется список допустимых дорожных сегментов. После этого из списка случайным образом выбирается один элемент с учетом его весового коэффициента и размещается в текущей ячейке.

Особенностью алгоритма является отсутствие механизма возврата к предыдущим решениям. Если для очередной ячейки не удается подобрать подходящий сегмент, она остается пустой. По этой причине алгоритм относится к жадным методам генерации. На рисунке 2 приведен пример города, сгенерированным данным методом.

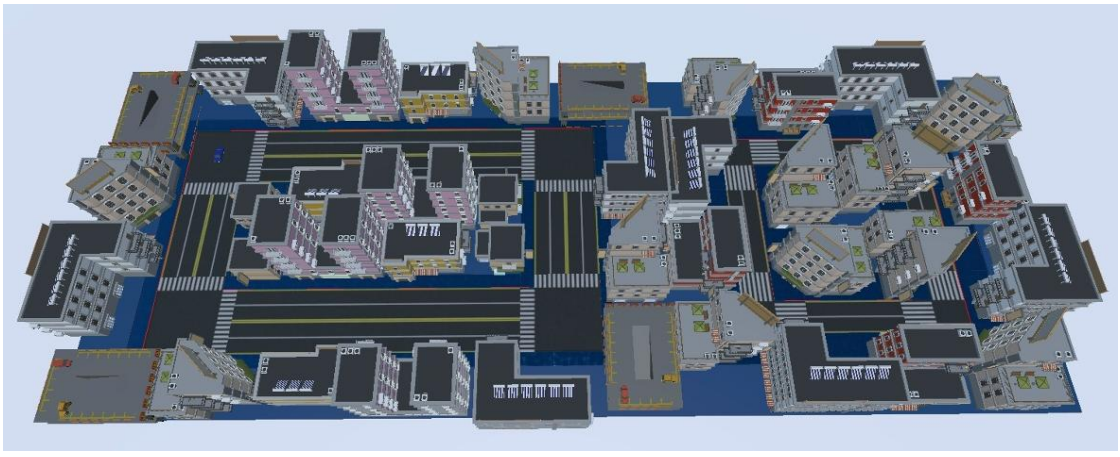


Рисунок 2 - Пример городской среды, сформированной алгоритмом регулярной сеточной генерации

Алгоритм Wafe Function Collapse

Алгоритм Wave Function Collapse основан на распространении ограничений между элементами генерируемой структуры [5, 6]. В отличие от алгоритма регулярной сеточной генерации, который принимает решение только на основании уже размещенных соседей, WFC хранит для каждой ячейки множество всех допустимых вариантов дорожных сегментов.

Перед началом генерации формируется таблица совместимости, содержащая информацию о допустимых сочетаниях сегментов по каждому из четырех направлений. Используемый механизм относится к классу задач удовлетворения ограничений [7]. На начальном этапе каждая ячейка может содержать любой элемент из набора, удовлетворяющий граничным условиям.

Далее выполняется итерационный процесс генерации. На каждом шаге выбирается ячейка с минимальной энтропией, то есть с наименьшим количеством допустимых вариантов размещения. Выбор такой ячейки позволяет в первую очередь разрешать наиболее ограниченные участки карты и снижать вероятность возникновения противоречий на поздних этапах генерации.

После выбора ячейки один из допустимых сегментов фиксируется случайным образом с учетом заданных весовых коэффициентов. Затем запускается процедура распространения ограничений. Для всех соседних ячеек пересчитываются множества допустимых вариантов, из которых удаляются элементы, несовместимые с выбранным сегментом. Если множество допустимых вариантов изменилось, процедура продолжается для следующих соседних ячеек до тех пор, пока система не достигнет согласованного состояния.

В процессе генерации возможно возникновение противоречий, при которых для некоторой ячейки не остается ни одного допустимого варианта размещения. Для обработки таких ситуаций используется механизм отката. В реализованной системе при обнаружении противоречия генерация начинается заново. Для предотвращения бесконечного выполнения алгоритма вводится ограничение на максимальное количество попыток построения карты.

На Рисунке 3 приведен пример города, который сгенерирован с помощью алгоритма Wave Function Collapse.

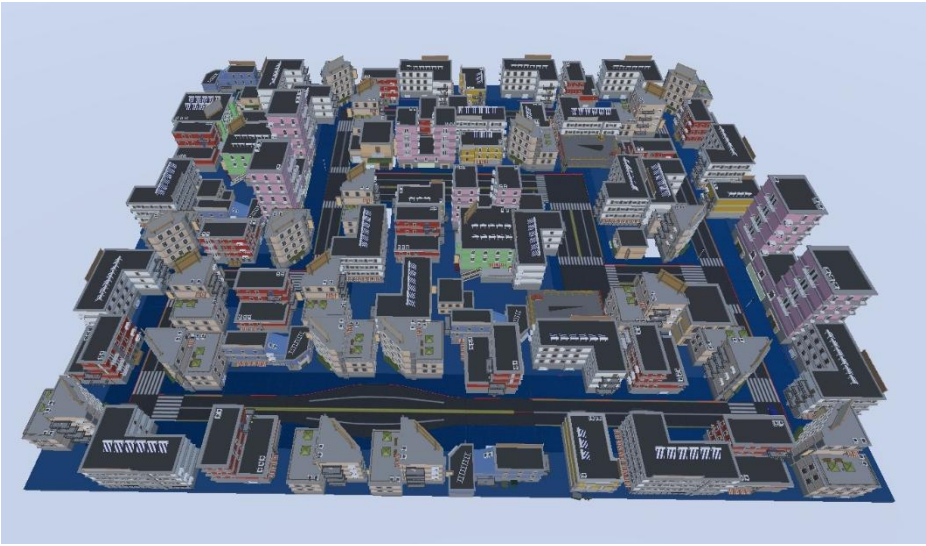


Рисунок 3 - Пример городской среды, сформированной алгоритмом WFC

Сравнительный анализ алгоритмов

Для оценки эффективности рассматриваемых алгоритмов был проведен ряд вычислительных экспериментов. Для каждого набора параметров производилось по 100 запусков генерации с последующим усреднением результатов. В качестве основных показателей были выбраны процент успешно заполненных ячеек, влияние размера сетки на устойчивость генерации, среднее количество откатов и время построения городской среды.

На первом этапе исследования была выполнена оценка устойчивости алгоритмов при изменении набора дорожных сегментов и введении дополнительных ограничений на границах города (Рисунок 4).

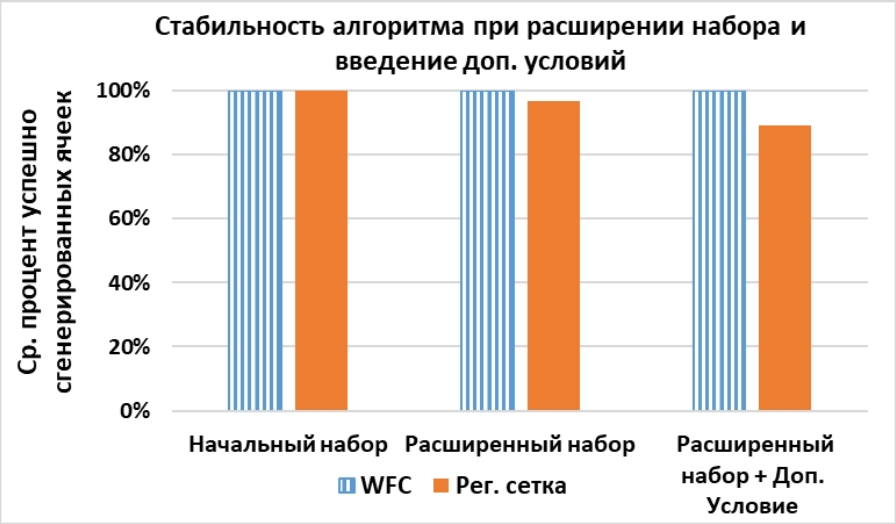


Рисунок 4 - Влияние размера набора сегментов и граничных условий на процент успешно заполненных ячеек

Полученные результаты показывают, что при использовании исходного набора дорожных сегментов оба алгоритма обеспечивают полное заполнение сетки. Однако после расширения набора и введения дополнительных граничных условий наблюдается снижение процента заполнения для алгоритма регулярной сеточной генерации. Причиной является отсутствие механизма исправления ранее принятых решений. В случае возникновения конфликтной ситуации отдельные ячейки остаются пустыми. Алгоритм WFC сохраняет высокий процент успешной генерации даже после усложнения условий. Это объясняется использованием механизма распространения ограничений, который позволяет учитывать совместимость сегментов на протяжении всего процесса построения карты.

Следующим этапом стало исследование зависимости качества генерации от количества ячеек в сетке (Рисунок 5).

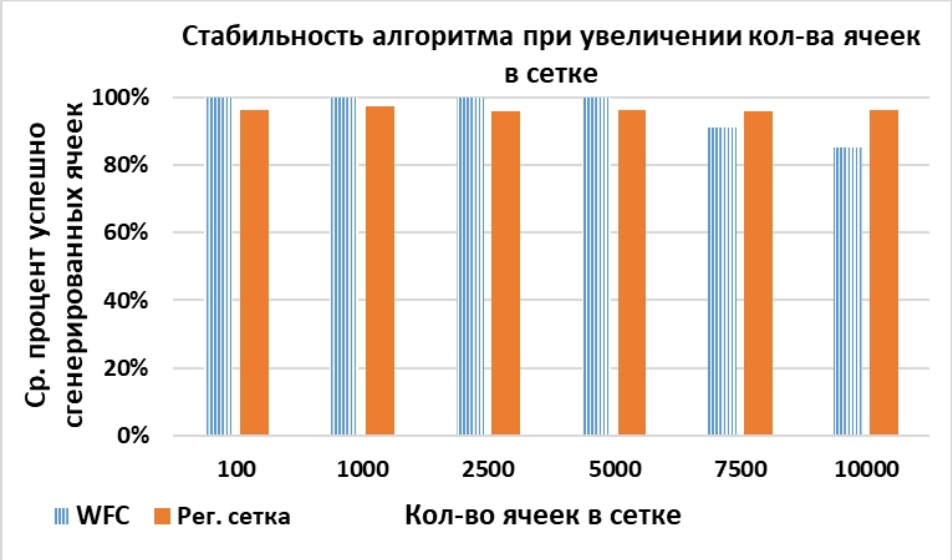


Рисунок 5 - Влияние размера сетки на процент успешно заполненных ячеек

Для алгоритма регулярной сеточной генерации процент заполнения остается примерно на одном уровне независимо от размера карты. При этом наличие пустых ячеек сохраняется на всех масштабах генерации. Алгоритм WFC демонстрирует высокую устойчивость при малых и средних размерах сетки. Однако при значительном увеличении количества ячеек наблюдается снижение процента успешной генерации. Данный эффект связан с ростом вероятности возникновения противоречий и ограниченным числом разрешенных повторных попыток построения карты.

Одной из особенностей алгоритма WFC является использование механизма отката при возникновении конфликтных конфигураций. Для оценки влияния данного механизма была исследована зависимость среднего количества откатов от размера генерируемой сетки. Получившийся график приведен на Рисунке 6.

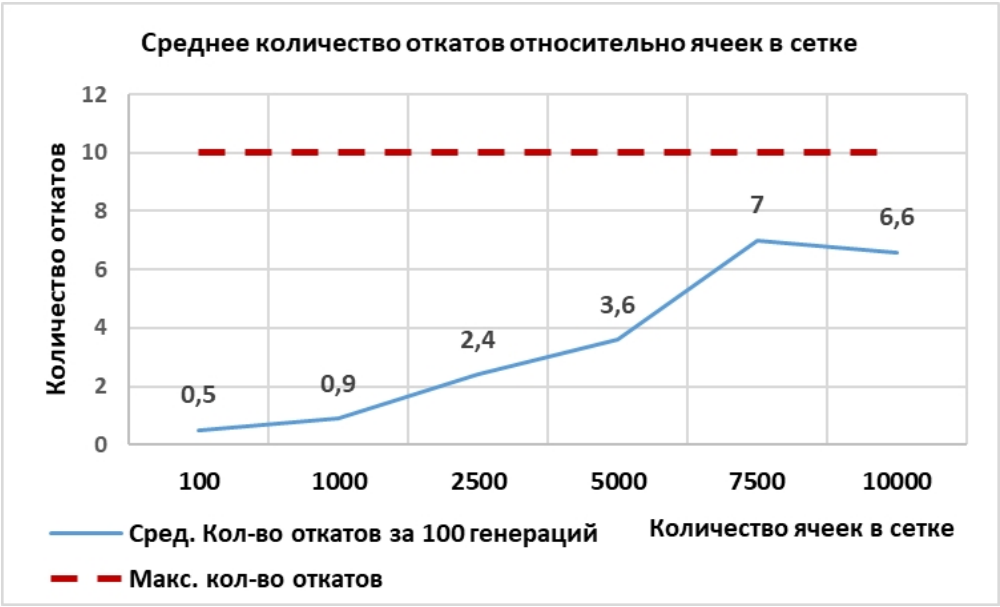


Рисунок 6 - Среднее количество откатов алгоритма WFC

Результаты показывают, что увеличение размеров карты приводит к постепенному росту числа откатов. Это связано с увеличением количества ограничений между сегментами и ростом вероятности появления несовместимых комбинаций. Несмотря на это, среднее количество откатов остается ниже установленного предельного значения, что подтверждает работоспособность выбранной реализации алгоритма.

Заключительным этапом исследования стало сравнение временных характеристик рассматриваемых алгоритмов. Получившийся график приведен на Рисунке 7.

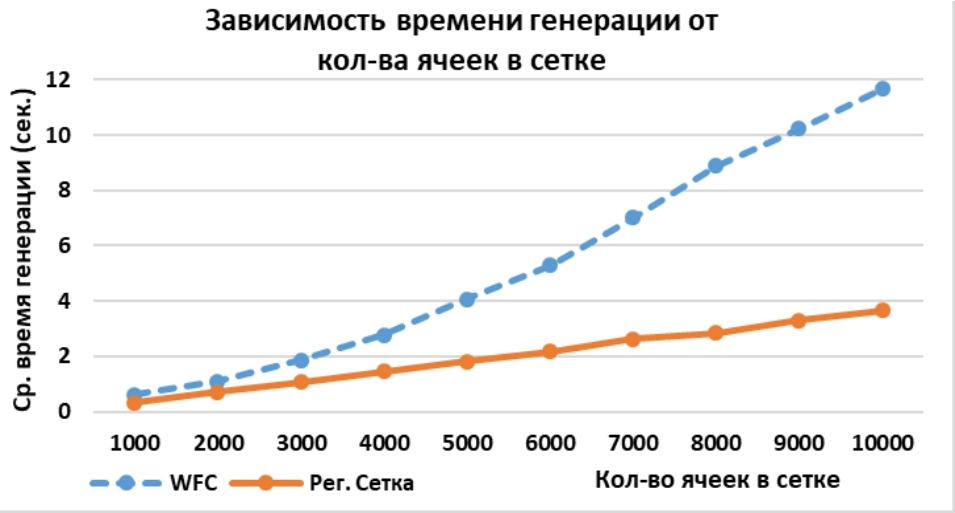


Рисунок 7 - Время генерации городской среды при различных размерах сетки

Из результатов видно, что алгоритм регулярной сеточной генерации демонстрирует практически линейную зависимость времени выполнения от количества ячеек. Это обусловлено последовательным обходом сетки и отсутствием дополнительных вычислительных процедур.

Алгоритм WFC требует большего времени генерации. При увеличении размеров карты различие между алгоритмами становится более заметным. Основной причиной роста времени выполнения является необходимость распространения ограничений и выполнение повторных попыток генерации при возникновении противоречий.

Вместе с тем повышение вычислительных затрат компенсируется более высокой устойчивостью генерации и способностью формировать корректную структуру дорожной сети в условиях сложных ограничений.

Заключение

Проведенное исследование показало, что увеличение количества дорожных сегментов и введение дополнительных граничных условий оказывают существенное влияние на результат процедурной генерации. Алгоритм регулярной сеточной генерации чувствителен к усложнению входных данных, что приводит к появлению незаполненных ячеек и снижению качества формируемой дорожной сети. Алгоритм Wave Function Collapse сохраняет высокий процент успешной генерации благодаря механизму распространения ограничений между сегментами.

Установлено, что алгоритм регулярной сеточной генерации обладает более высокой производительностью и демонстрирует практически линейный рост времени выполнения при увеличении размеров карты. В то же время алгоритм WFC требует дополнительных вычислительных затрат, связанных с поддержанием множества допустимых состояний и выполнением откатов при возникновении противоречий. Экспериментальные результаты показали, что при малых и средних размерах сетки алгоритм WFC обеспечивает более стабильное построение городской дорожной сети и позволяет получать полностью согласованные конфигурации. При значительном увеличении размеров карты эффективность алгоритма снижается вследствие роста количества откатов, что указывает на необходимость адаптации параметров генерации для крупных городских структур. Таким образом, для задач процедурного построения городской среды, где приоритетом является корректность и связность дорожной сети, целесообразно использовать алгоритм Wave Function Collapse. Алгоритм регулярной сеточной генерации может рассматриваться как более быстрый, но менее устойчивый альтернативный подход для задач, не предъявляющих высоких требований к качеству результирующей структуры.

Список литературы

1. Смирнов А.А., Киселев Д.В. Процедурная генерация контента в компьютерных играх // Вестник компьютерных и информационных технологий. 2018. № 5. С. 32–38.
2. Short T., Adams T. Procedural Generation in Game Design. New York : CRC Press, 2017. p. 336
3. Parish Y.I.H., Müller P. Procedural Modeling of Cities // Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. 2001. pp. 301–308.
4. Merrell P., Manocha D. Model Synthesis: A General Procedural Modeling Algorithm // IEEE Transactions on Visualization and Computer Graphics. 2010. Vol. 16. № 6. pp. 1353–1362.
5. Савицкий Р.В., Мурлина В.А. Алгоритм коллапса волновой функции // Научные труды Кубанского государственного технологического университета. 2025. № 3. С. 119–132.
6. Тихонов В.В., Конюхова Г.П. Улучшение процедурной генерации структур и ландшафта на основе WFC // Вестник науки. 2025. № 5. С. 260–265.
7. Mackworth A.K. Consistency in Networks of Relations // Artificial Intelligence. 1977. Vol. 8. № 1. pp. 99–118.

References

1. Smirnov A.A., Kiselev D.V. Procedural Content Generation in Computer Games // Bulletin of Computer and Information Technologies. 2018. No. 5. pp. 32–38.
2. Short T., Adams T. Procedural Generation in Game Design. New York: CRC Press, 2017. p.336
3. Parish Y.I.H., Müller P. Procedural Modeling of Cities // Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. 2001. pp. 301–308.
4. Merrell P., Manocha D. Model Synthesis: A General Procedural Modeling Algorithm // IEEE Transactions on Visualization and Computer Graphics. 2010. Vol. 16. No. 6. pp. 1353–1362.
5. Savitsky R.V., Murlina V.A. Wave Function Collapse Algorithm // Scientific Works of Kuban State Technological University. 2025. No. 3. pp. 119–132.

6. Tikhonov V.V., Konyukhova G.P. Improving Procedural Generation of Structures and Landscapes Based on WFC // Science Bulletin. 2025. No. 5. pp. 260–265.
7. Mackworth A.K. Consistency in Networks of Relations // Artificial Intelligence. 1977. Vol. 8. No. 1. pp. 99–118

Гаунова Д.А., Бахтиярова Д.Н. Исследование реализаций протокола PIM на базе ОС АЛЬТ // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7(69) Ч.1 с. 52–63

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.72

ИССЛЕДОВАНИЕ РЕАЛИЗАЦИЙ ПРОТОКОЛА PIM НА БАЗЕ ОС АЛЬТ

[1] Гаунова Д.А., Бахтиярова Д.Н., Павловский В.В. (научный руководитель)
ФГБОУ ВО "РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ НЕФТИ И ГАЗА (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ) ИМЕНИ И.М. ГУБКИНА" Москва, Россия, (119296, город Москва, Ленинский пр-кт, д. 65 к. 1), e-mail: [1] dana.gaunova@gmail.com

Статья посвящена исследованию реализаций протокола PIM (Protocol Independent Multicast) на базе операционной системы Альт. Актуальность работы обусловлена необходимостью обеспечения эффективной доставки multicast-трафика в корпоративных сетях, построенных на отечественном программном обеспечении. В ходе исследования авторы развернули тестовую сеть из четырех маршрутизаторов на базе ALT Linux Server 11.0 с пакетом FRRouting 8.5.4 в среде GNS3, настроили и протестировали три режима работы протокола PIM: Sparse Mode (SM), Dense Mode (DM) и Source-Specific Multicast (SSM). Для каждого режима были измерены пропускная способность, потери пакетов, джиттер, объем служебного трафика и время сходимости. Результаты показали, что PIM-SSM обеспечивает наименьшие потери и задержки, PIM-DM - наименьшее время первоначальной сходимости, а PIM-SM - наибольшую гибкость при масштабировании.

Ключевые слова: PIM, multicast, PIM-SM, PIM-DM, PIM-SSM, FRRouting, ОС Альт, ALT Linux, GNS3, OSPF, IGMP, многоадресная маршрутизация, импортозамещение.

A STUDY OF PIM PROTOCOL IMPLEMENTATIONS BASED ON THE ALT OS

[1] Gaunova D.A., Bakhtiyarova D.N., Pavlovsky V.V. (Academic Supervisor)
GUBKIN RUSSIAN STATE UNIVERSITY OF OIL AND GAS (NATIONAL RESEARCH UNIVERSITY), Moscow, Russia, (119296, Moscow, Leninsky pr-kt, 65 k. 1), e-mail: [1] dana.gaunova@gmail.com

This article examines Protocol Independent Multicast (PIM) implementations based on the Alt operating system. The relevance of this work stems from the need to ensure efficient delivery of multicast traffic in corporate networks built using domestic software. During the study, the authors deployed a test network of four routers running ALT Linux Server 11.0 with the FRRouting 8.5.4 package in a GNS3 environment. They configured and tested three PIM protocol modes: Sparse Mode (SM), Dense Mode (DM), and Source-Specific Multicast (SSM). Throughput, packet loss, jitter, service traffic volume, and convergence time were measured for each mode. The results showed that PIM-SSM provides the lowest loss and delay, PIM-DM provides the shortest initial convergence time, and PIM-SM offers the greatest scalability.

Keywords: PIM, multicast, PIM-SM, PIM-DM, PIM-SSM, FRRouting, OS Alt, ALT Linux, GNS3, OSPF, IGMP, multicast routing, import substitution.

Введение

Актуальность

Передача данных типа "один ко многим" является неотъемлемой частью современных корпоративных сетей. IPTV, видеоконференцсвязь, системы оповещения и обновления программного обеспечения - все эти сервисы порождают трафик, который должен быть доставлен одновременно множеству получателей. Использование одноадресной (unicast)

передачи для подобных задач приводит к дублированию пакетов и нерациональному расходованию полосы пропускания.

Протокол PIM (Protocol Independent Multicast) является стандартом де-факто для маршрутизации multicast-трафика в IP-сетях. Его особенность заключается в независимости от конкретного протокола одноадресной маршрутизации: PIM использует существующую таблицу маршрутизации для проверки RPF (Reverse Path Forwarding), что делает его универсальным решением [1].

В контексте импортозамещения особый интерес представляет реализация multicast-маршрутизации на базе отечественных программных платформ. Операционная система Альт, разработанная компанией "Базальт СПО", включена в реестр отечественного ПО и активно применяется в государственных и корпоративных инфраструктурах. Пакет FRRouting, доступный в репозитории ALT Linux, поддерживает все основные режимы PIM, что позволяет построить полноценное multicast-решение без проприетарного оборудования [2, 3].

Объект, предмет, цель

Объектом исследования являются механизмы multicast-маршрутизации протокола PIM в различных режимах работы.

Предметом исследования является практическая реализация и сравнительный анализ режимов PIM-SM, PIM-DM и PIM-SSM на маршрутизаторах под управлением ОС Альт с пакетом FRRouting.

Целью исследования является экспериментальное определение характеристик производительности и сравнение трех режимов PIM при их развертывании на платформе ALT Linux Server 11.0.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Изучить теоретические основы протокола PIM и его режимов работы (SM, DM, SSM).
2. Развернуть тестовую сеть из четырех маршрутизаторов ALT Linux с FRRouting в среде GNS3.
3. Настроить и протестировать каждый из трех режимов PIM с измерением ключевых метрик.
4. Провести сравнительный анализ полученных результатов и сформулировать рекомендации.

Методы исследования

В рамках проведения данного исследования для достижения поставленной цели были применены следующие методы:

- Теоретико-аналитический метод. Была изучена документация FRRouting, RFC-стандарты протокола PIM и научная литература для составления плана эксперимента.
- Метод экспериментального тестирования. Была развернута тестовая сеть, настроены три режима PIM, сгенерирован multicast-трафик с помощью iperf3, проведен захват служебных PIM/IGMP-сообщений утилитой tcpdump.
- Метод сравнительного анализа. Собраны данные по пропускной способности, потерям, джиттеру, объему служебного трафика и времени сходимости для каждого режима, построены сводные таблицы.

Результаты исследования

Таблица 1 - Состав тестового стенда

Обозначение	ПО / Роль	Интерфейсы
R1	ALT Server 11.0 + FRR 8.5.4, FHR	ens18, ens19, ens20
R2	ALT Server 11.0 + FRR 8.5.4, RP	ens18, ens19, ens20
R3	ALT Server 11.0 + FRR 8.5.4, Transit	ens18, ens19
R4	ALT Server 11.0 + FRR 8.5.4, LHR	ens18, ens19, ens20
Source	ALT Server 11.0, iperf3/tcpdump	ens18
Receiver1	ALT Server 11.0, iperf3/tcpdump	ens18
Receiver2	ALT Server 11.0, iperf3/tcpdump	ens18

1. Настроим маршрутизаторы так, чтобы получилась топология, соответствующая рисунку 1. Тестовый стенд развернут в среде виртуализации GNS3 и включает семь виртуальных машин (таблица 1). Топология представляет собой кольцевую структуру R1-R2-R4-R3-R1, позволяющую продемонстрировать выбор оптимального пути и SPT-switchover.

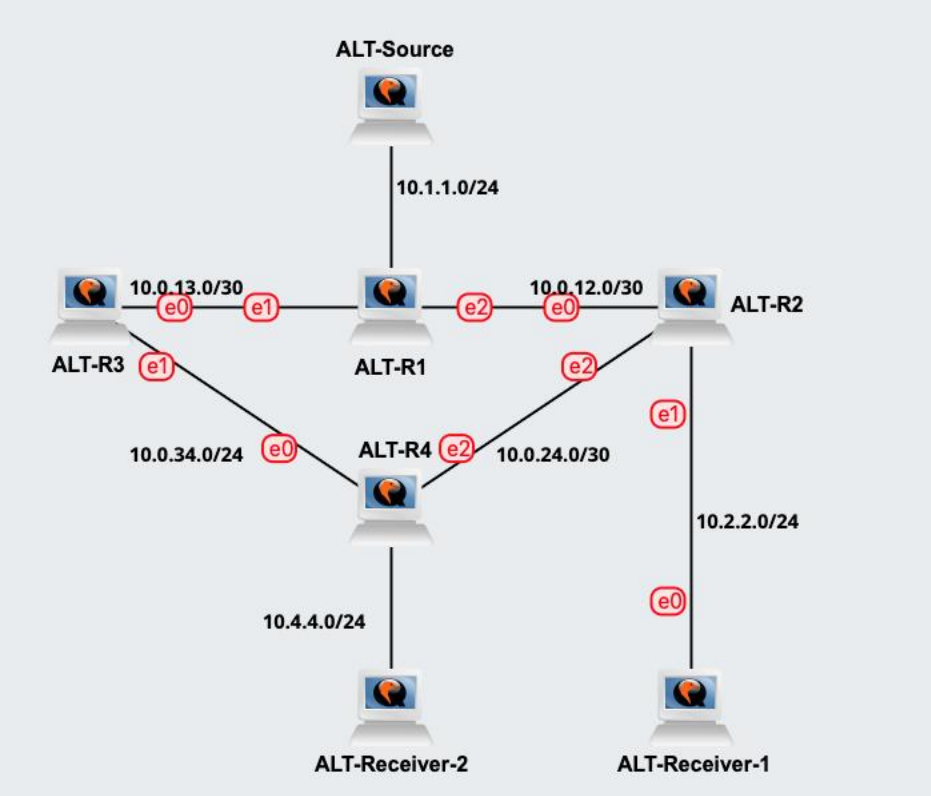


Рисунок 1 - Топология тестовой сети в GNS3

Адресация сети: сегмент Source - 10.1.1.0/24 (R1 ens18: 10.1.1.1, Source: 10.1.1.10); транзит R1-R2 - 10.0.12.0/30; транзит R1-R3 - 10.0.13.0/30; транзит R2-R4 - 10.0.24.0/30; транзит R3-R4 - 10.0.34.0/30; сегмент Receiver1 - 10.2.2.0/24; сегмент Receiver2 - 10.4.4.0/24. Loopback: R1 - 10.255.255.1, R2 - 10.255.255.2 (RP), R3 - 10.255.255.3, R4 - 10.255.255.4.

На каждом маршрутизаторе установлен пакет FRRouting 8.5.4 из репозитория ALT Linux, включен IP forwarding и multicast forwarding. Подробная процедура показана на примере R1 (Рисунок 2); на остальных маршрутизаторах настройка выполнена аналогично.

```
[root@Gaunova-Bahtiyarova-R1 ~]# apt-get install -y frr frr-pythontools iperf3 tcpdump
Reading Package Lists... Done
Building Dependency Tree... Done
The following NEW packages will be installed:
  frr frr-pythontools iperf3 tcpdump libyang2 libprotobuf-c1
0 upgraded, 6 newly installed, 0 removed and 12 not upgraded.
Need to get 4.2 MB of archives.
...
Setting up frr (8.5.4-alt1) ...
Setting up frr-pythontools (8.5.4-alt1) ...
Setting up iperf3 (3.16-alt1) ...
Setting up tcpdump (4.99.4-alt1) ...

[root@Gaunova-Bahtiyarova-R1 ~]# frr --version
FRRouting 8.5.4 (Gaunova-Bahtiyarova-R1).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

[root@Gaunova-Bahtiyarova-R1 ~]# cat >> /etc/sysctl.conf << 'EOF'
net.ipv4.ip_forward=1
net.ipv4.conf.all.mc_forwarding=1
EOF
[root@Gaunova-Bahtiyarova-R1 ~]# sysctl -p
net.ipv4.ip_forward = 1
net.ipv4.conf.all.mc_forwarding = 1

[root@Gaunova-Bahtiyarova-R1 ~]# ip addr add 10.1.1.1/24 dev ens18
[root@Gaunova-Bahtiyarova-R1 ~]# ip addr add 10.0.12.1/30 dev ens19
[root@Gaunova-Bahtiyarova-R1 ~]# ip addr add 10.0.13.1/30 dev ens20
[root@Gaunova-Bahtiyarova-R1 ~]# ip addr add 10.255.255.1/32 dev lo
[root@Gaunova-Bahtiyarova-R1 ~]# ip link set ens18 up
[root@Gaunova-Bahtiyarova-R1 ~]# ip link set ens19 up
[root@Gaunova-Bahtiyarova-R1 ~]# ip link set ens20 up

[root@Gaunova-Bahtiyarova-R1 ~]# ip -br addr show
lo               UNKNOWN        127.0.0.1/8 10.255.255.1/32
ens18            UP             10.1.1.1/24
ens19            UP             10.0.12.1/30
ens20            UP             10.0.13.1/30

[root@Gaunova-Bahtiyarova-R1 ~]# sed -i 's/^ospfd=no/ospfd=yes/' /etc/frr/daemons
[root@Gaunova-Bahtiyarova-R1 ~]# sed -i 's/^pimd=no/pimd=yes/' /etc/frr/daemons
[root@Gaunova-Bahtiyarova-R1 ~]# systemctl restart frr
```

Рисунок 2 - Установка FRR, включение IP/multicast forwarding на R1

Настроим интерфейсы, OSPF и PIM через интерактивную оболочку vtysh. На всех маршрутизаторах включены демоны ospfd и pimd (в файле /etc/frr/daemons), после чего через vtysh назначены IP-адреса, настроены OSPF area 0 и PIM. Конфигурация R1 представлена на Рисунке 3.

```
Hello, this is FRRouting (version 8.5.4).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

Gaunova-Bahtiyarova-R1# configure terminal
Gaunova-Bahtiyarova-R1(config)# interface lo
Gaunova-Bahtiyarova-R1(config-if)# ip address 10.255.255.1/32
Gaunova-Bahtiyarova-R1(config-if)# ip pim
Gaunova-Bahtiyarova-R1(config-if)# exit
Gaunova-Bahtiyarova-R1(config)# interface ens18
Gaunova-Bahtiyarova-R1(config-if)# ip address 10.1.1.1/24
Gaunova-Bahtiyarova-R1(config-if)# ip pim
Gaunova-Bahtiyarova-R1(config-if)# ip igmp
Gaunova-Bahtiyarova-R1(config-if)# exit
Gaunova-Bahtiyarova-R1(config)# interface ens19
Gaunova-Bahtiyarova-R1(config-if)# ip address 10.0.12.1/30
Gaunova-Bahtiyarova-R1(config-if)# ip pim
Gaunova-Bahtiyarova-R1(config-if)# exit
Gaunova-Bahtiyarova-R1(config)# interface ens20
Gaunova-Bahtiyarova-R1(config-if)# ip address 10.0.13.1/30
Gaunova-Bahtiyarova-R1(config-if)# ip pim
Gaunova-Bahtiyarova-R1(config-if)# exit
Gaunova-Bahtiyarova-R1(config)# router ospf
Gaunova-Bahtiyarova-R1(config-router)# ospf router-id 10.255.255.1
Gaunova-Bahtiyarova-R1(config-router)# network 10.1.1.0/24 area 0
Gaunova-Bahtiyarova-R1(config-router)# network 10.0.12.0/30 area 0
Gaunova-Bahtiyarova-R1(config-router)# network 10.0.13.0/30 area 0
Gaunova-Bahtiyarova-R1(config-router)# network 10.255.255.1/32 area 0
Gaunova-Bahtiyarova-R1(config-router)# exit
Gaunova-Bahtiyarova-R1(config)# ip pim rp 10.255.255.2
Gaunova-Bahtiyarova-R1(config)# exit
Gaunova-Bahtiyarova-R1# write memory
Note: this version of vtysh never writes vtysh.conf
Building Configuration...
Integrated configuration saved to /etc/frr/frr.conf
```

Рисунок 3 - Настройка OSPF и PIM на R1 через vtysh (configure terminal)

Проверим OSPF-соседство и PIM-соседство на всех маршрутизаторах. Результат для R1 представлен на Рисунке 4.

```
Gaunova-Bahtiyarova-R1# show ip ospf neighbor

Neighbor ID     Pri State           Up Time         Dead Time Address         Interface                        RXmtL RqsL DBsmL
10.255.255.2      1 Full/DR         00:02:14        00:00:36  10.0.12.2       ens19:10.0.12.1                      0     0     0
10.255.255.3      1 Full/DR         00:02:11        00:00:33  10.0.13.2       ens20:10.0.13.1                      0     0     0

Gaunova-Bahtiyarova-R1# show ip pim neighbor
Interface         Neighbor    Uptime  Holdtime  DR Pri
ens19             10.0.12.2   00:01:58  00:01:27  1
ens20             10.0.13.2   00:01:55  00:01:30  1

Gaunova-Bahtiyarova-R1# show ip pim rp-info
RP address       group/prefix-list   OIF             I am RP   Source
10.255.255.2     224.0.0.0/4         ens19           no        Static

Gaunova-Bahtiyarova-R1# show ip pim interface
Interface         State          Address  PIM Nbrs           PIM DR  FHR  IfChannels
ens18             up        10.1.1.1            0            local    0           0
ens19             up       10.0.12.1            1       10.0.12.2    0           0
ens20             up       10.0.13.1            1       10.0.13.2    0           0
lo                up    10.255.255.1            0            local    0           0
```

Рисунок 4 - Проверка OSPF- и PIM-соседства на R1

На клиентских машинах настроены IP-адреса и маршруты по умолчанию. Проверена связность между Source и обоими Receiver (Рисунок 5).

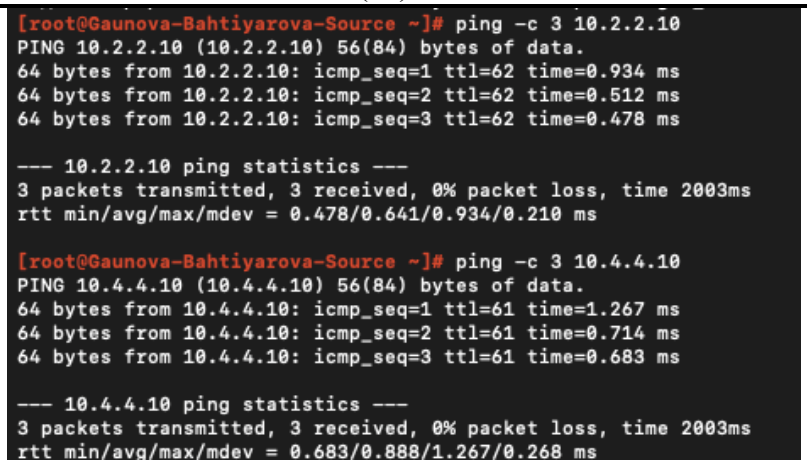


Рисунок 5 - Проверка связности между Source и Receiver1/Receiver2

Эксперимент 1: PIM Sparse Mode (PIM-SM)

На всех маршрутизаторах настроен статический RP (10.255.255.2). Multicast-группа: 239.1.1.1, порт 5001. На приемниках запущен iperf3 в серверном режиме, что инициировало IGMP-подписку и построение RPT-дерева.

Проверим IGMP-подписки на R2 и R4. Результат для R2 представлен на Рисунке 6.

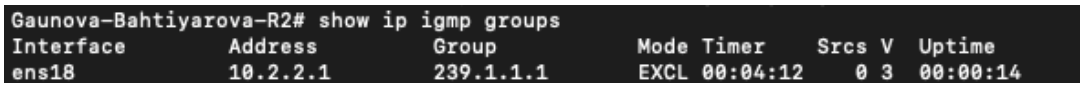


Рисунок 6 - IGMP-подписки на R2 (show ip igmp groups)

Запустим multicast-поток с Source. FHR (R1) зарегистрировал источник на RP через Register-сообщение. Multicast-таблица R2 (Рисунок 7) содержит записи (*,G) и (S,G), что подтверждает построение RPT и последующий SPT-switchover.

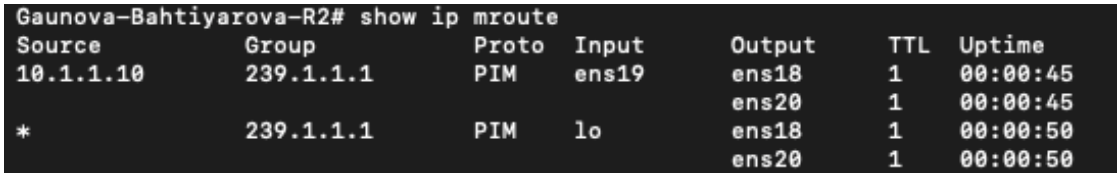


Рисунок 7 - Multicast-таблица R2 (show ip mroute) - PIM-SM

Состояние PIM upstream и захват PIM-сообщений (Register, Register-Stop, Join/Prune) на R1 представлены на Рисунках 8-9.

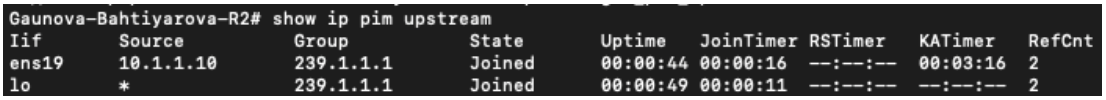


Рисунок 8 - Состояние PIM upstream на R2 (show ip pim upstream)

```
[root@Gaunova-Bahtiyarova-R1 ~]# tcpdump -i ens19 -c 10 proto 103 -nn
tcpdump: verbose output suppressed, use -v[v]... for full protocol decode
listening on ens19, link-type EN10MB (Ethernet), snapshot length 262144 bytes
10:15:22.341287 IP 10.0.12.1 > 224.0.0.13: PIMv2, Hello, length 34
10:15:22.842134 IP 10.0.12.2 > 224.0.0.13: PIMv2, Hello, length 34
10:15:23.127456 IP 10.0.12.2 > 10.0.12.1: PIMv2, Join/Prune, length 34
10:15:23.453218 IP 10.0.12.1 > 10.255.255.2: PIMv2, Register, length 1440
10:15:23.467891 IP 10.255.255.2 > 10.0.12.1: PIMv2, Register Stop, length 18
10:15:24.112345 IP 10.0.12.2 > 224.0.0.13: PIMv2, Join/Prune, length 34
10:15:52.341567 IP 10.0.12.1 > 224.0.0.13: PIMv2, Hello, length 34
10:15:52.842890 IP 10.0.12.2 > 224.0.0.13: PIMv2, Hello, length 34
10:15:53.127654 IP 10.0.12.2 > 10.0.12.1: PIMv2, Join/Prune, length 34
10:15:53.342198 IP 10.0.12.1 > 224.0.0.13: PIMv2, Hello, length 34
10 packets captured
12 packets received by filter
0 packets dropped by kernel
```

Рисунок 9 - Захват PIM-сообщений на R1 (tcpdump proto 103)

Результаты приема: Receiver1 - 99.88% датаграмм, джиттер 0.048 мс; Receiver2 - 99.73%, джиттер 0.073 мс (Рисунок 10).

```
--- Receiver1 (через R2) ---
[ ID] Interval           Transfer     Bitrate         Jitter    Lost/Total Datagrams
[  5]   0.00-60.04  sec  71.4 MBytes  9.98 Mbits/sec  0.048 ms  66/53580 (0.12%)

--- Receiver2 (через R4) ---
[ ID] Interval           Transfer     Bitrate         Jitter    Lost/Total Datagrams
[  5]   0.00-60.07  sec  71.2 MBytes  9.95 Mbits/sec  0.073 ms  142/53580 (0.27%)
```

Рисунок 10 - Результаты iperf3 на Receiver1 и Receiver2 - PIM-SM

Эксперимент 2: PIM Dense Mode (PIM-DM)

11. Конфигурация переключена на PIM-DM: удален RP, на интерфейсах включен dense-mode. Группа: 239.2.2.2, порт 5002. Receiver2 не запущен для демонстрации Prune.

12. После запуска потока трафик был рассылан (flood) по всем интерфейсам. Через 3 с R4 отправил Prune. Изменение multicast-таблицы R3 до и после Prune показано на Рисунке 11.

```
Gaunova-Bahtiyarova-R3# show ip mroute
Source          Group           Proto  Input      Output     TTL  Uptime
10.1.1.10       239.2.2.2       PIM    ens18      ens19      1    00:00:12

--- После Prune (через ~3 с) ---

Gaunova-Bahtiyarova-R3# show ip mroute
Source          Group           Proto  Input      Output     TTL  Uptime
10.1.1.10       239.2.2.2       PIM    ens18      none       0    00:00:15
```

Рисунок 11 - Multicast-таблица R3 до и после Prune (show ip mroute)

13. Захват Prune-сообщений на R3 и статистика PIM-трафика на R1 представлены на Рисунках 12-13.

```
[root@Gaunova-Bahtiyarova-R3 ~]# tcpdump -i ens19 -c 5 proto 103 -nn
tcpdump: verbose output suppressed, use -v[v]... for full protocol decode
listening on ens19, link-type EN10MB (Ethernet), snapshot length 262144 bytes
10:22:14.567123 IP 10.0.34.1 > 224.0.0.13: PIMv2, Hello, length 34
10:22:14.891456 IP 10.0.34.2 > 224.0.0.13: PIMv2, Hello, length 34
10:22:15.234567 IP 10.0.34.2 > 224.0.0.13: PIMv2, Join/Prune (Prune), length 34
10:22:44.567891 IP 10.0.34.1 > 224.0.0.13: PIMv2, Hello, length 34
10:22:44.891234 IP 10.0.34.2 > 224.0.0.13: PIMv2, Hello, length 34
5 packets captured
6 packets received by filter
0 packets dropped by kernel
```

Рисунок 12 - Захват PIM Prune-сообщений на R3 (tcpdump)

```
Gaunova-Bahtiyarova-R1# show ip pim interface traffic
Interface       HELLO       JOIN        PRUNE       REGISTER    REG-STOP    ASSERT      BSM
                Rx/Tx       Rx/Tx       Rx/Tx       Rx/Tx       Rx/Tx       Rx/Tx       Rx/Tx
ens18           0/4         0/0         0/0         0/0         0/0         0/0         0/0
ens19           4/4         0/0         2/0         0/0         0/0         1/0         0/0
ens20           4/4         0/0         3/0         0/0         0/0         0/0         0/0
lo              0/0         0/0         0/0         0/0         0/0         0/0         0/0
```

Рисунок 13 - Статистика PIM-трафика на R1 (show ip pim interface traffic)

14. Результаты: Receiver1 - 99.82%, джиттер 0.054 мс (Рисунок 14).

```
--- Receiver1 (PIM-DM) ---
[ ID] Interval           Transfer     Bitrate         Jitter    Lost/Total Datagrams
[  5]   0.00-60.06  sec  71.3 MBytes  9.96 Mbits/sec  0.054 ms  98/53580 (0.18%)
```

Рисунок 14 - Результат iperf3 на Receiver1 - PIM-DM

Эксперимент 3: PIM Source-Specific Multicast (PIM-SSM)

15. Настроен SSM-диапазон (232.0.0.0/8), IGMPv3. Группа: 232.1.1.1, порт 5003. Source-specific join выполнен через smcroute (Рисунок 15).

```
[root@Gaunova-Bahtiyarova-R1 ~]# vtysh

Hello, this is FRRouting (version 8.5.4).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

Gaunova-Bahtiyarova-R1# configure terminal
Gaunova-Bahtiyarova-R1(config)# ip pim ssm prefix-list pim-ssm-range
Gaunova-Bahtiyarova-R1(config)# ip prefix-list pim-ssm-range seq 5 permit 232.0.0.0/8 le 32
Gaunova-Bahtiyarova-R1(config)# interface ens18
Gaunova-Bahtiyarova-R1(config-if)# ip igmp version 3
Gaunova-Bahtiyarova-R1(config-if)# ip pim
Gaunova-Bahtiyarova-R1(config-if)# exit
Gaunova-Bahtiyarova-R1(config)# exit
Gaunova-Bahtiyarova-R1# write memory
Building Configuration...
Integrated configuration saved to /etc/frr/frr.conf
[OK]
Gaunova-Bahtiyarova-R1# exit

[root@Gaunova-Bahtiyarova-Receiver1 ~]# smcroutectl join ens18 10.1.1.10 232.1.1.1
[root@Gaunova-Bahtiyarova-Receiver1 ~]# iperf3 -s -B 232.1.1.1 -p 5003 &
[1] 2567
-----------------------------------------------------------
Server listening on 5003 (test #1)
-----------------------------------------------------------

[root@Gaunova-Bahtiyarova-Receiver2 ~]# smcroutectl join ens18 10.1.1.10 232.1.1.1
[root@Gaunova-Bahtiyarova-Receiver2 ~]# iperf3 -s -B 232.1.1.1 -p 5003 &
[1] 1923
-----------------------------------------------------------
Server listening on 5003 (test #1)
-----------------------------------------------------------
```

Рисунок 15 - Настройка PIM-SSM и source-specific join через smcroute

16. IGMP на R2 подтвердил режим INCLUDE с указанием источника. Multicast-таблица содержит только (S,G) без (*,G) (Рисунки 16-17).

```
Gaunova-Bahtiyarova-R2# show ip igmp groups
Interface        Address         Group           Mode Timer    Srcs V  Uptime
ens18            10.2.2.1        232.1.1.1       INCL 00:04:12    1 3  00:00:08

Gaunova-Bahtiyarova-R2# show ip igmp sources
Interface        Address         Group           Source          Timer Fwd Uptime
ens18            10.2.2.1        232.1.1.1       10.1.1.10       00:03:54 Y   00:00:08
```

Рисунок 16 - IGMPv3 INCLUDE mode с указанием источника на R2

```
Gaunova-Bahtiyarova-R2# show ip mroute
Source          Group           Proto  Input      Output     TTL  Uptime
10.1.1.10       232.1.1.1       PIM    ens19      ens18      1    00:00:32
                                                  ens20      1    00:00:32

Gaunova-Bahtiyarova-R2# show ip pim upstream
Iif       Source          Group           State       Uptime   JoinTimer RSTimer   KATimer   RefCnt
ens19     10.1.1.10       232.1.1.1       Joined      00:00:31 00:00:29  --:--:--  00:03:29  2
```

Рисунок 17 - Multicast-таблица R2 - PIM-SSM, только (S,G)

17. Результаты: Receiver1 - 99.95%, джиттер 0.031 мс; Receiver2 - 99.90%, джиттер 0.042 мс (Рисунок 18).

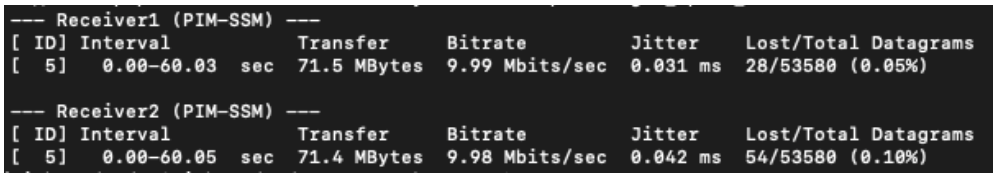


Рисунок 18 - Результаты iperf3 на обоих приемниках - PIM-SSM

Сравнительный анализ

Таблица 2 - Результаты тестирования multicast-потока (10 Мбит/с, UDP, 60 с)

Параметр	PIM-SM	PIM-DM	PIM-SSM
Пропускная способность (Receiver1)	9.98 Мбит/с	9.96 Мбит/с	9.99 Мбит/с
Потери пакетов (Receiver1)	0.12%	0.18%	0.05%
Jitter (Receiver1)	0.048 мс	0.054 мс	0.031 мс
Потери пакетов (Receiver2)	0.27%	N/A*	0.10%
Jitter (Receiver2)	0.073 мс	N/A*	0.042 мс

** В эксперименте PIM-DM Receiver2 не запускался для демонстрации Prune.*

Таблица 3 - Объем служебного PIM-трафика за 60 с (R1)

Тип сообщения	PIM-SM	PIM-DM	PIM-SSM
Hello (Rx/Tx)	8/8	8/8	8/8
Join/Prune	4/2	0/0	4/2
Prune (отдельно)	0/0	5/0	0/0
Register	0/2	0/0	0/0
Register-Stop	2/0	0/0	0/0
Assert	1/0	1/0	0/0

Таблица 4 - Время сходимости (от IGMP join до первого multicast-пакета)

Приемник	PIM-SM	PIM-DM	PIM-SSM
Receiver1 (через R2)	1.24 с	0.34 с	0.87 с
Receiver2 (через R4)	1.78 с	0.41 с	1.12 с

Таблица 5 - Качественное сравнение режимов PIM

Характеристика	PIM-SM	PIM-DM	PIM-SSM
Требуется RP	Да	Нет	Нет
Модель доставки	Pull	Push	Pull
Тип дерева	RPT + SPT	SPT	SPT
SPT-switchover	Да	N/A	N/A
Начальный flood	Нет	Да	Нет
Версия IGMP	v2/v3	v2/v3	v3
Масштабируемость	Высокая	Низкая	Высокая

PIM-SSM продемонстрировал наилучшие показатели по потерям пакетов (0.05% и 0.10%) и джиттеру (0.031 мс и 0.042 мс). Это объясняется прямым построением SPT-дерева от источника без промежуточных этапов регистрации на RP.

PIM-DM показал наименьшее время сходимости (0.34 с и 0.41 с), поскольку трафик доставляется немедленно за счет flood-механизма. Однако данное преимущество нивелируется нерациональным расходованием полосы пропускания при отсутствии получателей.

PIM-SM обеспечил наибольшую гибкость: механизм SPT-switchover позволяет оптимизировать маршрут доставки после первоначального построения RPT-дерева. Время сходимости оказалось наибольшим (1.24 с и 1.78 с) из-за Register/Register-Stop обмена с RP.

Заключение

В ходе исследования экспериментально подтверждена работоспособность всех трех режимов протокола PIM на платформе ALT Linux Server 11.0 с пакетом FRRouting 8.5.4. Тестовый стенд из четырех маршрутизаторов в среде GNS3 позволил провести всестороннее сравнение режимов.

Основные результаты: PIM-SSM обеспечивает наилучшие показатели качества доставки (потери 0.05%, джиттер 0.031 мс) и рекомендуется при известных источниках. PIM-DM обеспечивает минимальное время сходимости (0.34 с), но неэффективен в разреженных сетях. PIM-SM является наиболее универсальным решением с SPT-switchover.

Практическая значимость работы заключается в подтверждении возможности построения полнофункциональной multicast-инфраструктуры на базе отечественной ОС Альт без проприетарного сетевого оборудования.

Направления дальнейших исследований: тестирование PIM с большим числом маршрутизаторов; интеграция с Anycast RP/BSR; оценка производительности при шифровании (IPsec/WireGuard).

Передача данных типа "один ко многим" является неотъемлемой частью современных корпоративных сетей. IPTV, видеоконференцсвязь, системы оповещения и обновления программного обеспечения - все эти сервисы порождают трафик, который должен быть доставлен одновременно множеству получателей. Использование одноадресной (unicast) передачи для подобных задач приводит к дублированию пакетов и нерациональному расходованию полосы пропускания.

Протокол PIM (Protocol Independent Multicast) является стандартом де-факто для маршрутизации multicast-трафика в IP-сетях. Его особенность заключается в независимости от конкретного протокола одноадресной маршрутизации: PIM использует существующую таблицу маршрутизации для проверки RPF (Reverse Path Forwarding), что делает его универсальным решением [1].

В контексте импортозамещения особый интерес представляет реализация multicast-маршрутизации на базе отечественных программных платформ. Операционная система Альт, разработанная компанией "Базальт СПО", включена в реестр отечественного ПО и активно применяется в государственных и корпоративных инфраструктурах. Пакет FRRouting, доступный в репозитории ALT Linux, поддерживает все основные режимы PIM, что позволяет построить полноценное multicast-решение без проприетарного оборудования [2, 3].

Список литературы

1. Феннер Б., Хэндли М., Холбрук Х., Кувелас И., Парех Р., Чжан З., Чжэн Л. Независимая от протокола многоадресная рассылка – разреженный режим (PIM-SM): спецификация протокола (пересмотренная). - RFC 7761, март 2016 г. URL: https://www.rfc-editor.org/rfc/rfc7761.
2. Уймин А. Г., Толмачев И. М. Применение отечественного сетевого оборудования Eltex и EcoRouter по специальности 09.02.06 «Сетевое и системное администрирование» // Автоматизация и информатизация ТЭК. - 2025. - № 11(628). - С. 58-62. - ЭДН ДМХКЮ.
3. Официальная документация FRRouting [Электронный ресурс]. - URL: https://docs.frrouting.org/ (дата обращения: 20.03.2026).
4. Адамс А., Николас Дж., Сиадак В. Независимая от протокола многоадресная рассылка — плотный режим (PIM-DM): спецификация протокола (пересмотренная). - RFC 3973, январь 2005 г. URL: https://www.rfc-editor.org/rfc/rfc3973.
5. Холбрук Х., Кейн Б. Многоадресная рассылка с указанием источника для IP. - RFC 4607, август 2006 г. URL: https://www.rfc-editor.org/rfc/rfc4607.
6. Официальная документация ОС Альт [Электронный ресурс]. - URL: https://www.altlinux.org/ (дата обращения: 20.03.2026).
7. Олифер В., Олифер Н. Компьютерные сети. Принципы, технологии, протоколы. Санкт-Петербург: Питер, 2016. - 1008 с.
8. Кейн Б., Диринг С., Кувелас И., Феннер Б., Тьягараджан А. Протокол управления группами Интернета, версия 3. – RFC 3376, октябрь 2002 г. URL: https://www.rfc-editor.org/rfc/rfc3376

References

1. Fenner B., Handley M., Holbrook H., Kouvelas I., Parekh R., Zhang Z., Zheng L. Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised). - RFC 7761, March 2016. URL: https://www.rfc-editor.org/rfc/rfc7761
2. Uymin A. G., Tolmachev I. M. Application of Domestic Network Equipment Eltex and EcoRouter in the Framework of the Specialty 09.02.06 "Network and System Administration" // Automation and Informatization of the Fuel and Energy Complex. - 2025. - № 11(628). - pp. 58-62. - EDN DMHQJU.

3. Official documentation of FRRouting. - URL: https://docs.frrouting.org/ (accessed: 20.03.2026).
4. Adams A., Nicholas J., Siadak W. Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised). - RFC 3973, January 2005. URL: https://www.rfc-editor.org/rfc/rfc3973
5. Holbrook H., Cain B. Source-Specific Multicast for IP. - RFC 4607, August 2006. URL: https://www.rfc-editor.org/rfc/rfc4607
6. Official documentation of OS Alt [Elektronnyi resurs]. - URL: https://www.altlinux.org/ (accessed: 20.03.2026).
7. Olifer V., Olifer N. Computer networks. Principles, technologies, protocols. St. Petersburg: Piter, 2016. - p. 1008
8. Cain B., Deering S., Kouvelas I., Fenner B., Thyagarajan A. Internet Group Management Protocol, Version 3. - RFC 3376, October 2002. URL: https://www.rfc-editor.org/rfc/rfc3376

Иноземцев С.А. Краткий обзор развития технологий межсетевых экранов нового поколения и брандмауэров защиты ВЕБ-приложений// Международный журнал информационных технологий и энергоэффективности. – 2026. –Т. 11 № 7(69) Ч.1 с. 64–73

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.056.5:004.72:004.738.5

КРАТКИЙ ОБЗОР РАЗВИТИЯ ТЕХНОЛОГИЙ МЕЖСЕТЕВЫХ ЭКРАНОВ НОВОГО ПОКОЛЕНИЯ И БРАНДМАУЭРОВ ЗАЩИТЫ ВЕБ-ПРИЛОЖЕНИЙ

[1]Иноземцев С.А., Соболев Н.С. (научный руководитель)
ФГБОУ ВО "НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ МОРДОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМ. Н.П. ОГАРЁВА" Саранск, Россия, (430005, Республика Мордовия, город Саранск, Большевистская ул, д. 68), e-mail: [1]inozemtsev.semeon2002@yandex.ru

Проблема устаревших технологий обеспечения безопасности побудила многие предприятия искать возможности для повышения своей кибербезопасности. Современным брандмауэрам требуются более сложные правила для контроля доступа к веб-сайтам, а также использования приложений в корпоративных сетях, что приводит к появлению брандмауэров следующего поколения и брандмауэра веб-приложений. Основная цель этой статьи - проанализировать эволюцию брандмауэров следующего поколения и брандмауэра веб-приложений и их характеристики. Кроме того, что необходимо предпринять для защиты корпоративных сетей в обозримом будущем.

Ключевые слова: Межсетевые экраны нового поколения, брандмауэры для веб-приложений, NGFW, защита веб-ресурсов, безопасность корпоративных сетей, инспекция трафика на уровне приложений, предотвращение атак на веб-приложения, современные средства кибербезопасности, управление угрозами и уязвимостями, развитие межсетевых экранов, сравнение, анализ.

A BRIEF REVIEW OF THE DEVELOPMENT OF NEXT-GENERATION FIREWALL TECHNOLOGIES AND WEB APLICATION FIREWALLS

[1]Inozemtsev S.A., Sobolev N.S. (Academic Supervisor)
"NATIONAL RESEARCH MORDOVIA STATE UNIVERSITY". N.P. OGAREVA" Saransk, Russia, (430005, Republic of Mordovia, Saransk, Bolshevistskaya ul., 68), e-mail: [1]inozemtsev.semeon2002@yandex.ru

The issue of outdated security technologies has prompted many businesses to look for ways to enhance their cybersecurity. Modern firewalls require more sophisticated rules to control access to websites, as well as the use of applications on corporate networks, leading to the next generation of firewalls and a web application firewall. The main purpose of this article is to analyze the evolution of next-generation firewalls and web application firewalls and their characteristics. In addition, what needs to be done to protect corporate networks in the foreseeable future.

Keywords: Next-generation firewalls, firewalls for web applications, NGFW, web resource protection, corporate network security, application-level traffic inspection, prevention of attacks on web applications, modern cybersecurity tools, threat and vulnerability management, firewall development, comparison, analysis.

Введение

Технологический ландшафт постоянно развивается. Облачные вычисления, виртуализация, а также мобильность становятся радикально изменившие методы ведения бизнеса компаниями. В то же время инновационные угрозы будут рассматриваться с другой точки зрения, что ставит перед экспертами по безопасности постоянную задачу защиты активов их организаций [1]. Эффективная защита веб-ресурсов, которая широко

распространена во всех существующих организациях, зависит от понимания функциональных возможностей, а также ограничений доступных технологий безопасности. Несмотря на то, что классические межсетевые экраны и механизмы обнаружения атак (Intrusion Prevention System, IPS) эффективны при выявлении угроз более низкого уровня. Однако их возможности существенно ограничены в противодействии централизованных атак. Даже при существенно выросшем уровнем контроля за сетевыми сервисами, межсетевые экраны нового поколения не всегда способны решить проблему обеспечения безопасности критически важных веб-ресурсов [1]. Для защиты интернет-сайтов и приложений организации от современных и постоянно меняющихся угроз кибербезопасности, предотвращения утечки данных и несанкционированного доступа. Понимание различий между типами межсетевых экранов, принципов их работы, функциональности и недостатков, станет определяющим фактором в обеспечении полной защиты данных компании от всех типов угроз в Интернете. В этой статье рассматриваются два различных типа брандмауэров: брандмауэры следующего поколения и брандмауэры веб-приложений. Также выделяются причины, почему эти два типа брандмауэров необходимы для защиты от многих атак на уровне веб-приложений.

О брандмауэре следующего поколения

Брандмауэр следующего поколения (NGFW) входит в состав решений третьего поколения, объединяя классический брандмауэр с дополнительными возможностями фильтрации сетевой инфраструктуры, включая встроенные механизмы выявления атак (IPS) и расширенную инспекцию сетевых данных (DPI) [2]. Брандмауэр следующего поколения - это брандмауэры с глубокой проверкой пакетов (DPI), которые выходят за рамки проверки портов и протоколов, а также блокировки, добавляя проверку на уровне приложений, предотвращение вторжений и предоставление информации из-за пределов брандмауэра [2]. Традиционные брандмауэры функционировали на уровнях 3 и 4 модели OSI и разрешали или блокировали трафик в зависимости от порта и протокола, используя проверку с учетом состояния, и принимали решения, основанные на определенных политиках [2]. В наши дни уже нецелесообразно и не вызывает доверия применение политик безопасности таким негибким и непрозрачным способом [3].

По мере того как инциденты развивались и становились все более изощренными, злоумышленникам удавалось обходить брандмауэры с отслеживанием состояния, что значительно повышало важность повышения безопасности. Брандмауэр следующего поколения стал предоставлять все возможности традиционного брандмауэра, а также дополнительные возможности управления приложениями и интегрированного предотвращения вторжений [3]. Они также предоставили более детализированную функциональность для идентификации местоположения пользователя и приложения. Вместо использования множества различных точечных решений брандмауэр нового поколения значительно упрощает и повышает эффективность применения правил безопасности во все более сложном компьютерном мире [3].

Общие сведения и свойства брандмауэра

Брандмауэр - это устройство сетевой безопасности, которое отслеживает входящий и исходящий сетевой трафик и разрешает или запрещает передачу пакетов данных на основе набора политик безопасности. Его цель - создать барьер между внутренней сетью и входящим

трафиком из внешних источников (например, из Интернета), чтобы блокировать вредоносный трафик, такой как вирусы и хакерские атаки [4, 11]. Брандмауэры тщательно анализируют входящий трафик на основе заранее сформированных правил, а также фильтруют трафик, исходящий из небезопасных или подозрительных источников, чтобы помочь предотвратить атаки. Брандмауэр защищает трафик в точке входа компьютера, известной как порты, то есть там, где может осуществляться обмен информацией с внешними устройствами [4, 11].

Существует два основных типа брандмауэров - программный и аппаратный. Первый тип предусматривает развёртывание специализированного программного обеспечения непосредственно на каждом устройстве, который отслеживает и регулирует сетевые соединений через контроль портов и запущенных приложений. Аппаратный брандмауэр - это специализированное устройство в корпоративной сети, действующего как промежуточный элемент между внутренней инфраструктурой и внешними сетями. Наиболее эффективно применять обоих типа одновременно.

1) Брандмауэры с фильтрацией на уровне пакетов: представляет собой наиболее распространённый механизм сетевой защиты. Его принцип действия основан на анализе каждого передаваемого блока данных и блокировании элементов, не удовлетворяющих действующей политики безопасности. Система осуществляет верификацию параметров пакетов отправителя и получателя. В случае соответствия параметров установленному правилу брандмауэра, обмен данными разрешается.

2) Брандмауэры с фильтрацией пакетов подразделяются на два типа:

Бесконтекстные брандмауэры оценивают каждый сетевой пакет независимо от остальных. Серьезным недостатком является отсутствие механизма отслеживания взаимосвязей между передаваемыми блоками данных.

Контекстные брандмауэры ведут журнал всех сетевых операций, что обеспечивает более высокую защиту от злоумышленника.

Тем не менее, технология пакетной инспекции имеет критические недостатки. Основное ограничение заключается в неспособности интерпретировать контекст передаваемых пакетов и оценивать их. Например, если вредоносный трафик поступает через разрешенный канал от проверенного источника и вызывает негативные последствия (например, удаление базы данных), контекстный брандмауэр пропустит его без распознавания угрозы. Брандмауэры следующего поколения (NGFW) обладают расширенным функционалам для обнаружения подобных сложных атак.

3) Межсетевые экраны прокси-сервера играет роль промежуточного звена между двумя оконечными узлами и выполняют контроль потоков данных на прикладном уровне. Запрос от пользователя поступает на межсетевой экран, который анализирует его на предмет соблюдения установленных требований безопасности, после чего пропускает или блокирует. Помимо прочего, такие межсетевые экраны осуществляют инспекцию передаваемых данных на прикладном уровне (включая HTTP(s) и FTP), параллельно анализируя состояния соединений и содержимого пакетов.

4) Брандмауэры с трансляцией сетевых адресов (Network Address Translation, NAT) дают возможность устройствам с независимыми локальными адресами осуществлять выход в сеть посредством единого адреса, что гарантирует маскировку индивидуальных IP-адресов. Применение NAT-технологии дает дополнительный уровень защиты, поскольку злоумышленник, осуществляющие активное или пассивное сканирование адресного

пространства, лишаются возможности идентификации реальных сетевых адресов. Следует отметить функциональную схожесть между NAT-брандмауэрами и прокси-серверами: обе технологии реализуют концепцию посреднического узла (intermediary node). Она заключается в контролируемом взаимодействие между защищаемым сегментом локальной сети и внешним трафиком.

5) Брандмауэры с технологией многоуровневой инспекции состояния соединений (Stateful Multilayer Inspection, SMLI) анализируют трафик на канальном, транспортном и прикладном уровнях модели OSI. Особенность технологии заключается в последовательной верификации входящих пакет путём сопоставления их с авторизованным трафиком. Принципиальное сходство SMLI-брандмауэров с межсетевыми экранами следующего поколения (NGFW) проявляется в применении методологии глубокого анализа пакетов, когда исследуются не только заголовки, но и полезной нагрузки каждого сетевого пакета. Важной функцией выступает контроль состояния сеанса связи, позволяющий отслеживать статус соединения.

6) Брандмауэр с единым управлением угрозами (Unified threat management, UTM) - устройство UTM, которое, как правило, сочетает в себе функции брандмауэра с отслеживанием состояния, предотвращения вторжений и антивирусной защиты. Оно также может включать дополнительные службы и часто облачное управление. UTM концентрируются на том, чтобы быть простыми и удобными в использовании [5].

7) Брандмауэры следующего поколения (NGFW) представляют собой комплексные решения, объединяя в единую архитектуру модули обнаружения вредоносного ПО, подсистемы активного противодействия кибератакам, средства криптоанализа защищённых каналов и технологии контекстного анализа трафика. Отличительной особенностью данного поколения брандмауэров является внедрение механизма глубокой инспекции пакетов (DPI). Если традиционные межсетевые экраны ограничивались обработкой пакетов на сетевом и транспортном уровнях модели OSI, то DPI-технология обеспечивает полноценный анализ на прикладном уровне [5].

8) NGFW, ориентированный на угрозы (threat-centric) имеют расширенный механизм выявления и нейтрализации инцидентов безопасности. Данный класс брандмауэров имеют следующие преимущества [5]:

- Внедрение систем автоматизированного реагирование на инциденты, которые осуществляют динамическое конфигурацию защитных политик на основе потоков угроз;
- Позволяют в режиме реального времени выполнять пространственную и временную кластеризацию критичных узлов инфраструктуры и профилирование сетевой активности;
- Значительно сокращается время от обнаружения до устранения неполадок благодаря ретроспективной системе безопасности, которая постоянно отслеживает подозрительные действия и выполняет их даже после первоначальной проверки;
- Упрощается администрирование и уменьшаются трудности благодаря унифицированным политикам, обеспечивающим защиту от всех атак.

О брандмауэрах для веб-приложений (WAF)

Межсетевой экран веб-приложений (WAF) - это специализированный межсетевой экран, предназначенный для фильтрации и управления HTTP(s)-трафиком в Интернете между веб-клиентами и серверами приложений. Межсетевые экраны веб-приложений (WAF) являются важным компонентом для обеспечения надежной безопасности приложений. Традиционные сетевые межсетевые экраны работают на сетевом и транспортном уровнях, а также контролируют передачу пакетов и данных. WAF, в свою очередь, обеспечивает безопасность на уровне 7 модели OSI, обычно располагаясь между межсетевым экраном и веб-сервером или сервером приложений. В отличие от своих предшественников, межсетевые экраны веб-приложений, подключаемые к портам, делают шаг вперед в обеспечении безопасности приложений, обслуживаемых через интернет [6].

В свою очередь прокси-сервера реализуют принцип клиентской анонимности путём маскировки исходных адресов через механизм посредничества. Напротив, WAF-технологии функционируют в режиме реверсивного прокси, позиционируясь между внешней сетью и серверной инфраструктурой как обязательная точка инспекции. Данная топология обеспечивает непосредственную защиту серверных ресурсов от раскрытия, требуя предварительной проверку каждого клиентского запроса перед его маршрутизацией к целевому приложению [7]. Основная задача WAF - защита конкретных приложений от веб-атак на уровне приложения. WAF продолжают внедрять расширенные функции, такие как анализ угроз, балансировка нагрузки, предотвращение вторжений, единое управление угрозами (UTM) и многое другое [6]. Все эти возможные угрозы объясняют необходимость в специализированной технологии межсетевых экранов веб-приложений [3]. В итоге, WAF закрывают пробел в защите, оставленный традиционными межсетевыми экранами при решении проблем безопасности приложений [6].

Функции брандмауэрах для веб-приложений (WAF)

Межсетевой экран веб-приложений (WAF) обеспечивает защиту посредством фильтрации, мониторинга и блокировки всего вредоносного HTTP/HTTPS-трафика, поступающего в веб-приложения, и помогает предотвратить выход несанкционированных данных. Это достигается за счет соблюдения набора политик, которые помогают определить, какой трафик является вредоносным, а какой - безопасным [8]. Подобно тому, как прокси-сервер выступает в качестве посредника для защиты личности клиента, межсетевой экран веб-приложений (WAF) работает аналогичным образом, но в обратном порядке называется обратным прокси-сервером, который действует как посредник, защищающий сервер веб-приложений от потенциально вредоносного клиента [6, 9].

Брандмауэры веб-приложений (WAF) могут быть программными, аппаратными или предоставляться как услуга. Политики безопасности могут быть настроены в соответствии с уникальными требованиями веб-приложения или их набора. Хотя многие WAF требуют частого обновления политик для устранения новых уязвимостей, инновации в машинном обучении позволяют некоторым межсетевым экранам веб-приложений (WAF) автоматически обновляться [8]. Эта автоматизированная система становится все более важной, поскольку ландшафт угроз продолжает усложняться и становиться все более неопределенным.

Предотвращение атак с помощью брандмауэров для веб-приложений (WAF)

Следует отметить, что межсетевые экраны веб-приложений (WAF) предназначены для работы в сочетании с комплексным набором продуктов безопасности, таких как традиционные межсетевые экраны, а также системы предотвращения вторжений (IPS). В то же время, хотя они очень эффективны в блокировании определенных типов атак, они не предназначены для защиты от всех угроз. С учетом этого, наиболее распространенные кибератаки на приложения, которые помогают предотвратить WAF приведены в Таблице 1.

Таблица 1 - Предотвращение атак с помощью (WAF)

Вид кибератаки	**Описание**
SQL-инъекция	Это разновидность атаки, при которой используется SQL-запрос, позволяющий хакерам считывать конфиденциальные данные, подделывать идентификационные данные, изменять данные, выполнять административные задачи и иногда отдавать приказы операционной системе.
Межсайтовый скриптинг (Cross-site scripting, XSS)	Это разновидность инъекции, при которой вредоносные скрипты внедряются на доверенные веб-сайты. Хакеры рассылают вредоносный код, часто это скрипты на стороне браузера, конечному пользователю. Если код будет обнаружен, пользователь не сможет понять, что скрипт ненадежен и будет его выполнять. Вредоносный скрипт может получить доступ к конфиденциальной информации или изменить ее содержимое.
Атаки прикладного уровня или DDoS-атаки	DDoS-атаки на прикладной уровень связаны с вредоносным поведением, направленным на «самый верхний» уровень модели OSI, на котором происходят обычные интернет запросы, такие как HTTP GET и HTTP POST. Такие атаки на 7-й уровень модели OSI, в отличие от атак на сетевом уровне, таких как DNS, особенно эффективны благодаря использованию серверных ресурсов в дополнение к сетевым.
Веб-скрапинг	Веб-скрейпинг - это метод получения определенной информации с веб-сайта, часто используемый для собственных нужд. Межсетевой экран веб-приложений (WAF) может предотвратить получение данных с веб-сайта скриптами или другими средствами.
Заражение файлами cookie	Это своего рода кибератака, при которой хакеры манипулируют файлом cookie, который будет возвращен на сервер. Измененным файлом cookie можно манипулировать для обхода системы безопасности или кражи конфиденциальной информации.
Неподтвержденный ввод	Злоумышленники изменяют HTTP-запросы (а также URL-адрес, заголовки и поля формы), чтобы обойти методы защиты веб-сайта.
Использование уязвимостей переполнения буфера	Переполнение буфера - это ошибка программного кода, которая возникает, когда в буфер помещается больше данных, чем он может обработать, что приводит к утечке данных в соседнее хранилище. Переполнение буфера может создать точку входа для киберугроз, позволяющую хакеру перезаписать часть памяти приложения.

Преимущества межсетевой экран веб-приложений (WAF)

Брандмауэры веб-приложений специально разработаны, чтобы помочь защитить сеть от потенциальных угроз, которые еще не были обнаружены, а это означает, что внедрение этого решения может спасти организацию или пользователя от угроз нулевого дня, SQL-инъекций, атак с использованием межсайтовых сценариев, уязвимости в системе безопасности и других видов угроз [10]. Правильно настроенные брандмауэры беспроводных приложений также участвуют в действиях по предотвращению атак ботов или других ненужных инцидентов с трафиком. WAF будет поддерживать чистый поток трафика приложений и в то же время защищать все потоки вредоносных данных. В Таблице 2 приведен подробный обзор функциональных возможностей каждого типа брандмауэра.

Таблица 2 - Функциональные возможности каждого типа брандмауэра

Функция брандмауэра	**Межсетевой экран веб-приложений (WAF)**	**Межсетевой экран нового поколения (NGFW)**
Охват модели OSI	Уровень 7	Уровни 3-7
Типовое развертывание	Обратный прокси	Шлюз
Защита от сетевых угроз	Поддерживается только HTTP(s)	Полностью поддерживается
Ограничение трафика с помощью политик безопасности	Поддерживается только HTTP(s)	Полностью поддерживается
Предотвращение вредоносного ПО	С интегрированными системами обнаружения и предотвращения вторжений (IDS/IPS)	С интегрированными системами обнаружения и предотвращения вторжений (IDS/IPS)
Защита от DDos	Уровни 3-7 с CDN	Только уровни 3 и 4
Защита веб-приложений	Полностью поддерживается	Базовая
OWASP Top 10 покрытие (открытый проект по обеспечению безопасности веб-приложений)	Полностью поддерживается	Базовая
Настраиваемые правила веб-приложений	Полностью поддерживается	Нет
Эвристический метод сопоставления с веб-уязвимостями	Полностью поддерживается	Нет
Моделирование веб-приложений для дополнительной защиты	Полностью поддерживается	Нет

Наиболее подходящая межсетевой экран для организации

Не существует абсолютно универсального решения, которое могло бы удовлетворить исключительные требования к безопасности каждой организации. Более того, каждый из различных типов брандмауэров имеет свои преимущества и ограничения. Брандмауэры с фильтрацией пакетов являются базовыми, но обеспечивают ограниченную безопасность, в то

время как проверка состояния, а также прокси-брандмауэры могут снижать эффективность сети. Брандмауэры следующего поколения представляют собой полноценный пакет, хотя не все организации располагают бюджетом или ресурсами для их успешной настройки и управления ими. Поскольку атаки становятся все более изощренными, средства защиты организации должны оказывать противодействие на должном уровне. Одного брандмауэра, защищающего внутреннюю сеть от внешних угроз, будет недостаточно. Каждый ресурс в частной сети также нуждается в индивидуальной защите. Лучше всего использовать многоуровневый подход к обеспечению безопасности, а не полагаться на возможности конкретного брандмауэра. Нет необходимости в выборе одного конкретного межсетевого экрана, когда речь идет о том, как использовать преимущества нескольких брандмауэров в отношении архитектуры, улучшенной специально для удовлетворения потребностей организации в области безопасности [10].

Заключение

Принимая во внимание угрозы, с которыми сталкиваются корпоративные сети сегодня, совершенно очевидно, что организации хотят пересмотреть свои методы обеспечения безопасности. Таким образом, брандмауэры следующего поколения (NGFW) по-прежнему способны предотвращать многие виды сетевых угроз для защиты от несанкционированного доступа и изменения характеристик внутри сети организации. Ранее основные защитные средства, такие как классические межсетевые экраны и системы обнаружения и предотвращения вторжений (IDS/IPS), обеспечивали необходимый уровень безопасности для ограниченного набора веб-сервисов. Рост критической зависимости бизнес-процессов от веб-приложений, сопровождается параллельной эволюцией методологии кибератак. Тем не менее, брандмауэр веб-приложений предлагает более обширные правила, предназначенные для блокировки уровней веб-приложений, и является чрезвычайно важным условием, если организация хочет гарантировать безопасность и надежность своих общедоступных веб-сайтов или веб-приложений. Данная статья является кратким анализом брандмауэров и межсетевых экранов нового поколения. Проведя обзор, можно подчеркнуть, что межсетевые экраны следующего поколения (NGFW) ориентированы на обеспечение комплексной защиты сети в целом, тогда как брандмауэры веб-приложений (WAF) специализируются на защите конкретных веб-ориентированных программных систем. Организации, которые зависят от веб-приложений, могут получить большую выгоду от WAF. Для таких клиентов в большинстве случаев рекомендуется использовать оба решения.

Список литературы

1. What is a Web Application Firewall (WAF)? // F5 Distributed Cloud Services URL: f5.com/glossary/web-application-firewall-waf (дата обращения: 05.03.2026).
2. What Is a Next-Generation Firewall? // Zscaler URL: zscaler.com/resources/security-terms-glossary/what-is-next-generation-firewall (дата обращения: 05.03.2026).
3. FAW vs NGFW // Orange Cyberdefense URL: orangecyberdefense.com/be/blog/waf-vs-ngfw (дата обращения: 21.03.2026).
4. Заборовский В. С. Методы и средства защиты компьютерной информации. Межсетевое экранирование [Текст]: учеб. пособие // В. С. Заборовский, В.А. Мулюха, А.Г. Новопашенный, Ю.Е. Подгурский, В.С. Заборовский. - СПб.: СПбГПУ, 2010 - 91 с.

5. What is a firewall? // Cisco URL: cisco.com/site/us/en/learn/topics/security/what-is-a firewall.html#~types-of-firewalls (дата обращения: 21.03.2026).
6. Курамбаев Й.Б. Особенности использования WAF FIREWALL технологий в защите данных // Научный журнал «A POSTERIORI». - 2023. - №4. URL: elibrary.ru/download/elibrary_50488430_63401467.pdf (дата обращения: 21.03.2026).
7. Баранов П.А., Бейбутов Э.Р. Обеспечение информационной безопасности информационных ресурсов с помощью межсетевых экранов для веб-приложений // Business Informatics. - 2015. - №4.
8. Azure Web Application Firewall // Microsoft Azure URL: azure.microsoft.com/en-us/products/web-application-firewall/ (дата обращения: 10.04.2026).
9. Шатурный М.В. Анализ актуальных угроз и разработка подходов к защите веб - приложений // Инженерный вестник Дона. - 2024. - №7. URL: ivdon.ru/back_media/uploads/article/pdf/IVD_58N7y24_Shaturniy.pdf_84babdb54b.pdf (дата обращения: 10.04.2026).
10. Кацупеев А.А., Щербакова Е.А., Воробьёв С.П. Постановка и формализация задачи формирования информационной защиты распределённых систем // Инженерный вестник Дона. - 2015. - №1, ч.2. URL: ivdon.ru/back_media/uploads/article/pdf/IVD_86_katsupeev.pdf_79178b4853.pdf (дата обращения: 17.04.2026).
11. Пестов, И. Е. Использование брандмауэра для защиты информации / И. Е. Пестов, Ю. О. Качуровский // Инновационные технологии, экономика и менеджмент в промышленности: сборник научных статей по итогам XII международной научной конференции, Волгоград, 23-24 декабря 2021 года / НПП Медпромдеталь. Том Часть 1. - Волгоград: ООО "Конверт", 2021. - С. 203-204. URL: elibrary.ru/download/elibrary_47554440_45787637.pdf (дата обращения: 17.04.2026).

References

1. What is a Web Application Firewall (WAF)? // F5 Distributed Cloud Services URL: f5.com/glossary/web-application-firewall-waf (accessed: 03/05/2026).
2. What Is a Next-Generation Firewall? // Zscaler URL: zscaler.com/resources/security-terms-glossary/what-is-next-generation-firewall (accessed: 03/05/2026).
3. FAW vs NGFW // Orange Cyberdefense URL: orangecyberdefense.com/be/blog/waf-vs-ngfw (date of request: 03/21/2026).
4. Zaborovskiy V. S. Methods and means of computer information protection. Inter-network shielding [Text]: textbook. the manual // V. S. Zaborovsky, V.A. Mulukha, A.G. Novopashenny, Yu.E. Podgursky, V.S. Zaborovsky. - St. Petersburg: SPbGPU, 2010 - 91 p.
5. What is a firewall? // Cisco URL: cisco.com/site/us/en/learn/topics/security/what-is-a firewall.html #~types-of-firewalls (accessed: 03/21/2026).
6. Kurambaev J.B. Features of using WAF FIREWALL technologies in data protection // Scientific journal "A POSTERIORI". - 2023. - No. 4. URL: elibrary.ru/download/elibrary_50488430_63401467.pdf (date of reference: 03/21/2026).
7. Baranov P.A., Beybutov E.R. Ensuring information security of information resources using firewalls for web applications // Business Informatics, 2015, No. 4.

8. Azure Web Application Firewall // Microsoft Azure URL: azure.microsoft.com/en-us/products/web-application-firewall / (date of access: 04/10/2026).
9. Shaturny M.V. Analysis of current threats and development of approaches to web application protection // Engineering Bulletin of the Don. - 2024. - No. 7. URL: ivdon.ru/back_media/uploads/article/pdf/IVD_58N7y24_Shaturniy.pdf_84babdb54b.pdf (date of reference: 04/10/2026).
10. Katsupeev A.A., Shcherbakova E.A., Vorobyev S.P. Formulation and formalization of the task of forming information security of distributed systems // Engineering Bulletin of the Don. - 2015. - No. 1, part 2. URL: ivdon.ru/back_media/uploads/article/pdf/IVD_86_katsupeev.pdf_79178b4853.pdf (date of request: 04/17/2026).
11. Pestov, I. E. Using a firewall to protect information / I. E. Pestov, Yu. O. Kachurovsky // Innovative technologies, economics and management in industry: a collection of scientific articles based on the results of the XII International Scientific conference, Volgograd, December 23-24, 2021 / NPP Medpromdetal. Volume Part 1. - Volgograd: LLC "Envelope", 2021. - pp. 203-204. URL: elibrary.ru/download/elibrary_47554440_45787637.pdf (date of request: 04/17/2026).

Теплая Н.А., Пронин В.Ю. Сравнительный анализ систем управления проектами JIRA, TRELLO и ASANA: возможности, ограничения и области применения // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7(69) Ч.1 с. 74–80

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 005.8:004.9

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СИСТЕМ УПРАВЛЕНИЯ ПРОЕКТАМИ JIRA, TRELLO И ASANA: ВОЗМОЖНОСТИ, ОГРАНИЧЕНИЯ И ОБЛАСТИ ПРИМЕНЕНИЯ

Теплая Н.А., Пронин В.Ю.
ФГБОУ ВО "МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕХНОЛОГИЙ И УПРАВЛЕНИЯ ИМЕНИ К.Г. РАЗУМОВСКОГО (ПЕРВЫЙ КАЗАЧИЙ УНИВЕРСИТЕТ)" Москва, Россия, (109004, город Москва, ул. Земляной Вал, д.73), e-mail: [1] *dana.gaunova@gmail.com*

Статья посвящена сравнительному анализу трёх наиболее распространённых систем управления проектами: Jira, Trello и Asana. Рассматриваются ключевые функциональные возможности, интерфейсные особенности, модели ценообразования и оптимальные области применения каждой платформы. На основе анализа выявлены сильные и слабые стороны систем, определены критерии их выбора для проектов различной сложности и масштаба. Результаты исследования ориентированы на руководителей ИТ-проектов, скрам-мастеров и специалистов, принимающих решение о внедрении инструментов управления задачами в гибких средах.

Ключевые слова: Управление проектами, Jira, Trello, Asana, Agile, Scrum, Kanban, сравнительный анализ.

COMPARATIVE ANALYSIS OF JIRA, TRELLO, AND ASANA PROJECT MANAGEMENT SYSTEMS: OPPORTUNITIES, LIMITATIONS, AND APPLICATIONS

[1] **Teplaya N.A.**, **Pronin V.Yu**.
GUBKIN RUSSIAN STATE UNIVERSITY OF OIL AND GAS (NATIONAL RESEARCH UNIVERSITY), Moscow, Russia, (119296, Moscow, Leninsky pr-kt, 65 k. 1), e-mail: [1] *dana.gaunova@gmail.com*

The article provides a comparative analysis of three widely used project management systems: Jira, Trello, and Asana. Key functional capabilities, interface features, pricing models, and optimal application areas of each platform are examined. The analysis identifies strengths and weaknesses of the systems and defines criteria for their selection for projects of varying complexity and scale. The findings are intended for IT project managers, Scrum masters, and specialists making decisions on implementing task management tools in agile environments.

Keywords: Project management, Jira, Trello, Asana, Agile, Scrum, Kanban, comparative analysis.

Цифровая трансформация бизнеса и массовое распространение гибких методологий разработки программного обеспечения выдвигают на первый план задачу рационального выбора инструментов управления проектами. Рынок предлагает широкий спектр решений, различающихся по сложности, цене и целевой аудитории. На фоне активного импортозамещения и перехода российских компаний на отечественные аналоги интерес к объективному сравнению мировых лидеров сохраняется, поскольку многие команды продолжают использовать привычные зарубежные продукты [1].

Среди наиболее популярных платформ выделяются три системы, фактически ставшие индустриальным стандартом: Jira (Atlassian), Trello (также продукт Atlassian) и Asana (Asana, Inc.). По данным опросов 2025 года, Jira доминирует в ИТ- и DevOps-среде, Trello остаётся востребованным в малом бизнесе и стартапах, а Asana занимает промежуточное положение, привлекая команды, ориентированные на сбалансированное сочетание функциональности и удобства [2, 3].

Целью настоящей работы является проведение всестороннего сравнительного анализа Jira, Trello и Asana для систематизации информации, необходимой для обоснованного выбора инструмента в зависимости от специфики проекта и характеристик команды.

Обзор рынка и предпосылки выбора. Рынок систем управления проектами продолжает расти. Основными драйверами выступают ускорение разработки программного обеспечения, распространение удалённой работы и потребность в прозрачных инструментах координации распределённых команд. В российском сегменте наряду с мировыми лидерами активно развиваются отечественные платформы (Яндекс Трекер, YouGile, Kaiten), однако многие организации, особенно с устоявшимися процессами, сохраняют приверженность Jira и ищут компромиссные варианты среди Trello и Asana [4].

Jira была создана в 2002 году как баг-трекер для разработчиков Java и впоследствии переросла в полнофункциональную систему управления проектами и задачами. Trello, запущенный в 2011 году как побочный продукт Fog Creek Software, предложил предельно простой подход на основе Kanban-досок и быстро завоевал популярность благодаря интуитивной понятности. Asana, основанная в 2008 году Дастином Московицем и Джастином Розенштейном, изначально проектировалась как инструмент для повышения прозрачности и командной работы, объединяющий списки задач, диаграммы Ганта и стратегическое целеполагание. Цифровые инструменты, такие как Jira, Trello и Asana, способствуют оптимизации рабочих процессов, повышению эффективности команд и сокращению сроков реализации проектов [2].

Методология исследования. Сравнение проводилось по пяти ключевым критериям: функциональность, простота освоения, поддержка Agile-методологий, интеграционные возможности и экономические параметры. Источниками послужили официальная документация Atlassian и Asana, аналитические обзоры 2024–2026 годов [2, 3, 4, 5], а также пользовательские рейтинги на платформах G2 и Capterra.

Результаты

Jira: профессиональный инструмент для разработки. Jira обладает наиболее широким функционалом среди рассматриваемых систем. Она поддерживает полноценное управление спринтами, бэклогами, пользовательскими историями и включает мощные средства визуализации хода работ — burndown-диаграммы, roadmaps, настраиваемые дашборды [6]. Платформа позволяет создавать гибкие рабочие процессы, автоматизировать действия через встроенные правила и получать детальную аналитику по качеству и производительности команды. Более 6000 интеграций с GitHub, GitLab, Confluence, Slack и другими сервисами делают Jira центральным узлом сквозных DevOps-циклов [2, 4].

Вместе с тем интерфейс Jira часто критикуют за перегруженность, а настройка корректных workflow требует участия опытного администратора [3]. Лицензионная политика

Atlassian, особенно после приостановки продаж в России, и высокая стоимость коммерческих тарифов ограничивают её применение небольшими стартапами. Бесплатная версия доступна только для команд до 10 участников, а стандартный тариф начинается от $9,05 за пользователя в месяц [6].

Trello: простота и визуальная наглядность. Trello реализует Kanban-подход в наиболее наглядном виде: доски, списки и карточки, перемещаемые между колонками, интуитивно понятны даже неподготовленным пользователям [7]. Платформа практически не требует обучения и отлично подходит для оперативного управления задачами, личной организации и координации небольших команд. В 2025 году Trello получил крупное обновление, добавив AI-функции для автоматического создания карточек на основе входящих писем, расширенные интеграции с почтовыми сервисами и улучшенную связку с Jira [5, 7].

Однако функциональность Trello остаётся ограниченной базовым управлением задачами. Отсутствуют встроенные средства отслеживания спринтов, аналитические отчёты и глубокая настройка рабочих процессов. Расширение возможностей возможно через Power-Ups и автоматизацию Butler, но для сложных проектов с большим количеством зависимостей платформа оказывается недостаточной [3, 5].

Asana: универсальность и сбалансированный подход. Asana занимает промежуточное положение, предлагая широкий спектр представлений: списки, доски Kanban, временные шкалы (Timeline) и стратегические цели (Goals). Платформа изначально проектировалась как средство обеспечения прозрачности и подотчётности в командах любого размера, независимо от применяемой методологии [8]. Интеграции с Slack, Google Drive, Microsoft Teams и инструментами разработки позволяют встроить Asana в существующую инфраструктуру.

Преимуществами Asana являются развитые средства командной работы (комментарии, вложения, назначение ответственных, уведомления) и гибкость визуализации. Сдерживающим фактором выступает относительно высокая стоимость доступа к продвинутым функциям для небольших команд, а также некоторая «перегруженность» интерфейса, которая может показаться избыточной для простых проектов [3, 4].

Обобщённые результаты сравнения приведены в Таблице 1.

Таблица 1 — Сравнительный анализ Jira, Trello, Asana

Критерий	**Jira**	**Trello**	**Asana**
Функциональность	Максимальная: Scrum, Kanban, отчётность, roadmaps [6]	Базовая: Kanban-доски [7]	Средняя: списки, доски, Timeline, Goals [8]
Простота освоения	Высокий порог входа [3	Минимальный порог [5]	Умеренный порог [4]
Поддержка Agile	Полноценная [2]	Отсутствует	Ограниченная [8
Интеграции	>6000 (GitHub, Confluence, Slack) [2]	Power-Ups, Butler [7]	>1000 (Slack, Google Drive) [8]
Цена (на одного пользователя в месяц)	Бесплатно до 10 польз., Standard от $9.05 [6]	Бесплатный тариф, Standard от $5 [7]	Бесплатно до 15 польз., Starter от $10.99 [8]

Помимо основных критериев, важными для конечного выбора оказываются такие параметры, как доступность мобильных приложений, возможность экспорта данных в случае миграции, а также официальная техническая поддержка на русском языке. Jira и Trello, будучи продуктами Atlassian, не предоставляют локализованную поддержку для российских пользователей, тогда как Asana имеет ограниченный русскоязычный интерфейс и сообщество.

Обсуждение. Проведённый анализ демонстрирует, что выбор подходящей системы напрямую зависит от контекста использования.

Jira остаётся безальтернативным выбором для профессиональных команд разработки, практикующих Scrum и Kanban в зрелой форме. Её глубина проработки спринтов, бэклогов и аналитики оправдывает затраты на внедрение и обучение в средних и крупных ИТ-компаниях [2, 6]. Вместе с тем Atlassian объявила о прекращении поддержки Data Center к 2029 году, что в совокупности с геополитическими рисками стимулирует часть пользователей рассматривать миграцию на другие платформы [1].

Trello оптимален для стартапов, небольших отделов и индивидуальной работы, где критически важны скорость внедрения и наглядность. При этом по мере усложнения проектов команды часто переходят с Trello на более мощные системы — Jira или Asana [5, 7]. Показателен опыт многих маркетинговых агентств, которые начинали с Trello для планирования кампаний, но по мере роста портфеля проектов внедряли Asana для лучшей приоритезации и отслеживания сроков.

Asana представляет собой сбалансированное решение для кросс-функциональных проектов, маркетинга, операционной деятельности и гибридных методологий. Её способность объединять стратегические цели с повседневными задачами делает платформу удобной для менеджмента среднего звена и руководителей, не погружённых в технические детали Agile [8]. Компании, практикующие гибридные Waterfall-Agile-подходы, часто отмечают, что Asana позволяет одновременно управлять и долгосрочными планами, и спринтами.

Следует отметить, что границы между платформами постепенно стираются: обновлённый Trello с AI-функциями и интеграцией с Jira становится более «умным», а Asana

активно внедряет автоматизацию рабочих процессов [5, 8]. Этот тренд повышает конкуренцию и расширяет возможности выбора для конечных пользователей. Кроме того, развитие экосистемы дополнений и ботов (например, Butler для Trello, Asana Rules) позволяет частично преодолеть функциональные ограничения, хотя и требует дополнительных усилий по настройке.

Таким образом, универсального «лучшего» инструмента не существует. Решение должно приниматься коллегиально, с учётом текущей зрелости процессов, предполагаемого роста команды и бюджета. Рекомендуется также проводить пилотное тестирование 2–3 платформ на реальных задачах перед окончательным выбором. Важнейшим фактором успеха является адаптивность и постоянное совершенствование процессов, что подтверждается исследованиями в области гибких методов управления проектами [9].

Заключение. Сравнительный анализ показал, что Jira, Trello и Asana реализуют три различные философии управления проектами: Jira ориентирована на максимальную функциональность и контроль, Trello — на простоту и скорость, Asana — на универсальность и командную коллаборацию. Оптимальный выбор определяется масштабом проекта, зрелостью процессов, бюджетом и компетенциями команды. Для разработки программного обеспечения по Agile предпочтительна Jira, для быстрых стартов и визуального управления — Trello, а для сбалансированной работы смешанных команд — Asana. Практическая значимость работы заключается в систематизации критериев выбора и предоставлении обоснованных рекомендаций для руководителей проектов и ИТ-специалистов, принимающих решения о внедрении или смене инструментальной платформы.

Список литературы

1. Zhaksylykov N. Digitalization of Project Management in the IT Sector: Case Studies and Models // Вестник науки Казахского агротехнического университета им. С. Сейфуллина. — 2025. — URL: https://cyberleninka.ru/article/n/digitalization-of-project-management-in-the-it-sector-case-studies-and-models (дата обращения: 12.05.2026).
2. Milojević D., Macuzic I., Djordjevic A., Savković M., Djapan M. Comparative Analysis of Software Tools for Agile Project Management // International Conference on Production Engineering (Serbia). — 2023. — URL: https://scidar.kg.ac.rs/handle/123456789/18405 (дата обращения: 12.05.2026).
3. Сергеева А. И., Коняхина Е. А. Управление инновационными ИТ-проектами: квантовые и прорывные технологии как вызов для проектного менеджмента // Молодой ученый. — 2025. — № 51 (602). — С. 285–291. — URL: https://moluch.ru/archive/602/131800 (дата обращения: 12.05.2026).
4. Борзых Н. Ю., Смоленцева Т. Е., Смирнов М. В. К вопросу управления входным набором критериев при выборе стратегии проектирования корпоративных информационных систем // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и Технические Науки. — 2023. — № 06. — С. 61–67. — DOI: 10.37882/2223-2966.2023.06.06.
5. Курочкин Д. Э. Современные проблемы управления инновационными проектами информатизации предприятия // Современные проблемы науки и образования. — 2015.

— № 1. — URL: https://science-education.ru/article/view?id=16867 (дата обращения: 12.05.2026).

6. Цыплов Е. А., Новиков В. А., Хайитов Х. О. Применение информационных систем управления проектами в повышении эффективности системы управления качеством // Вестник Тольяттинского государственного университета. — URL: https://cyberleninka.ru/article/n/primenenie-informatsionnyh-sistem-upravleniya-proektami-v-povyshenii-effektivnosti-sistemy-upravleniya-kachestvom (дата обращения: 12.05.2026).
7. Пащук Е. О. Agile-подходы в проектировании информационно-компьютерных систем: проблемы и пути их решения // Электронные системы и технологии: сборник материалов 61-й научной конференции аспирантов, магистрантов и студентов БГУИР. — Минск, 2025. — С. 760–763.
8. Зарубина П. А., Кравчук А. И., Тутынина Д. А. Влияние Agile и её фреймворка Scrum на эффективность управления ИТ-проектами // Сборник тезисов докладов конгресса молодых ученых. — СПб: Университет ИТМО, 2025. — URL: https://kmu.itmo.ru/digests/article/15998 (дата обращения: 12.05.2026).
9. Буталенко А. А., Елецких П. Е. Гибкие методы управления проектами: сравнительный анализ // Актуальные вопросы экономики и информационных технологий: сборник материалов 61-й научной конференции аспирантов, магистрантов и студентов БГУИР. — Минск, 2025. — С. 616–620.

References

1. Zhaksylykov N. Digitalization of Project Management in the IT Sector: Case Studies and Models // Bulletin of Science of the Kazakh Agrotechnical University named after S. Seifullin. — 2025. — URL: https://cyberleninka.ru/article/n/digitalization-of-project-management-in-the-it-sector-case-studies-and-models (date of request: 05/12/2026).
2. Milojević D., Macuzic I., Djordjevic A., Savković M., Djapan M. Comparative Analysis of Software Tools for Agile Project Management // International Conference on Production Engineering (Serbia). - 2023. — URL: https://scidar.kg.ac.rs/handle/123456789/18405 (date of access: 05/12/2026).
3. Sergeeva A. I., Konyakhina E. A. Management of innovative IT projects: quantum and breakthrough technologies as a challenge for project management // Young Scientist. — 2025. — № 51 (602). — pp. 285-291. — URL: https://moluch.ru/archive/602/131800 (date of request: 05/12/2026).
4. Borzykh N. Yu., Smolentseva T. E., Smirnov M. V. On the issue of managing the input set of criteria when choosing a strategy for designing corporate information systems // Modern science: actual problems of theory and practice. Series: Natural and Technical Sciences. — 2023. — No. 06. — pp. 61-67. — DOI: 10.37882/2223-2966.2023.06.06.
5. Kurochkin D. E. Modern problems of management of innovative projects of enterprise informatization // Modern problems of science and education. — 2015. — No. 1. — URL: https://science-education.ru/article/view?id=16867 (date of request: 05/12/2026).
6. Tsyplov E. A., Novikov V. A., Khayitov Kh. O. Application of project management information systems in improving the effectiveness of the quality management system // Bulletin of Tolyatti State University. — URL: https://cyberleninka.ru/article/n/primenenie-

informatsionnyh-sistem-upravleniya-proektami-v-povyshenii-effektivnosti-sistemy-upravleniya-kachestvom (date of request: 05/12/2026).

7. Pashchuk E. O. Agile approaches in the design of information and computer systems: problems and solutions // Electronic systems and technologies: collection of materials of the 61st scientific conference of graduate students, undergraduates and students of BSUIR. Minsk, 2025. pp. 760-763.
8. Zarubina P. A., Kravchuk A. I., Tutynina D. A. The influence of Agile and its Scrum framework on the effectiveness of IT project management // Collection of abstracts of the Congress of Young Scientists. — St. Petersburg: ITMO University, 2025. — URL: https://kmu.itmo.ru/digests/article/15998 (date of request: 05/12/2026).
9. Butalenko A. A., Yeletskikh P. E. Flexible methods of project management: a comparative analysis // Actual issues of economics and information technology: a collection of materials from the 61st scientific conference of graduate students, undergraduates and students of BSUIR. — Minsk, 2025. — pp. 616-620.

Мухина К.Е., Митянина А.В. Разработка и экспериментальная оценка графовой долговременной памяти LLM-агента для персонализации диалогов // Международный журнал информационных технологий и энергоэффективности.– 2026. – Т. 11 № 7(69) Ч. 1 с. 81–90

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.89:004.822:004.93'1

РАЗРАБОТКА И ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА ГРАФОВОЙ ДОЛГОВРЕМЕННОЙ ПАМЯТИ LLM-АГЕНТА ДЛЯ ПЕРСОНАЛИЗАЦИИ ДИАЛОГОВ

[1] **Мухина К.Е.,** [2]**Митянина А.В.**
ФГБОУ ВО "ЧЕЛЯБИНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ", Челябинск, Россия (454001, Челябинская область, город Челябинск, ул Братьев Кашириных, д. 129), e-mail: kseniamuhina16@gmail.com, [2]nastya.mityanina@gmail.com

В статье рассматривается разработка долговременной памяти для персонализации диалоговых агентов на основе больших языковых моделей. Показано, что контекстное окно, дообучение модели и классический Retrieval-Augmented Generation имеют ограничения при работе с длительными многосессионными диалогами. Предложен механизм памяти, объединяющий векторное хранилище и граф знаний о пользователе. Векторное хранилище используется для поиска по смысловому сходству, а граф знаний – для хранения фактов. Экспериментальная оценка показала, что предложенный механизм достигает метрики accuracy 88%, превышая результат контекстного окна на 13 процентных пунктов и классического RAG на 25 процентных пунктов.

Ключевые слова: Большие языковые модели, LLM-агенты, долговременная память, персонализация диалога, Retrieval-Augmented Generation, GraphRAG, граф знаний, пользовательский контекст.

DEVELOPMENT AND EXPERIMENTAL EVALUATION OF GRAPH-BASED LONG-TERM MEMORY FOR LLM AGENTS FOR DIALOGUE PERSONALIZATION

[1] **Mukhina K.E.**, [2]**Mityanina A. V**.
CHELYABINSK STATE UNIVERSITY, Chelyabinsk, Russia (454001, Chelyabinsk region, Chelyabinsk city, Bratyev Kashirinykh st., 129,), e-mail: [1]kseniamuhina16@gmail.com, [2]nastya.mityanina@gmail.com

The article presents the development of long-term memory for personalizing dialogue agents based on large language models. It shows that the context window, fine-tuning, and classical Retrieval-Augmented Generation have limitations in long multi-session dialogues. The proposed memory mechanism combines a vector store and a user knowledge graph. The vector store is used for semantic similarity search, while the knowledge graph stores facts. The experimental evaluation showed that the proposed mechanism achieves 88% accuracy, outperforming the context window mode by 13 percentage points and classical RAG by 25 percentage points.

Keywords: Large language models, LLM agents, long-term memory, dialogue personalization, Retrieval-Augmented Generation, GraphRAG, knowledge graph, user context.

Введение

Современные диалоговые агенты на основе больших языковых моделей все чаще используются не только для ответа на отдельные запросы, но и для длительного взаимодействия с пользователем. В таких сценариях от системы ожидается учет ранее сообщенных сведений: целей, ограничений, предпочтений, профессиональной области, языка

общения и истории предыдущих решений. Если эти данные не сохраняются между сессиями, пользователь вынужден повторять одну и ту же информацию.

Задача долговременной памяти особенно актуальна для прикладных сервисов, где диалоговый агент должен поддерживать персонализированное взаимодействие: помогать с обучением, планированием, подбором услуг, заполнением форм или работой с корпоративными данными. В этих случаях память становится не дополнительной функцией, а отдельным архитектурным компонентом, влияющим на качество и устойчивость диалога.

Существующие способы учета пользовательского контекста имеют существенные ограничения. Передача полной истории диалога в контекст модели плохо масштабируется: длинный контекст увеличивает стоимость инференса, а релевантные факты могут использоваться моделью нестабильно. В частности, исследования показывают, что качество извлечения информации снижается, если нужные сведения расположены в середине длинного контекста [1]. Дообучение модели или параметрически эффективная адаптация, например Low-Rank Adaptation (LoRA), полезны для настройки модели под домен или стиль, но плохо подходят для оперативного хранения изменяемых пользовательских фактов [2]. Retrieval-Augmented Generation позволяет хранить данные во внешнем хранилище и извлекать их при формировании ответа [3], однако классический RAG обычно работает с набором независимых текстовых фрагментов. Для персонализации этого недостаточно, поскольку пользовательский контекст включает не только отдельные сведения, но и связи между ними, источники, уточнения и актуальные версии фактов.

Цель работы – разработать и экспериментально оценить механизм графовой долговременной памяти в задаче персонализации длинных диалогов.

В данной работе предложен механизм долговременной памяти LLM-агента, основанный на совместном использовании векторного хранилища и графа знаний пользователя. Векторное хранилище отвечает за поиск релевантных сведений по смысловому сходству, а граф знаний позволяет хранить пользовательский контекст как связанную структуру: сущности, факты, отношения, исходные реплики и обновленные версии данных.

1. Анализ предметной области и существующих подходов

LLM-агентом в данной работе называется диалоговая система, в которой большая языковая модель используется не только для генерации текста, но и для выбора действий, обращения к инструментам, работы с внешними данными и формирования ответа с учетом дополнительного контекста. Одним из ключевых компонентов такого агента является память – механизм сохранения и повторного использования информации, полученной в предыдущих взаимодействиях.

Персонализация LLM-агента означает адаптацию его ответов к конкретному пользователю на основе ранее известных сведений о нем. Память не должна быть копией всей переписки. Она должна выделять значимые факты, связывать их между собой, обновлять при появлении новых данных и использовать только в релевантных ситуациях.

Основные подходы к организации памяти имеют разные ограничения. Контекстное окно просто в реализации, но становится дорогим и ненадежным на длинных диалогах. Суммаризация истории уменьшает объем передаваемого контекста, но приводит к потере деталей. Fine-tuning и LoRA позволяют адаптировать модель к домену или стилю, но не

предназначены для частого добавления и изменения индивидуальных пользовательских фактов [2]. Классический RAG делает систему более гибкой, но в типичной реализации извлекает независимые текстовые фрагменты и не отражает структуру пользовательских знаний [3].

Графовые структуры знаний, включая GraphRAG, позволяют перейти от поиска изолированных текстовых фрагментов к работе со связанным представлением данных [4]. В задаче персонализации это особенно важно, поскольку сведения о пользователе редко существуют независимо друг от друга. Графовое представление дает возможность не только найти релевантное утверждение, но и проверить, к какой сущности оно относится, из какой реплики было получено и не было ли оно обновлено позднее.

Таблица 1 – Сравнение подходов к организации памяти LLM-агента

Подход	**Преимущество**	**Ограничение для персонализации**
Контекстное окно	Простота реализации	Высокая стоимость, ограничение длины, потеря фактов в длинном контексте
Суммаризация	Сжатие длинного диалога	Потеря деталей, накопление ошибок
Fine-tuning	Адаптация к домену или стилю	Не подходит для частого обновления пользовательских фактов
Классический RAG	Внешнее обновляемое хранилище	Фрагменты извлекаются без явных связей и актуальных версий
Граф знаний пользователя	Хранение связей между фактами, сущностями и репликами	Требует отдельной логики извлечения, обновления и разрешения противоречий

Как показано в Таблице 1, каждый из существующих подходов решает только часть задачи. Поэтому для персонализации диалогового агента необходимо разработать такой механизм, который выделяет значимые пользовательские факты, связывает их с сущностями и исходными репликами, учитывает обновления и возвращает модели компактный согласованный контекст.

2. Архитектура предлагаемого решения

В работе предлагается механизм долговременной памяти LLM-агента, основанный на совместном использовании векторной базы данных и графовой базы данных. Пользовательские сведения сохраняются в двух взаимодополняющих формах: как векторные представления для семантического поиска и как структурированные факты со связями между сущностями, репликами и изменениями пользовательской информации. Процессы обращения к памяти и ее обновления представлены на Рисунке 1.

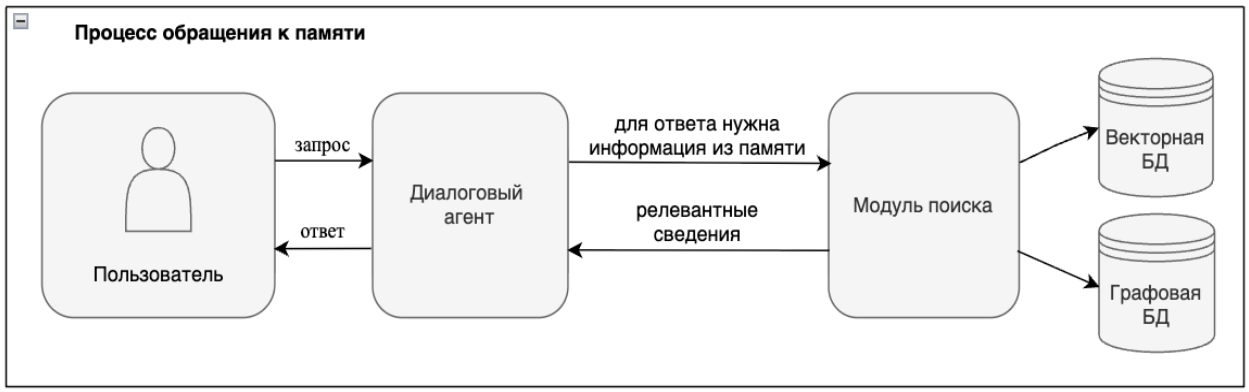


Рисунок 1 – Процессы обращения к долговременной памяти

На Рисунке 1 показан процесс использования памяти при формировании ответа. Пользователь направляет запрос диалоговому агенту. Если для ответа недостаточно информации из текущего контекста и требуются ранее сохраненные сведения, агент передает запрос в модуль поиска. Модуль поиска обращается к векторной и графовой базам данных: векторная база позволяет найти семантически близкие пользовательские сведения, а графовая база – получить связанные с ними факты и их контекст. Отобранные релевантные сведения возвращаются диалоговому агенту и используются при генерации ответа пользователю.

На Рисунке 2 показан отдельный процесс обновления памяти, который запускается при обработке сообщения пользователя. Сообщение поступает в оркестратор памяти, координирующий работу компонентов сохранения информации. Анализатор контекста определяет, содержит ли реплика сведения, значимые для будущих взаимодействий: предпочтения, ограничения, цели, интересы, данные о деятельности пользователя или ранее совершенных действиях. Это позволяет не сохранять в долговременную память служебные и ситуативные реплики, не несущие устойчивой пользовательской информации.

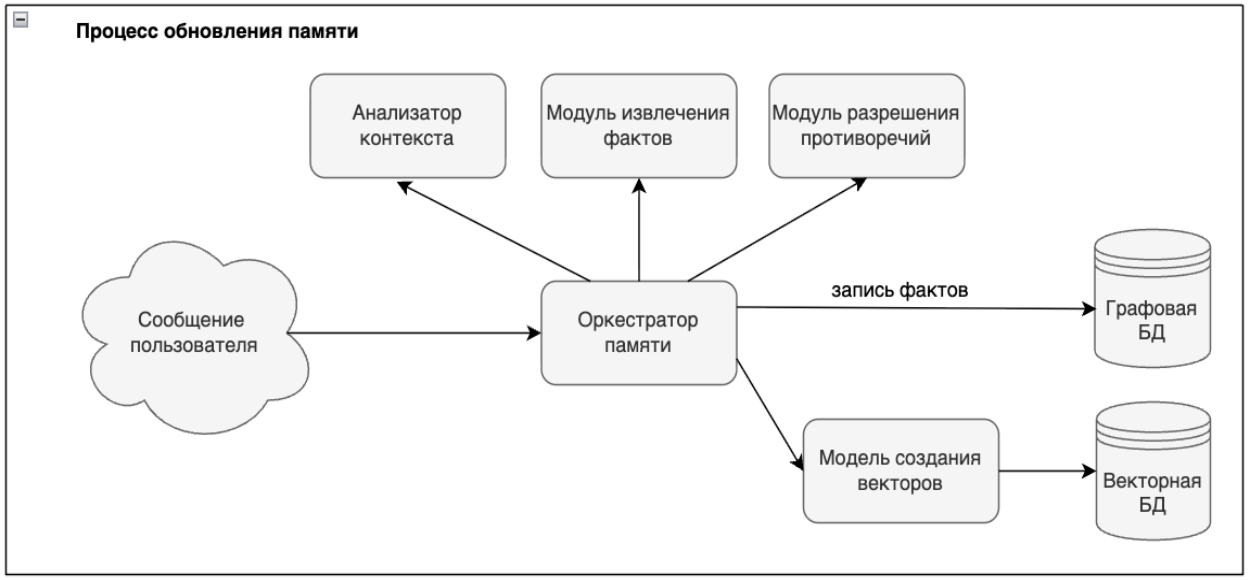


Рисунок 2 – Процесс обновления долговременной памяти

Модуль извлечения фактов преобразует значимые сведения из естественно-языкового сообщения в структурированное представление. Факт описан в форме связи между сущностью, отношением и значением, например: «пользователь – занимается футболом» или «пользователь – работает – сварщиком». Вместе с фактом может сохраняться связь с исходной репликой, благодаря чему в дальнейшем можно установить источник сведения и контекст, в котором оно было сообщено.

Модуль разрешения противоречий необходим для обработки изменений пользовательских сведений. При поступлении нового факта он сопоставляется с ранее сохраненной информацией: новый факт может подтверждать существующий, уточнять его либо заменять как более актуальный. Например, изменение места работы или отказ от прежнего предпочтения не должны приводить к одновременному использованию устаревшей и актуальной информации при формировании ответа.

После обработки структурированные факты записываются в графовую базу данных, где сохраняются связи между пользователем, фактами, исходными сообщениями и их изменениями во времени. Параллельно модель создания векторов формирует векторное представление сохраняемой информации и помещает его в векторную базу данных. Таким образом, графовая база обеспечивает структурное и согласованное хранение фактов, а векторная база – быстрый поиск сведений, близких по смыслу к новому запросу.

В классическом RAG-подходе поиск обычно ограничивается возвращением текстовых фрагментов, наиболее близких к запросу по векторному представлению. В предлагаемом механизме найденные сведения дополнительно рассматриваются с учетом связей и актуальности фактов в графовой базе данных. За счет этого агент получает не всю историю диалога и не набор изолированных фрагментов, а компактный набор релевантных и согласованных пользовательских сведений, необходимых для формирования персонализированного ответа.

3. Детали реализации

Прототип реализован на языке Python как отдельный сервис долговременной памяти, взаимодействующий с диалоговым агентом через REST API на базе FastAPI [5]. Для извлечения пользовательских фактов и генерации ответов используется мультиязычная модель qwen3.5-9b [6]. Для построения векторных представлений фактов и запросов используется модель эмбеддингов bge-m3, поддерживающая более 100 языков и различные режимы информационного поиска [7].

При обработке пользовательской реплики модель на основе системного промпта извлекает структурированные сведения: сущность, отношение, значение и источник факта. Выделенные сведения сохраняются в графовой базе данных Neo4j, основанной на представлении данных в виде узлов, связей и их свойств [8]. В графе фиксируются связи фактов с сущностями и исходными репликами, а также обновления ранее сохраненной информации.

Embedding-векторы фактов сохраняются в векторной базе данных Qdrant совместно с идентификаторами соответствующих узлов Neo4j. Qdrant используется для семантического поиска релевантных фактов по векторному представлению пользовательского запроса [9].

Найденные кандидаты дополнительно уточняются по графу с учетом связей и актуальности сведений.

Отобранные факты добавляются в системный промпт qwen3.5-9b как персонализирующий контекст для генерации ответа. При отсутствии подтвержденных данных модель получает инструкцию не формировать предположений о пользователе. Таким образом, сервис объединяет LLM-извлечение фактов, графовое хранение памяти в Neo4j, семантический поиск в Qdrant и генерацию ответа с опорой на извлеченный пользовательский контекст.

4. Методология эксперимента

Экспериментальная оценка проводилась как сравнительный анализ трёх подходов к организации пользовательской памяти в многосессионном диалоге: контекстного окна, классического RAG и предложенного метода на основе графового представления памяти.

В качестве данных использовались диалоги из соревнования AI Journey Contest 2025, посвященного задачам долговременной памяти [10]. В датасете четыре диалога, каждый диалог содержит около 100 000 токенов (300 000 символов). Для оценки подготовлены 100 контрольных вопросов (по 25 вопросов на каждый из диалогов). Вопросы требовали обращения к сведениям, ранее встречавшимся в диалоге: фактам о пользователе, его интересах и т.д. Для каждого вопроса задан эталонный ответ или эталонный факт. Вопросы делятся на 4 типа для возможности глубокой оценки системы. В таблице 2 приведена подробная информация о типах вопросов.

Таблица 2 – Типы вопросов

Тип вопроса	Что проверяет?
fact_equal_session	Извлечение факта, содержащегося в одной конкретной сессии
info_consolidation	Объединение сведений из нескольких сессий
info_updating	Учет изменения фактов во времени и выбор актуальной версии
no_info	Определение отсутствия ответа в истории диалога

Таким образом, как видно из Таблицы 2, для достижения высокого значения метрики необходимо не только находить релевантные факты в диалоге, но и определять случаи, когда необходимая информация не была предоставлена. Это позволяет снизить вероятность галлюцинаций модели.

Отдельное требование к эксперименту состояло в использовании локально разворачиваемых моделей малого и среднего размера. Такой выбор связан с прикладным сценарием работы модуля памяти. Во-первых, отдельный модуль памяти не должен резко увеличивать стоимость инференса. Во-вторых, открытые модели доступны компаниям независимо от конкретного облачного провайдера и могут использоваться в разных юрисдикциях. В-третьих, локальное развертывание снижает риски передачи пользовательских и корпоративных данных третьим лицам. В-четвертых, малые и средние компании не всегда имеют ресурсы для постоянного использования крупных коммерческих моделей.

Качество ответов оценивалось с использованием подхода LLM-as-a-judge. Для каждого вопроса сгенерированный ответ сопоставлялся с эталонным ответом, сформированным на основе разметки диалога. Ответ считался корректным, если он содержал требуемую информацию, не противоречил эталону и не добавлял неподтверждённых сведений. Основной метрикой выступала accuracy, отражающая долю вопросов, на которые система дала корректный персонализированный ответ. В качестве судьи в работе использовалась модель от OpenAI ChatGPT 5.5 [11].

5. Результаты исследования и их анализ

Результаты экспериментального сравнения приведены в Таблице 3.

Таблица 3 – Результаты экспериментального сравнения механизмов памяти

Подход	**Accuracy, %**	**Интерпретация результата**
Контекстное окно	75	Модель находит часть фактов, но теряется в большой количестве деталей
Классический RAG	63	Извлекаются тематически близкие фрагменты, но сведения не всегда согласованы
Предложенный механизм	88	Извлекается связанный и актуальный набор пользовательских фактов

Как видно из Таблицы 3, предложенный механизм достиг точности 88%. Это на 13 процентных пунктов выше результата режима с использованием контекстного окна и на 25 процентных пунктов выше результата классического RAG. Следовательно, для задачи персонализации важно не только вынести память за пределы контекстного окна, но и структурировать пользовательские сведения.

Результат классического RAG оказался ниже результата контекстного окна. Это объясняется тем, что при работе с длинными диалогами текстовый поиск часто возвращает фрагменты, близкие к запросу по теме, но не обязательно содержащие правильный факт о конкретном пользователе. Если в диалоге обсуждались похожие темы, другие люди или несколько разных ситуаций, RAG может извлечь лексически близкую информацию без нужной персональной связи.

Контекстное окно в части случаев оказывается устойчивее, потому что модель видит больше исходной истории и может использовать косвенные признаки. Однако этот подход плохо масштабируется: модель путается в деталях, хуже работает с фактами из середины длинной последовательности и может смешивать сведения из разных частей диалога.

Предложенный механизм показывает лучший результат за счет сочетания двух принципов. Векторный поиск находит кандидатов на релевантные сведения, а граф знаний восстанавливает связи между фактами и выбирает согласованный контекст. Например, если вопрос касается работы пользователя, система извлекает не просто фрагмент, где встречается слово «работа», а факт, связанный именно с сущностью пользователя и его профессией.

Анализ ошибок показал, что для режима контекстного окна характерны ошибки, связанные с перегрузкой длинным контекстом: модель путается в многочисленных деталях, забывает сведения из середины диалога или дает слишком общий ответ. Для классического RAG основная проблема заключается в смешении близких по теме фрагментов: система может извлечь данные о другом человеке, другом эпизоде или похожей теме. Для предложенного механизма ошибки связаны преимущественно с выбором нужного набора текстовых реплик. Необходимо совершенствовать и улучшать методы поиска по сформированным фактам о пользователе.

В ходе эксперимента рассмотрены отдельные примеры ответов. На вопрос «Каким спортом я занимаюсь?» система с предложенной памятью корректно отвечала, что пользователь занимается футболом и играет в него около 15 лет. На вопрос «Кем я работаю?» система извлекала факт о работе сварщиком на нефтеперерабатывающем заводе. На вопрос о породе собаки система находила сведения о мальтийской болонке. В случае вопроса о предпочтениях в сигаретах система не выдумывала факт, а отвечала, что такой информации в памяти нет. Последний пример показывает, что качественная память должна не только извлекать сохраненные факты, но и не создавать ложную персонализацию при отсутствии данных.

Таким образом, эксперимент подтвердил, что графовое представление пользовательских знаний повышает точность персонализированных ответов и делает работу памяти более управляемой.

Заключение

В работе предложен и экспериментально проверен механизм долговременной памяти LLM-агента, сочетающий векторный поиск и граф знаний пользователя. Такой подход позволяет хранить пользовательский контекст не как набор разрозненных текстовых фрагментов, а как связанную структуру фактов, сущностей, реплик и актуальных версий сведений.

Эксперимент на длинных многосессионных диалогах показал, что предложенный механизм достигает accuracy 88% и превосходит контекстное окно и классический RAG. Полученные результаты подтверждают, что для персонализации LLM-агента важен не только поиск релевантной информации, но и ее структурирование с учетом связей, источников и обновлений.

Список литературы

1. Лю Н. Ф. Затерянный посередине: как языковые модели используют длинные контексты / Н. Ф. Лю, К. Лин, Дж. Хьюитт, А. Паранджапе, М. Бевилаква, Ф. Петрони, П. Лян // Труды Ассоциации компьютерной лингвистики. — 2024. — Том 12. — С. 157-173. — arXiv: 2307.03172.
2. Ху Э. Дж. ЛоРА: низкоуровневая адаптация больших языковых моделей / Э. Дж. Ху, Ю. Шен, П. Уоллис, З. Аллен-Чжу, Ю. Ли, С. Ван, Л. Ванг, У. Чен. - 2021. — arXiv: 2106.09685.
3. Льюис П. Поисково-расширенная генерация для наукоемких задач НЛП / П. Льюис, Э. Перес, А. Пиктус, Ф. Петрони, В. Карпухин, Н. Гоял, Х. Кюттлер, М. Льюис, У. Йих, Т.

Роктешель, С. Ридель, Д. Кила // Достижения в области нейронных систем обработки информации. — 2020. — Том 33. — С. 9459-9474.

4. Эдж Д. От локального к глобальному: графический подход к обобщению, ориентированному на запросы / Д. Эдж, Х. Тринь, Н. Ченг, Дж. Брэдли, А. Чао, А. Моуди, С. Труитт, Д. Метрополитански, Р. О. Несс, Дж. Ларсон. - 2024. — arXiv: 2404.16130.
5. FastAPI : официальная документация. — URL: https://fastapi.tiangolo.com/ (дата обращения: 06.05.2026). — Текст : электронный.
6. Команда Qwen. Qwen3: Думай глубже, действуй быстрее : [сайт]. — 2025. — URL: https://qwenlm.github.io/blog/qwen3 / (дата обращения: 06.05.2026). — Текст : электронный.
7. Чен Дж. BGE M3-Встраивание: Многоязычное, многофункциональное, многозернистое встраивание текстов посредством самопознания / Дж. Чен, С. Сяо, П. Чжан, К. Луо, Д. Лиан, З. Лю // Выводы Ассоциации компьютерной лингвистики: ACL 2024. — 2024. — arXiv: 2402.03216.
8. Neo4j. Концепции графовых баз данных : официальная документация. — URL: https://neo4j.com/docs/getting-started/graph-database/ (дата обращения: 06.05.2026). — Текст : электронный.
9. Qdrant. Документация : официальная документация. — URL: https://qdrant.tech/documentation/ (дата обращения: 06.05.2026). — Текст : электронный.
10. Набор данных памяти AIJ25 : [сайт] / DSWorks. — URL: https://dsworks.ru/en/champ/aij25-memory/data (дата обращения: 06.05.2026). — Текст : электронный.
11. Чат : официальный сайт OpenAI. — URL: https://chatgpt.com / (дата обращения: 06.05.2026). — Текст : электронный.

References

1. Liu N. F. Lost in the Middle: How Language Models Use Long Contexts / N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang // Transactions of the Association for Computational Linguistics. — 2024. — Vol. 12. — P. 157–173. — arXiv:2307.03172.
2. Hu E. J. LoRA: Low-Rank Adaptation of Large Language Models / E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. — 2021. — arXiv:2106.09685.
3. Lewis P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 9459–9474.
4. Edge D. From Local to Global: A Graph RAG Approach to Query-Focused Summarization / D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson. — 2024. — arXiv:2404.16130.
5. FastAPI : official documentation. — URL: https://fastapi.tiangolo.com/ (дата обращения: 06.05.2026). — Текст : электронный.
6. Qwen Team. Qwen3: Think Deeper, Act Faster : [сайт]. — 2025. — URL: https://qwenlm.github.io/blog/qwen3/ (дата обращения: 06.05.2026). — Текст : электронный.

7. Chen J. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation / J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu // Findings of the Association for Computational Linguistics: ACL 2024. — 2024. — arXiv:2402.03216.
8. Neo4j. Graph database concepts : official documentation. — URL: https://neo4j.com/docs/getting-started/graph-database/ (дата обращения: 06.05.2026). — Текст : электронный.
9. Qdrant. Documentation : official documentation. — URL: https://qdrant.tech/documentation/ (дата обращения: 06.05.2026). — Текст : электронный.
10. AIJ25 Memory Dataset : [сайт] / DSWorks. — URL: https://dsworks.ru/en/champ/aij25-memory/data (дата обращения: 06.05.2026). — Текст : электронный.
11. ChatGPT : официальный сайт OpenAI. — URL: https://chatgpt.com/ (дата обращения: 06.05.2026). — Текст : электронный.

Волегов А.А. Совершенствование складской логистики на основе роботизации и технологии цифрового двойника (на примере предприятия пищевой промышленности)// Международный журнал информационных технологий и энергоэффективности.– 2026. – Т. 11 № 7(69) Ч. 1 с. 91–98

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.94: 621.865.8: 664

СОВЕРШЕНСТВОВАНИЕ СКЛАДСКОЙ ЛОГИСТИКИ НА ОСНОВЕ РОБОТИЗАЦИИ И ТЕХНОЛОГИИ ЦИФРОВОГО ДВОЙНИКА (НА ПРИМЕРЕ ПРЕДПРИЯТИЯ ПИЩЕВОЙ ПРОМЫШЛЕННОСТИ)

Волегов А.А.
АНО ВО "МОСКОВСКИЙ МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ", Москва, Россия (125040, город Москва, Ленинградский пр-кт, д. 17), e-mai: andreye308@gmail.com

Статья подготовлена в рамках магистерского исследования и посвящена вопросам дальнейшего повышения эффективности складской логистики предприятий пищевой промышленности в условиях цифровой трансформации. В работе рассмотрены перспективные направления развития складских систем после внедрения WMS-технологий, включая роботизацию складских операций и применение технологии цифрового двойника. Предложена концепция интеллектуальной модели управления складом, предусматривающая использование автономных мобильных роботов для перемещения грузов и цифровой модели склада для имитационного моделирования и оптимизации логистических потоков. Проведена оценка потенциального экономического эффекта от внедрения данных решений, включая рост производительности, снижение операционных затрат и повышение точности выполнения складских операций. Результаты исследования могут быть использованы при разработке программ автоматизации и модернизации складской инфраструктуры предприятий пищевой промышленности.

Ключевые слова: Складская логистика, роботизация склада, цифровой двойник, автоматизация складских процессов, цифровая трансформация логистики, операционная эффективность.

IMPROVEMENT OF WAREHOUSE LOGISTICS BASED ON ROBOTICS AND DIGITAL TWIN TECHNOLOGY (USING THE EXAMPLE OF A FOOD INDUSTRY ENTERPRISE)

Volegov A.A.
MOSCOW INTERNATIONAL UNIVERSITY, Moscow, Russia (125040, Moscow, Leningradskiy prospekt, 17), e-mail: andreye308@gmail.com

The article is prepared within the framework of a master's degree research and focuses on further improving the efficiency of warehouse logistics in food industry enterprises under conditions of digital transformation. The study examines advanced directions for the development of warehouse systems following the implementation of WMS technologies, including warehouse robotics and the use of digital twin technology. A concept of an intelligent warehouse management model is proposed, involving the use of autonomous mobile robots for cargo movement and a digital warehouse model for simulation and optimization of logistics flows. The potential economic effect of implementing these solutions is evaluated, including increased productivity, reduced operating costs, and improved accuracy of warehouse operations. The results of the study can be applied in the development of automation and modernization programs for warehouse infrastructure in food industry enterprises.

Keywords: Warehouse logistics, warehouse robotics, digital twin, warehouse automation, logistics digital transformation, operational efficiency.

Введение

Развитие складской логистики в условиях цифровой трансформации становится одним из ключевых факторов повышения конкурентоспособности предприятий пищевой

промышленности. Рост объёмов товарных потоков, усложнение логистических цепей и необходимость строгого соблюдения требований к срокам годности продукции требуют повышения точности, скорости и прозрачности складских операций. Особую значимость данные факторы приобретают в пищевом секторе, специфика которого диктует необходимость соблюдения жестких температурных режимов (управление «холодовой цепью»), высокой скорости оборачиваемости запасов (FMCG) и беспрекословного следования принципу FEFO (First Expire, First Out).

Даже при повсеместном внедрении WMS-систем сохраняется критическая зависимость физических процессов от человеческого фактора. В условиях современного нарастающего дефицита линейного складского персонала и тяжелых условий труда (например, работа в холодильных камерах), классическая автоматизация учета исчерпывает свой потенциал. Это ограничивает возможности дальнейшего роста производительности и качества выполнения операций.

В последние годы в научной и прикладной литературе активно рассматриваются вопросы роботизации складских комплексов и применение технологий цифровых двойников (Digital Twin) для моделирования логистических процессов. Использование автономных мобильных роботов (AMR), автоматизированных систем перемещения грузов и имитационного моделирования позволяет значительно повысить эффективность управления материальными потоками, снизить количество ошибок и повысить устойчивость логистических систем к пиковым сезонным нагрузкам. Однако практические аспекты синергетического внедрения роботизированных решений и технологий цифрового моделирования (когда виртуальная модель в реальном времени управляет физическими роботами) на предприятиях пищевой промышленности остаются недостаточно исследованными.

Целью данного исследования является разработка концепции дальнейшего развития складской логистики предприятия пищевой промышленности на основе роботизации складских операций и применения технологии цифрового двойника.

Для достижения поставленной цели в работе решаются следующие задачи: рассматриваются возможности интеграции роботизированных систем перемещения с существующей WMS-инфраструктурой, а также обосновывается использование цифровой модели склада для анализа, предиктивного прогнозирования и оптимизации логистических потоков. Объектом исследования выступает складской комплекс предприятия пищевой промышленности, предметом — процесс трансформации логистических операций на базе технологий Индустрии 4.0.[1]

Реализация предложенных решений направлена на повышение производительности складских операций, снижение операционных затрат и формирование устойчивой интеллектуальной архитектуры управления складским комплексом.

1. Методология и организация исследования

Исследование выполнено на основе ретроспективного анализа логистических процессов складского хозяйства предприятия пищевой промышленности, на котором ранее была реализована цифровизация складских операций посредством внедрения WMS-системы. В качестве методологической базы использован системный подход, позволяющий

рассматривать складской комплекс как интегрированную логистическую систему, в которой материальные, информационные и технологические потоки тесно взаимосвязаны. Для достижения целей исследования применялись методы процессного анализа, хронометража складских операций, дискретно-событийного имитационного моделирования, а также инструменты экономической оценки инвестиционных проектов.

Работа проводилась в несколько этапов. На первом этапе осуществлён анализ функционирования складской системы после внедрения WMS-технологии. Информационной базой послужили массивы исторических данных (логи), выгруженные из системы управления складом. [2] Были изучены показатели производительности складских операций, структура трудозатрат персонала, интенсивность перемещения грузов внутри склада и уровень загрузки складской инфраструктуры с учетом соблюдения требований к условиям хранения пищевой продукции. Проведённый анализ позволил выявить ограничения, связанные с высокой долей ручных операций и зависимостью эффективности процессов от человеческого фактора.

На втором этапе сформирована концептуальная модель дальнейшего развития складской системы, предусматривающая внедрение роботизированных средств перемещения грузов (в частности, автономных мобильных роботов) и использование технологии цифрового двойника склада. Цифровой двойник представляет собой виртуальную модель складского комплекса, которая посредством двустороннего обмена данными с WMS воспроизводит структуру склада, размещение зон хранения, движение материальных потоков и работу оборудования в режиме реального времени. Применение данной технологии позволяет проводить имитационное моделирование логистических процессов, анализировать различные сценарии организации складских операций (включая стресс-тестирование при пиковых сезонных нагрузках) и определять наиболее эффективные конфигурации логистической инфраструктуры до их практического внедрения.

На заключительном этапе выполнена оценка потенциального экономического эффекта от реализации предложенных решений. [3] В рамках анализа рассмотрены показатели изменения производительности труда, снижения операционных затрат (OPEX), сокращения времени выполнения складских операций и повышения точности обработки заказов. Для оценки целесообразности капитальных вложений использовались показатели окупаемости инвестиций (ROI) и чистого дисконтированного дохода (NPV). Полученные результаты позволили математически обосновать практическую значимость внедрения роботизированных технологий и инструментов цифрового моделирования для дальнейшего повышения эффективности складской логистики предприятий пищевой промышленности.

2. Концепция роботизации складских процессов

Анализ складских операций показал, что наибольшая доля трудозатрат приходится на внутрискладскую транспортировку продукции (горизонтальную интралогистику) между зонами приёмки, хранения, комплектации и отгрузки. [4] Значительная часть времени персонала расходуется на перемещение паллет с использованием ручных гидравлических тележек и классической погрузочной техники, что классифицируется как непродуктивные потери рабочего времени. Это приводит к увеличению цикла выполнения операций, неравномерной загрузке сотрудников и росту вероятности операционных ошибок.

Для повышения эффективности складских процессов предложено внедрение автономных мобильных роботов (AMR), выполняющих функции внутрискладской транспортировки. Роботы перемещают паллеты между основными функциональными зонами склада: от зоны приёмки к адресным ячейкам хранения, из зоны хранения к участкам комплектации, а также к зоне отгрузки. Задания на перемещение формируются WMS-системой автоматически на основе транзакций (приёмка, пополнение, комплектация заказов) через специализированный программный интерфейс (API). Навигация роботов осуществляется с использованием встроенных датчиков (LiDAR, видеокамеры) и технологии одновременной локализации и картирования (SLAM). В отличие от традиционных AGV-систем, требующих физических направляющих, AMR-роботы способны динамически перестраивать маршруты при возникновении препятствий, что позволяет им безопасно перемещаться в общей рабочей зоне с персоналом и погрузочной техникой.

Внедрение роботизированной транспортной системы позволяет радикально сократить время внутрискладских перемещений, снизить долю ручного труда и повысить стабильность выполнения логистических операций. Автоматизация транспортировки обеспечивает более равномерную загрузку складской инфраструктуры и персонала, а также создаёт гибкие условия для дальнейшего масштабирования складских процессов при росте объёмов товарооборота.

3. Применение технологии цифрового двойника склада

Для анализа и оптимизации логистических процессов предложено использование технологии цифрового двойника складского комплекса. Цифровой двойник представляет собой виртуальную модель склада, воспроизводящую его планировку, размещение зон хранения, параметры складского оборудования и движение материальных потоков на основе реальных данных из WMS. Такая модель позволяет проводить имитационное моделирование складских операций и предиктивно оценивать последствия изменений до их физического внедрения в реальную систему.[5]

Специфика предприятия пищевой промышленности, производящего кондитерские изделия (печенье), заключается в необходимости оперативной обработки больших партий готовой продукции, обеспечении сохранности хрупкого груза и строгом контроле сроков годности. В условиях высокой интенсивности отгрузок важно обеспечить равномерное распределение потоков между зонами приёмки, хранения и комплектации, а также безукоризненное соблюдение принципа FEFO при отборе продукции. Цифровой двойник склада позволяет моделировать различные сценарии размещения партий печенья, конфигурации зон хранения и маршрутов внутрискладской транспортировки.

На основе виртуальной модели возможно анализировать загрузку складских зон, интенсивность перемещения грузов и время выполнения ключевых операций. Моделирование позволяет в рамках сценарного анализа «что-если» выявлять узкие места в организации складских процессов, оценивать влияние внедрения роботизированных систем и оптимизировать маршруты перемещения продукции при пиковых (например, предпраздничных) нагрузках. Использование цифрового двойника создаёт мощный предиктивный инструмент для обоснованного принятия управленческих решений и

позволяет минимизировать финансовые и операционные риски при модернизации складской инфраструктуры предприятия.

4. Оценка экономической эффективности

Для обоснования экономической целесообразности предложенных решений была проведена оценка ключевых показателей работы склада готовой продукции (кондитерские изделия — печенье). Оценка базируется на сравнении модели «As-Is» (текущее состояние с ручным трудом) и модели «To-Be» (целевое состояние с применением AMR и цифрового двойника).

Анализ текущего состояния (Модель «As-Is»): на текущий момент базовая внутрискладская логистика (перемещения между зонами приёмки с производства, хранения, комплектации и отгрузки) осуществляется персоналом с использованием ручных гидравлических тележек и вилочных погрузчиков. В среднем за одну смену выполняется 180–220 перемещений паллет с готовой продукцией. Учитывая время на поиск техники, маневрирование и сам процесс движения, среднее время одного перемещения составляет 4–5 минут. В результате суммарные непродуктивные трудозатраты исключительно на транспортировку достигают 14–16 человеко-часов за смену (что эквивалентно постоянной занятости двух сотрудников в смену, выполняющих только функцию перемещения грузов).

Эффект от внедрения роботизации (Модель «To-Be»): внедрение системы из трёх автономных мобильных роботов (AMR) позволяет полностью автоматизировать данный процесс. Производительность одного робота составляет 35–40 транспортных операций в час. Парк из трёх единиц способен выполнять до 300 операций за смену, что формирует резерв мощности (+35% к текущей потребности) для покрытия пиковых сезонных нагрузок. Время выполнения одной операции сокращается до 2–3 минут за счёт алгоритмической оптимизации маршрутов WMS-системой и отсутствия простоев техники.[6]

Трансформация операционных показателей конвертируется в следующие прямые экономические эффекты:

1. Оптимизация фонда оплаты труда (ФОТ) - высвобождение 14–16 человеко-часов в смену позволяет перераспределить персонал на участки, требующие сложной моторики и когнитивных решений (например, штучная комплектация сборных заказов), либо оптимизировать штатное расписание, снизив расходы на ФОТ и сопутствующие налоги.

2. Снижение потерь от брака (порчи продукции) - специфика кондитерских изделий (печенья) требует бережного отношения к грузу. Плавность хода AMR-роботов, контролируемая программно, практически исключает риск падения паллет и повреждения хрупкой продукции, что характерно для ручного управления погрузчиками (снижение статьи расходов на списание брака).

3. Минимизация расходов на эксплуатацию ПТО - сокращается износ парка классической погрузочной техники и затраты на её техническое обслуживание.

Использование цифровой модели выступает катализатором эффективности роботизированной системы. Симуляция потоков готовой продукции позволяет найти оптимальные топологические решения без физического вмешательства в работу склада. Моделирование показало, что динамическое перераспределение наиболее оборачиваемых партий печенья ближе к зоне отгрузки и устранение маршрутных пересечений

(предотвращение заторов роботов) позволяет дополнительно сократить общее время обработки заказов на 15–20 %. Кроме того, цифровой двойник строго контролирует соблюдение партионного учета (FEFO), исключая человеческий фактор при отборе паллет, что для пищевой промышленности.

Заключение

В ходе проведённого исследования решена актуальная научно-практическая задача по совершенствованию складской логистики предприятия пищевой промышленности (на примере производства кондитерских изделий). Анализ действующей системы складского хозяйства показал, что классическая автоматизация на базе WMS-системы достигла предела своей эффективности. Выявлены критические ограничения, связанные с высокой долей ручных операций, значительными затратами времени на внутрискладскую горизонтальную транспортировку и зависимостью логистических процессов от человеческого фактора.

Для преодоления выявленных ограничений и перехода предприятия к стандартам Индустрии 4.0 была разработана целевая интеллектуальная модель управления складом. Ключевым направлением модернизации стала синергия трех элементов: существующей WMS-системы, роботизации внутрискладской транспортировки (с применением автономных мобильных роботов — AMR) и технологии цифрового двойника. Использование виртуальной модели позволило реализовать предиктивный подход к управлению складскими потоками, динамически оптимизировать адресное пространство и гарантировать строгий контроль отгрузок по принципу FEFO.

Реализация предложенного комплекса мероприятий позволяет радикально сократить цикл выполнения складских операций, нивелировать трудоёмкость ручных перемещений и повысить точность комплектации заказов. Исключение человеческого фактора при отборе партий обеспечивает безукоризненное управление сроками годности скоропортящейся продукции, а плавность работы роботизированной техники минимизирует бой хрупких кондитерских изделий, что суммарно ведет к резкому снижению финансовых потерь от списаний.

Проведённая оценка экономической эффективности подтвердила, что интеграция AMR-систем и цифрового моделирования обеспечивает переход склада на интенсивный путь развития. Несмотря на капитальные затраты на этапе внедрения, решения гарантируют существенное снижение операционных издержек (OPEX) за счет оптимизации ФОТ и роста общей производительности.

Практическая значимость исследования заключается в формировании универсальной цифровой архитектуры логистических процессов. Разработанные подходы и алгоритмы могут быть масштабированы и использованы при модернизации складских комплексов любых предприятий пищевой промышленности (FMCG-сектора), характеризующихся массовыми товарными потоками, жесткими требованиями к скорости обработки и строгим контролем температурных режимов и сроков годности.

Список литературы

1. Столяров А. Д., Файзуллина А. М., Абрамов В. И. Цифровая трансформация логистики предприятия с использованием цифровых двойников // Beneficium. 2024. №2 (51). URL:

https://cyberleninka.ru/article/n/tsifrovaya-transformatsiya-logistiki-predpriyatiya-s-ispolzovaniem-tsifrovyh-dvoynikov (дата обращения: 18.04.2026).

2. Кархова Ирина Юрьевна Применение цифровых технологий в логистике // Российский внешнеэкономический вестник. 2025. №7. URL: https://cyberleninka.ru/article/n/primenenie-tsifrovyh-tehnologiy-v-logistike (дата обращения: 18.04.2026).
3. Юсуфова Ольга Михайловна, Шиболденков Владимир Александрович, Андреева Анна Андреевна АНАЛИЗ ТЕХНОЛОГИЙ ЦИФРОВОЙ ЛОГИСТИКИ ДЛЯ АВТОМАТИЗАЦИИ И СЕРВИСНОЙ ИНТЕГРАЦИИ СКЛАДСКИХ ПРОЦЕССОВ ОРГАНИЗАЦИИ // Вопросы инновационной экономики. 2020. №3. URL: https://cyberleninka.ru/article/n/analiz-tehnologiy-tsifrovoy-logistiki-dlya-avtomatizatsii-i-servisnoy-integratsii-skladskih-protsessov-organizatsii (дата обращения: 18.04.2026).
4. Ковалева И. В. Эффективное управление складской логистикой организации: WMS-система // Экономика и бизнес: теория и практика. — 2024. — № 1-1 (107). — URL: https://cyberleninka.ru/article/n/effektivnoe-upravlenie-skladskoy-logistikoy-organizatsii-wms-sistema (дата обращения: 27.04.2026).
5. Абросимов, М. А. Складская автоматизация на основе искусственного интеллекта / М. А. Абросимов. — Текст : непосредственный // Молодой ученый. — 2024. — № 38 (537). — С. 1-3. — URL: https://moluch.ru/archive/537/117715.
6. Имаи М. Гемба кайдзен: путь к снижению затрат и повышению качества / М. Имаи ; пер. с англ. Т. Гутман, Д. Савченко. — Москва : Альпина Паблишер, 2022. — 344 с.

References

1. Stolyarov A.D., Fayzullina A.M., Abramov V. I. Digital transformation of enterprise logistics using digital twins // Beneficium. 2024. №2 (51). URL: https://cyberleninka.ru/article/n/tsifrovaya-transformatsiya-logistiki-predpriyatiya-s-ispolzovaniem-tsifrovyh-dvoynikov (date of reference: 04/18/2026).
2. Karkhova Irina Yurievna Application of digital technologies in logistics // Russian Foreign Economic Bulletin. 2025. No. 7. URL: https://cyberleninka.ru/article/n/primenenie-tsifrovyh-tehnologiy-v-logistike (date of request: 04/18/2026).
3. Yusufova Olga Mikhailovna, Shiboldenkov Vladimir Alexandrovich, Andreeva Anna Andreevna ANALYSIS OF DIGITAL LOGISTICS TECHNOLOGIES FOR AUTOMATION AND SERVICE INTEGRATION OF WAREHOUSE PROCESSES OF THE ORGANIZATION // Issues of innovative economics. 2020. No. 3. URL: https://cyberleninka.ru/article/n/analiz-tehnologiy-tsifrovoy-logistiki-dlya-avtomatizatsii-i-servisnoy-integratsii-skladskih-protsessov-organizatsii (date of request: 04/18/2026).
4. Kovaleva I. V. Effective management of warehouse logistics of the organization: WMS-system // Economics and Business: theory and practice. — 2024. — № 1-1 (107). — URL: https://cyberleninka.ru/article/n/effektivnoe-upravlenie-skladskoy-logistikoy-organizatsii-wms-sistema (date of access: 04/27/2026).
5. Abrosimov, M. A. Warehouse automation based on artificial intelligence / M. A. Abrosimov. — Text : direct // Young scientist. — 2024. — № 38 (537). — pp. 1-3. — URL: https://moluch.ru/archive/537/117715 .

6. Imai M. Gemba kaizen: the way to reduce costs and improve quality / M. Imai; translated from English by T. Gutman, D. Savchenko. Moscow : Alpina Publisher, 2022. p.344

Соловяненко О.Ю. Иерархическая мультиагентная архитектура с супервайзором, специализированными агентами и адресной маршрутизацией итеративной самокоррекции для автоматизированной генерации регламентированных кредитных заключений// Международный журнал информационных технологий и энергоэффективности.– 2026. – Т. 11 № 7(69) Ч. 1 с. 99–109

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.855.5:336.648

ИЕРАРХИЧЕСКАЯ МУЛЬТИАГЕНТНАЯ АРХИТЕКТУРА С СУПЕРВАЙЗОРОМ, СПЕЦИАЛИЗИРОВАННЫМИ АГЕНТАМИ И АДРЕСНОЙ МАРШРУТИЗАЦИЕЙ ИТЕРАТИВНОЙ САМОКОРРЕКЦИИ ДЛЯ АВТОМАТИЗИРОВАННОЙ ГЕНЕРАЦИИ РЕГЛАМЕНТИРОВАННЫХ КРЕДИТНЫХ ЗАКЛЮЧЕНИЙ

Соловяненко О.Ю.
ФГАОУ ВО "КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ" , Москва, Россия (420008, Республика Татарстан, г. Казань, Кремлевская ул., д. 18 к. 1), e-mai: olegsolovyanenko@gmail.com

Ансамблевые модели градиентного бустинга обеспечивают высокую точность прогнозирования корпоративного дефолта, однако их «непрозрачность» ограничивает применение в комплаенсе и риск-менеджменте, а использование облачных больших языковых моделей для текстовой генерации недопустимо из-за требований банковской тайны. Известный в литературе подход на основе одиночного реактивного агента ReAct, координирующего вызовы прогнозного ансамбля и алгоритма аддитивных объяснений SHAP, обеспечивает заземление генерации на фактических данных, но оставляет нерешенными две методологические проблемы: разделение отраслевой интерпретации и регуляторной классификации по Положению Банка России № 590-П между специализированными агентами и адресную маршрутизацию замечаний критика, неприменимую в каноническом методе итеративной самокоррекции. В статье предложена иерархическая мультиагентная архитектура с явной графовой моделью состояний, в которой делиберативный супервайзор координирует четырех специализированных под-агентов, а вердикт критика тегируется одним из десяти нормированных классов и указывает целевой узел графа по зоне ответственности за обнаруженную проблему. Узлы конвейера разделены на критические и пропускаемые, а таблица каскадной маршрутизации придает конвейеру устойчивость при сбоях отдельных компонентов. Для эталонного корпуса XBRL-документов сформулирована многоуровневая методика валидации, опирающаяся на архитектурные гарантии уровня кода, критерии приемки автоматического критика, методику пилотного прогона и протокол приемочной экспертной оценки в локальном вычислительном контуре, исключающем передачу конфиденциальных данных во внешние сервисы.

Ключевые слова: Мультиагентные Системы; большие языковые модели; итеративная самокоррекция; адресная маршрутизация; прогнозирование банкротства; кредитное заключение; корпоративные финансы; банковская тайна.

A HIERARCHICAL SUPERVISOR–SPECIALIST MULTI-AGENT ARCHITECTURE WITH ADDRESSED SELF-REFINE ROUTING FOR AUTOMATED GENERATION OF REGULATED CREDIT REPORTS

Solovyanenko O.Yu.
"KAZAN (VOLG) FEDERAL UNIVERSITY" , Moscow, Russia (420008, Republic of Tatarstan, g. Kazan, Kremlevskaya ul., d. 18 k. 1), e-mai: olegsolovyanenko@gmail.com

Gradient-boosting ensemble models deliver high accuracy in predicting corporate defaults, yet their opacity constrains their use in compliance and risk management, while cloud-based large language models cannot be used for text generation owing to banking-secrecy requirements. The approach known in the literature based on a single

Соловяненко О.Ю. Иерархическая мультиагентная архитектура с супервайзором, специализированными агентами и адресной маршрутизацией итеративной самокоррекции для автоматизированной генерации регламентированных кредитных заключений// Международный журнал информационных технологий и энергоэффективности.– 2026. – Т. 11 № 7(69) Ч. 1 с. 99–109

ReAct reactive agent that orchestrates calls to a gradient-boosting ensemble and the SHAP additive-explanation algorithm provides grounded generation but leaves two methodological problems unresolved: the separation of industry-specific interpretation and regulatory classification under Bank of Russia Regulation No. 590-P across specialized agents, and the addressed routing of the critic's remarks that is unavailable in the canonical self-refine method. This paper proposes a hierarchical multi-agent architecture with an explicit state-graph model, in which a deliberative supervisor coordinates four specialized sub-agents, and the critic's verdict is tagged with one of ten normalized classes and points to the target graph node by area of responsibility for the detected problem. Pipeline nodes are split into critical and skippable, and a cascade routing table ensures pipeline resilience under component failures. A multi-layer evaluation methodology is formulated for a reference corpus of XBRL documents, combining code-level architectural guarantees, acceptance criteria of the automatic critic, the protocol of a pilot run, and a planned expert acceptance review, all confined to a local computational environment that excludes transmission of confidential data to external services.

Keywords: Multi-agent systems; large language models; self-refine; addressed routing; bankruptcy prediction; credit report; corporate finance; banking secrecy.

Введение

Автоматизированная генерация кредитных заключений по корпоративному заемщику средствами больших языковых моделей выходит за рамки прямого применения сетевой генеративной модели. Прямой инференс LLM по сырой бухгалтерской отчетности сопряжен с высокими рисками сразу по нескольким соображениям. Машиночитаемый формат XBRL требует специализированных правил извлечения и к естественно-языковому пониманию не сводится. Численная оценка вероятности дефолта должна оставаться продуктом откалиброванной статистической модели - стохастическая генерация токенов порождает риск фактически недостоверных значений. Наконец, объяснение предсказания должно опираться на фактическую SHAP-декомпозицию, а не на ретроспективную аргументацию, реконструированную моделью постфактум. Эти ограничения мотивируют переход к парадигме, в которой языковая модель координирует инструменты: за ней закрепляется функция планировщика и вербализатора, тогда как предметные операции делегируются специализированным программным компонентам. Канонической реализацией такой парадигмы выступает паттерн ReAct [12]. Применимость связки градиентного бустинга, SHAP-объяснений и локально развернутой языковой модели под управлением одиночного реактивного агента для построения связного аналитического текста по российской бухгалтерской отчетности показана в [3].

Одиночный реактивный агент, однако, оставляет открытыми два методологических вопроса. Первый из них - содержательная декомпозиция аналитической записки по зонам ответственности. Кредитное заключение соединяет два разнородных языка: отраслевой, оперирующий бенчмарками и нормами финансовых коэффициентов, и нормативный, оперирующий категориями качества ссуды Положения Банка России № 590-П [2] и соответствующими диапазонами расчетного резерва. Когда обе аргументации совмещаются в одном вызове модели, они начинают интерферировать друг с другом, и эту интерференцию крайне затруднительно снять исключительно подбором промпта - ее корни лежат в архитектуре. Естественным ответом становится переход к иерархической мультиагентной системе [9], в которой задача распадается на отраслевую интерпретацию, регуляторную классификацию и синтез финального отчета. Среду для такой декомпозиции дает шаблон Blackboard [4]: специалисты обмениваются результатами через общее состояние, не вступая в прямую коммуникацию.

Второй вопрос связан с итеративной самокоррекцией сгенерированного текста. Метод итеративной самокоррекции, формализованный в [7], разворачивает цикл «генерация - критика - улучшение». В его основе лежит эмпирическое наблюдение: задачи проверки текста на соответствие фиксированным критериям систематически проще задач генерации того же текста с нуля. Каноническая реализация, однако, не отвечает на ключевой вопрос мультиагентной системы - куда возвращать управление при обнаружении ошибки. У линейной цепочки специалистов сразу три ограничения. Адресная коррекция в ней структурно ограничена - критик, обнаруживший регуляторное несоответствие, вынужден направлять замечание общему автору, а не в регуляторный узел. Цепочка демонстрирует уязвимость к сбоям, поскольку отказ любого делиберативного агента способен прервать ее целиком. Наконец, она затрудняет проведение аудита, не оставляя формальной трассы переходов и распределения вердиктов критика по типам нарушений.

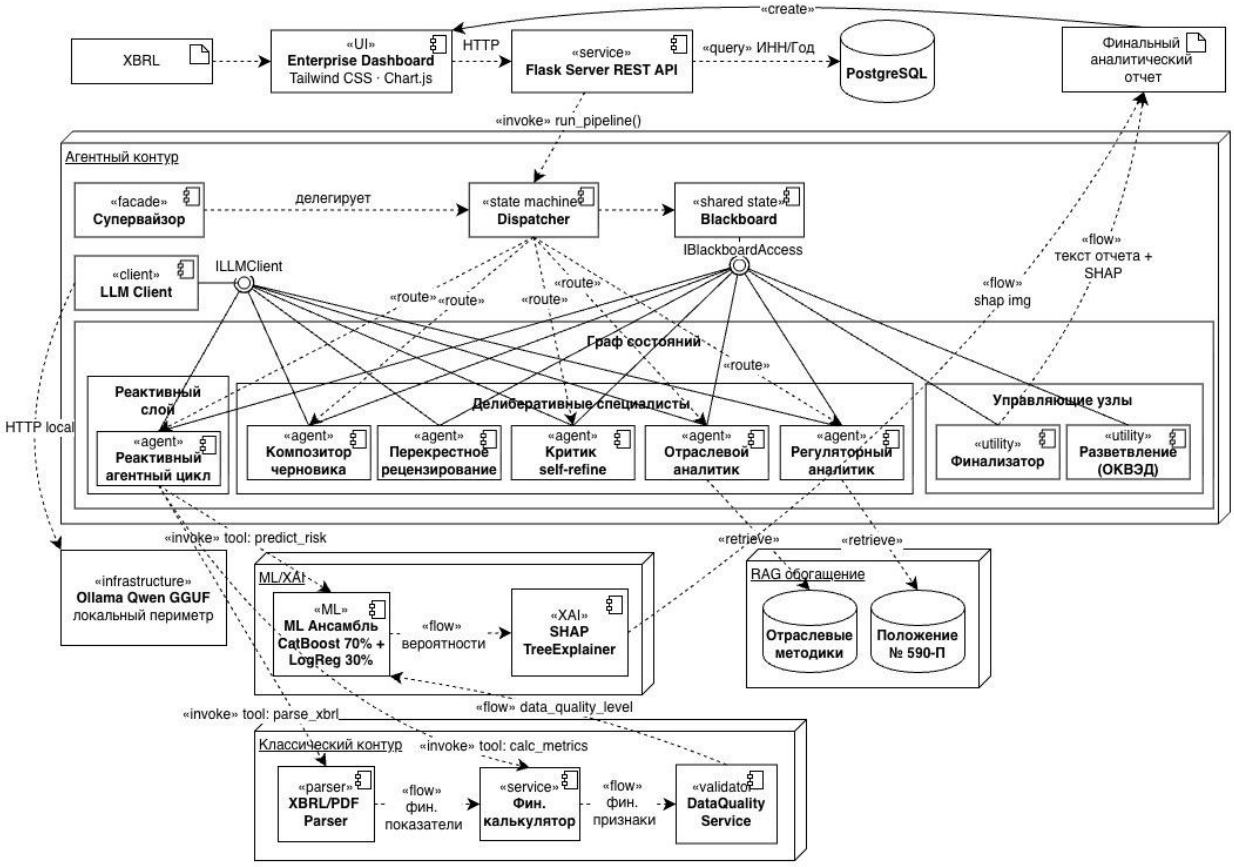
Целью настоящего исследования является разработка и валидация иерархической мультиагентной архитектуры с делиберативным супервайзором, специализированными отраслевым и регуляторным агентами и расширенным методом итеративной самокоррекции с адресной маршрутизацией. Научная новизна работы складывается из трех компонентов. Первый - явная графовая модель состояний с разделением узлов на критические и пропускаемые и с таблицей каскадной маршрутизации, придающей конвейеру устойчивость к сбоям отдельных компонентов. Второй - расширение канонического метода итеративной самокоррекции на мультиагентный контур: замечания критика адресно возвращаются в тот узел графа, на чьей зоне ответственности обнаружена проблема. Третий - интеграция механизма с RAG-обогащением [6] через раздельные обогащающие корпуса методических материалов и нормативных документов, обеспечивающие контекстную специализацию каждого агента.

Материалы и методы.

Базовый прогнозный слой системы реализован как ансамбль градиентного бустинга на категориальных признаках [8] и логистической регрессии, объединенных мягким голосованием с весами 0,7 и 0,3 соответственно. SHAP-объяснения [3] строятся через древовидный разложитель по бустинговой компоненте. Поверх прогнозного слоя функционирует одиночный реактивный агент ReAct [12] с делегированием предметных операций четырем детерминированным инструментам: парсингу XBRL-отчетности, валидации фактов, расчету двадцати финансовых коэффициентов и инференсу прогнозного ансамбля. Языковая модель работает в локальном вычислительном контуре в соответствии с требованиями статьи 26 Федерального закона № 395-1 о банковской тайне [1]. Конструкция реактивного слоя и протокол его взаимодействия с прогнозным ядром описаны в [3].

Над одиночным реактивным агентом надстраивается мультиагентный контур, организованный как явная графовая модель состояний из восьми именованных узлов: реактивный цикл, узел разветвления для извлечения кода отраслевой принадлежности, отраслевой и регуляторный анализ, опциональное перекрестное рецензирование, композиция чернового отчета, критик и финализация. Каждый узел реализуется чистой функцией над общим состоянием и сам возвращает имя следующего узла, а диспетчер лишь исполняет это

решение. Так логика маршрутизации локализуется в самих узлах. Состояние между узлами передается единым мутабельным объектом-доской [4]: каждый узел читает нужные ему поля и записывает свои результаты, не зная о соседях. Промежуточные артефакты - отраслевая и регуляторная записки, вердикт критика, черновик отчета - представлены строго типизированными моделями, что блокирует распространение некорректных выходов модели в нижестоящие узлы. Общая структурная схема контура представлена на Рисунке 1.



Рисуснок 1 - Структурная схема системы

Делиберативные специалисты выполняют ровно один вызов модели и возвращают строго типизированный результат, в отличие от реактивного агента, действующего в цикле. Все они реализуют общий шаблон: системный промпт, формирование пользовательского сообщения из доски, опциональное обогащение выдержками из собственного корпуса RAG [6], единственный вызов модели и парсинг ответа в типизированный объект с ограниченным числом повторных попыток.

Отраслевой специалист принимает на вход код отраслевой принадлежности и компактную сводку прогноза - вероятность дефолта, риск-уровень, по три ведущих фактора устойчивости и риска. Код предварительно проходит через статический справочник, из которого извлекается старший двузначный раздел с человекочитаемым названием. Агент опирается на выделенный корпус RAG отраслевых методических материалов. Результатом становится записка с риск-профилем отрасли, бенчмарками и интерпретацией прогноза сквозь отраслевую призму. Системный промпт явно запрещает агенту называть конкретные

численные значения коэффициентов, упоминать прогнозный аппарат и предлагать регуляторные категории.

Регуляторный специалист получает намеренно минимальный вход - вероятность дефолта, риск-уровень, уровень качества данных и до трех первых предупреждений - и опирается на собственный корпус RAG с выдержками из Положения № 590-П. Результат - записка с категорией качества ссуды, минимальным резервом, перечнем замечаний о соответствии Положению со ссылками на конкретные пункты и развернутой интерпретацией. Низкое качество входных данных Положение рассматривает как повод для консервативной классификации, и агент явно проинструктирован ужесточать категорию при низком уровне качества. Промпт запрещает изобретать абсолютные суммы долга и самостоятельно оценивать качество обслуживания долга, поскольку эти данные системе недоступны.

После отработки реактивного агента и специалистов доска содержит четыре артефакта - численную сводку, отраслевую и регуляторную записки, черновой текст реактивного цикла, - ни один из которых сам по себе финальным отчетом не является. Композитор выполняет акт синтеза: формирует отчет в жанре банковского кредитного отчета, обращенный к кредитному комитету. Отчет состоит из четырех связных блоков - резюме, драйверы устойчивости, зоны риска, рекомендации, - заданных типизированным DTO с количественными требованиями. Эмпирически установлено, что без явных количественных требований модель выдает минимально достаточные ответы, поэтому промпт композитора фиксирует строгие нормы: резюме 220–360 символов, блоки драйверов и зон риска - 450–650 символов из трех предложений каждый, рекомендации - 1200–1700 символов из 9–11 предложений с покрытием одиннадцати тематических пунктов. Промпт дополнен списком запретов: упоминание прогнозного аппарата, филлерные конструкции, изобретение абсолютных рублевых сумм, отступление от численной сводки.

Качество одиночной генерации композитора проверяется делиберативным критиком, реализующим расширенную петлю итеративной самокоррекции. В отличие от канонического метода [7], возвращающего текст одному автору, критик выступает не общим «улучшителем», а строгим валидатором по шести категориям критериев: числовая точность вероятности и риск-уровня, соответствие структурным требованиям к полям, отсутствие запрещенных конструкций, согласованность с отраслевой запиской, покрытие требований Положения № 590-П и корректность типов величин. Каждое замечание критика тегируется одним из десяти нормированных классов, а целевой узел возврата - композитор, отраслевой узел, регуляторный узел или реактивный цикл - определяется зоной ответственности за обнаруженную проблему. Схема расширенной петли с адресными возвратами представлена на Рисунке 2.

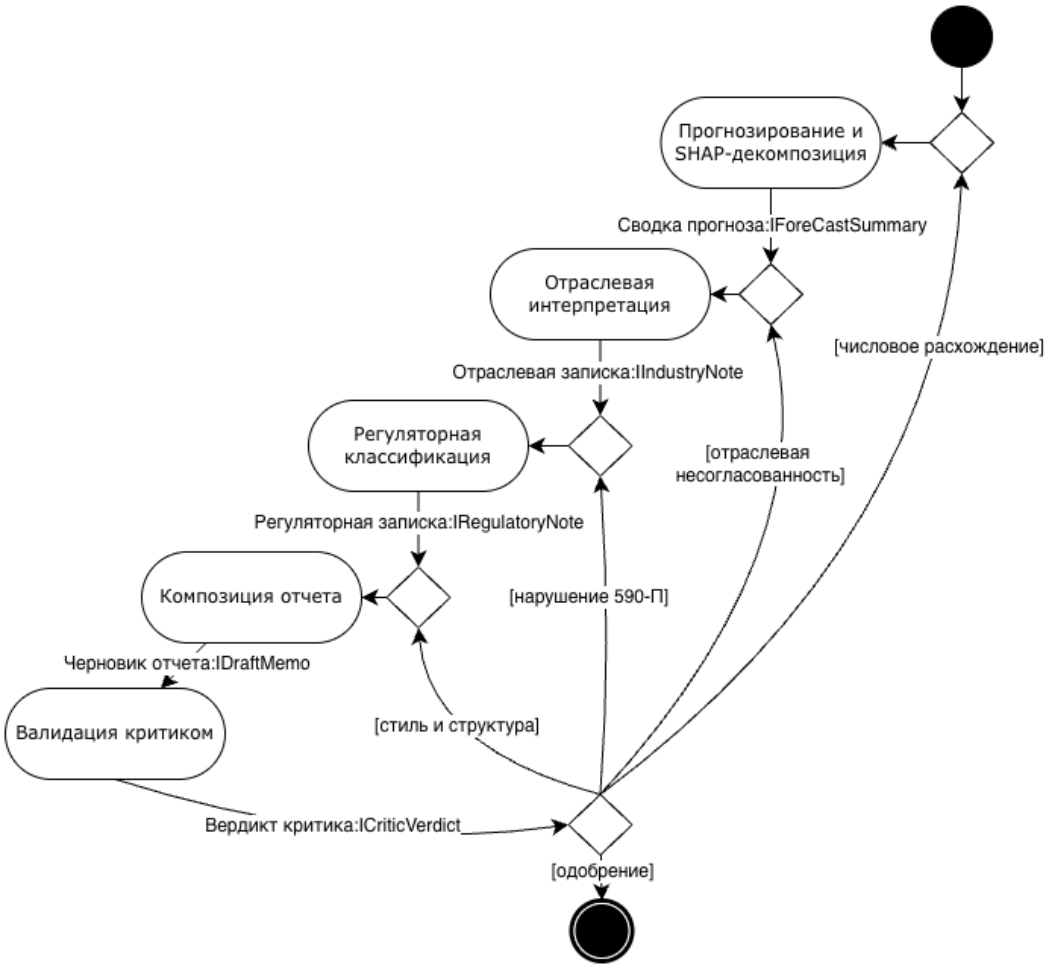


Рисунок 2 - Расширенная петля итеративной самокоррекции с адресной маршрутизацией

В каноническом методе итеративной самокоррекции [7] критик возвращает текст одному автору, поэтому при обнаружении регуляторного несоответствия повторно перегенерируется весь финальный отчет, включая корректные отраслевые и стилистические блоки. Адресная маршрутизация затрагивает только релевантный узел: то же регуляторное несоответствие приводит к повторному запуску одного регуляторного специалиста, а не всего конвейера, что снижает совокупную стоимость самокоррекции и устраняет источник бесконечных циклов пересмотра. Незавершимость петли исключается бюджетом раундов: при достижении лимита вердикт принудительно интерпретируется как одобряющий, а факт принудительного завершения фиксируется в журнале.

Устойчивость конвейера и каскадная деградация. В локальной банковской среде сетевой инстанс языковой модели нестабилен, поэтому простейшая схема, при которой сбой любого делиберативного агента прерывает весь конвейер, неприемлема: совокупная доступность мультипликативно деградирует по числу последовательных вызовов. Архитектурным ответом служит разделение узлов графа на два класса по их роли в формировании отчета. Критические узлы - реактивный цикл и узел разветвления - образуют

минимальный остов конвейера: без численной сводки и кода отраслевой принадлежности дальнейшие специалисты теряют входные данные, и сбой такого узла прерывает исполнение с явным кодом ошибки. Пропускаемые узлы - делиберативные специалисты, критик и опциональное перекрестное рецензирование - обогащают отчет, но не определяют сам факт его возможности. Их сбой обрабатывается маршрутизацией к следующему по конвейеру узлу по детерминированной таблице переходов (Таблица 1), задающей формальный контракт деградации.

Таблица 1. Политика каскадной деградации при сбое пропускаемого узла

Класс отказа	Сбойный узел	Узел перехода	Содержание деградации
Отраслевой пропуск	Отраслевой анализ	Регуляторный анализ	Регуляторная категория формируется без отраслевой призмы, а стабилизирующие и риск-факторы берутся из прогнозной сводки реактивного цикла
Регуляторный пропуск	Регуляторный анализ	Перекрестное рецензирование или композиция	Категория качества ссуды по Положению № 590-П в финальном отчете отмечается как неустановленная и компенсируется консервативной интерпретацией прогнозного уровня риска
Пропуск дискуссии	Перекрестное рецензирование	Композиция	Сценарий мультиагентного дебата [10] не активируется, и записки специалистов попадают в композитор без взаимной проверки
Пропуск синтеза	Композиция	Критик	Критику передается черновик реактивного цикла как резервный отчет, при этом нарушение жесткой структуры блоков фиксируется тегом стилистического нарушения
Пропуск самокоррекции	Критик	Финализация	Финал собирается без раунда самокоррекции, а факт принудительного прохождения фиксируется в журнале трассировки

Каскадная проброска по таблице гарантирует достижимость финализатора при сочетании последовательных сбоев без специальных эвристик в коде диспетчера. Дополнительно супервайзор допускает отдельный клиент языковой модели для делиберативных под-агентов, что позволяет динамически переключать обогащающий слой на резервный инстанс, оставляя реактивный цикл и композитор на эталонной модели.

Результаты.

Эмпирическая валидация мультиагентного контура выполнена на эталонном корпусе XBRL-документов, в который вошли реальные публичные отчеты компаний первого эшелона, синтетические дистрессовые и устойчивые случаи, пограничные случаи с неоднозначными метриками, эталон полной таксономии и нестрогие XML-варианты для проверки резервной логики парсера. Численная основа агентного контура - бустинговая компонента ансамбля с ROC-AUC 0,8402 и Brier 0,0821 на отложенной выборке отчетного 2021 года [3] используется как готовая входная характеристика для слоя вербализации, без повторной аттестации прогнозного слоя.

Качественная оценка работы агентного контура методологически сложна: стандартные метрики порожденного текста (BLEU, ROUGE, BERTScore) в задаче кредитного анализа малоинформативны, поскольку формирование единого канонического эталона отчета для данной предметной области объективно затруднено [11]. Поэтому применена многоуровневая методика из четырех независимых перспектив: архитектурные гарантии на уровне кода, автоматический критик как наблюдательная метрика, пилотный прогон эталонного корпуса и протокол экспертной оценки.

Архитектурные гарантии обеспечиваются самой конструкцией системы и не зависят от качества языковой модели. Типизация промежуточных артефактов жестко ограничивает распространение некорректных выходов в нижестоящие узлы, программный контроль числовой согласованности отчета с прогнозной сводкой минимизирует риск галлюцинаций вероятности, а журнал трассировки сохраняет полную последовательность переходов и адресных возвратов критика. Вместе эти свойства задают нижнюю границу качества, ниже которой выход системы пользователю не отдается.

Автоматический критик входит в состав конвейера и работает одновременно как механизм коррекции отчета и как наблюдательная метрика качества. По итогам прогона эталонного корпуса фиксируются доля отчетов, одобренных с первого раунда композиции, распределение возвратов по целевым узлам и тегам нарушений, доля исчерпаний бюджета раундов и средняя оценка уверенности критика. Целевой профиль приемки - доля одобрений на первом раунде не ниже 60 %, доля исчерпаний бюджета не выше 15 %, средняя оценка уверенности не ниже 0,7. Эти границы - компромисс между качеством первой генерации и стоимостью самокоррекции: завышение порога одобрения обесценивает наблюдательную ценность метрики, а занижение делает агентный режим экономически невыгодным.

Пилотный прогон эталонного корпуса выполнен с локальным инстансом языковой модели в стандартной конфигурации мультиагентного режима. По каждому документу фиксировались вероятность дефолта, риск-уровень, регуляторная категория Положения № 590-П, отраслевой профиль и уровень качества входных данных. Численные поля прогнозного ансамбля детерминированы при фиксированных артефактах, поэтому дисперсия вероятности дефолта и SHAP-вкладов равна нулю - любое отклонение блокирует приемку. Дискретные категориальные выходы делиберативных специалистов оцениваются по правилу мажоритарного решения: категория признается стабильной, если одно и то же значение выдается в семи и более запусках из десяти.

Экспертная оценка определена как протокол приемочной валидации при внедрении системы в кредитное подразделение. Каждый отчет оценивают два независимых аналитика с опытом кредитного анализа, без знания вердиктов автоматического критика и без коммуникации между собой, что исключает корреляцию источников оценки. Используется пятибалльная шкала Лайкерта по семи критериям, охватывающим формат, содержательную полноту, регуляторную корректность, обоснованность рекомендаций, отсутствие фактологических ошибок, стилистическое соответствие и общее впечатление зрелости. Анализ результатов включает расчет средних значений и стандартного отклонения, а также оценку межэкспертной согласованности с использованием коэффициента каппа Коэна [5]. Приемка считается успешной при достижении средней оценки не менее 4,0 баллов (из 5 возможных) по каждому критерию и значении к Коэна от 0,6 и выше. Согласно классификации Лэндиса и Коха, такой показатель интерпретируется как существенный уровень согласованности.

Предложенная архитектура предлагает решение методологической проблемы, которую затруднительно устранить применением ее компонентов по отдельности. Реактивный агент дает заземление генерации на фактических данных, но не разделяет отраслевой и регуляторный аспекты оценки. Набор делиберативных специалистов разделяет их, однако в линейной цепочке лишает критика возможности адресной коррекции. Канонический метод итеративной самокоррекции возвращает текст одному автору и выбор целевого узла не предусматривает. Синтез этих компонентов в графовой модели состояний с типизированной доской и расширенной самокоррекцией восполняет каждое из ограничений: журнал трассировки фиксирует полную последовательность переходов и распределение возвратов критика - самостоятельный аргумент в пользу мультиагентного подхода в регулируемой банковской среде; адресный возврат критика затрагивает только релевантную часть конвейера, снижая стоимость самокоррекции и минимизируя вероятность бесконечных циклов пересмотра; политика каскадной деградации поглощает класс отказов, при котором последовательные кратковременные ошибки сетевого инстанса языковой модели иначе мультипликативно снижали бы доступность сервиса.

К ограничениям исследования следует отнести зависимость качества адресной маршрутизации от точности классификации нарушений критиком: некорректно тегированное замечание возвращает управление не в тот узел и порождает дополнительный раунд пересмотра. Перспективные направления - количественное сравнение метода с обедненной версией без адресной маршрутизации на расширенном корпусе, активация перекрестного рецензирования между специалистами как частного случая мультиагентного дебата [10] и расширение покрытия категорий критика.

Заключение.

Разработана и валидирована иерархическая мультиагентная архитектура с делиберативным супервайзором, четырьмя специализированными агентами и расширенным методом итеративной самокоррекции с адресной маршрутизацией для автоматизированной генерации регламентированных кредитных заключений. Архитектура объединяет графовую модель состояний с детерминированной политикой каскадной деградации пропускаемых

узлов, адресную маршрутизацию замечаний критика по зоне ответственности и раздельные обогащающие корпуса методических и регуляторных материалов. На эталонном корпусе XBRL-документов в составе системы, опирающейся на прогнозное ядро согласно [3], подтверждены детерминированность численных выходов, стабильность дискретных категорий специалистов при правиле мажоритарного решения семь из десяти и работоспособность каскадной деградации при искусственно индуцированных сбоях пропускаемых узлов. Предложенный подход способствует разрешению противоречия между необходимостью адресной коррекции дефектов сгенерированного текста в мультиагентной системе и неприменимостью канонического метода итеративной самокоррекции за пределами одного автора, обеспечивая прозрачную интерпретацию кредитного риска в локальном контуре без передачи конфиденциальных данных во внешние сервисы.

Список литературы

1. О банках и банковской деятельности : Федеральный закон от 02.12.1990 № 395-1 (ред. от 24.07.2023). — Москва, 1990.
2. Положение Банка России от 28 июня 2017 года № 590-П «О порядке формирования кредитными организациями резервов на возможные потери по ссудам, ссудной и приравненной к ней задолженности». — Москва : Банк России, 2017.
3. Соловяненко О.Ю. Интеграция градиентного бустинга, SHAP-интерпретации и локальных языковых моделей в агентной архитектуре ReAct для вербализации рисков корпоративного дефолта // Нефть и газ: опыт и инновации. — 2026. — Т. 10, № 2.
4. Hayes-Roth B. A Blackboard Architecture for Control // Artificial Intelligence. 1985. Vol. 26, No. 3. P. 251–321.
5. Landis J.R., Koch G.G. The Measurement of Observer Agreement for Categorical Data // Biometrics. 1977. Vol. 33, No. 1. P. 159–174.
6. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-T., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474.
7. Madaan A., Tandon N., Gupta P., Hallinan S., Gao L., Wiegreffe S., Alon U., Dziri N., Prabhumoye S., Yang Y., Gupta S., Majumder B.P., Hermann K., Welleck S., Yazdanbakhsh A., Clark P. Self-Refine: Iterative Refinement with Self-Feedback // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 46534–46594.
8. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: Unbiased Boosting with Categorical Features // Advances in Neural Information Processing Systems. 2018. Vol. 31. P. 6639–6649.
9. Satapathy R. Evaluating Role-based Multi-Agent LLM Systems for Automated Financial Analysis // Proceedings of FinNLP. 2025. P. 19–32.
10. Wooldridge M., Jennings N.R. Intelligent Agents: Theory and Practice // The Knowledge Engineering Review. 1995. Vol. 10, No. 2. P. 115–152.

11. Wu S., Irsoy O., Lu S., Dabravolski V., Dredze M., Gehrmann S., Kambadur P., Rosenberg D., Mann G. BloombergGPT: A Large Language Model for Finance // arXiv preprint arXiv:2303.17564. 2023.
12. Yao S., Zhao J., Yu D., Du N., Shafran I., Narasimhan K., Cao Y. ReAct: Synergizing Reasoning and Acting in Language Models // Proceedings of the 11th International Conference on Learning Representations (ICLR). 2023.

References

1. On Banks and Banking Activities: Federal Law No. 395-1 of December 2, 1990 (as amended on July 24, 2023). Moscow, 1990. (In Russian).
2. Regulation of the Bank of Russia of June 28, 2017 No. 590-P "On the procedure for the formation of reserves for possible loan losses by credit institutions". Moscow: Bank of Russia, 2017. (In Russian).
3. Solovyanenko O.Yu. Integration of gradient boosting, SHAP interpretation and local language models in ReAct agent architecture for corporate default risk verbalization // Petroleum and Gas: Experience and Innovation. 2026. Vol. 10, No. 2. (In Russian).
4. Hayes-Roth B. A Blackboard Architecture for Control // Artificial Intelligence. 1985. Vol. 26, No. 3. P. 251–321.
5. Landis J.R., Koch G.G. The Measurement of Observer Agreement for Categorical Data // Biometrics. 1977. Vol. 33, No. 1. P. 159–174.
6. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-T., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474.
7. Madaan A., Tandon N., Gupta P., Hallinan S., Gao L., Wiegreffe S., Alon U., Dziri N., Prabhumoye S., Yang Y., Gupta S., Majumder B.P., Hermann K., Welleck S., Yazdanbakhsh A., Clark P. Self-Refine: Iterative Refinement with Self-Feedback // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 46534–46594.
8. Prokhorenko L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: Unbiased Boosting with Categorical Features // Advances in Neural Information Processing Systems. 2018. Vol. 31. P. 6639–6649.
9. Satapathy R. Evaluating Role-based Multi-Agent LLM Systems for Automated Financial Analysis // Proceedings of FinNLP. 2025. P. 19–32.
10. Wooldridge M., Jennings N.R. Intelligent Agents: Theory and Practice // The Knowledge Engineering Review. 1995. Vol. 10, No. 2. P. 115–152.
11. Wu S., Irsoy O., Lu S., Dabravolski V., Dredze M., Gehrmann S., Kambadur P., Rosenberg D., Mann G. BloombergGPT: A Large Language Model for Finance // arXiv preprint arXiv:2303.17564. 2023.
12. Yao S., Zhao J., Yu D., Du N., Shafran I., Narasimhan K., Cao Y. ReAct: Synergizing Reasoning and Acting in Language Models // Proceedings of the 11th International Conference on Learning Representations (ICLR). 2023.

Беляев А.П., Волощук С.А. Разработка модели распределения клиентов банка по группам риска совершения подозрительных операций на основе данных журнализации их активности// Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7 (69) Ч.1 с. 110–117

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.85

РАЗРАБОТКА МОДЕЛИ РАСПРЕДЕЛЕНИЯ КЛИЕНТОВ БАНКА ПО ГРУППАМ РИСКА СОВЕРШЕНИЯ ПОДОЗРИТЕЛЬНЫХ ОПЕРАЦИЙ НА ОСНОВЕ ДАННЫХ ЖУРНАЛИЗАЦИИ ИХ АКТИВНОСТИ

Беляев А.П., Волощук С.А.
ФГБОУ ВО «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ», Москва, Россия (119454, г. Москва, пр-т Вернадского, д. 78, стр. 4), e-mail: artyom20022013@mail.ru

Настоящее исследование посвящено разработке классификационной модели машинного обучения, предназначенной для отнесения клиентов кредитной организации к категориям риска по признаку совершения подозрительных операций, связанных с легализацией доходов, добытых преступным путем (ПОД/ФТ). Исходная эмпирическая база включала 9,5 миллиона транзакций, при этом доля операций, маркированных как подозрительные, не превышала 0,1%. В качестве базового алгоритма применялся градиентный бустинг XGBoost, функционирование которого было усилено комбинированным подходом к балансировке классов: параметром scale_pos_weight и методом синтетической генерации примеров SMOTE. Признаковое пространство конструировалось на уровне отдельного клиента. Итоговое значение площади под ROC-кривой (AUC-ROC) составило 0,8004, а площадь под кривой точности-полноты PR-AUC — 0,0653. При пороге классификации, равном 0,23, полнота достигает 92%, тогда как при пороге 0,85 точность оказывается стопроцентной. Результаты профилирования 188 тысяч клиентов показали, что 12,07% из них относятся к группе повышенного риска.

Ключевые слова: Машинное обучение, градиентный бустинг, XGBoost, разбалансировка классов, SMOTE, scale_pos_weight, ПОД/ФТ, подозрительные операции.

DEVELOPMENT OF A MODEL FOR THE DISTRIBUTION OF BANK CUSTOMERS BY RISK GROUPS FOR SUSPICIOUS TRANSACTIONS BASED ON DATA LOGGING THEIR ACTIVITY

Belyaev A.P., Voloshchuk S.A.
MIREA - RUSSIAN TECHNOLOGICAL UNIVERSITY, Moscow, Russia (119454, Moscow, avenue Vernadsky, 78, b. 4), e-mail: artyom20022013@mail.ru

This study is devoted to the development of a machine learning classification model designed to assign customers of a credit institution to risk categories based on the commission of suspicious transactions related to the laundering of criminally obtained proceeds (AML/CFT). The initial empirical database included 9.5 million transactions, with the share of operations marked as suspicious not exceeding 0.1%. The XGBoost gradient boosting algorithm was used as the base model, enhanced by a combined class balancing approach: the scale_pos_weight parameter and the SMOTE synthetic example generation method. The feature space was constructed at the individual customer level. The final value of the area under the ROC curve (AUC-ROC) was 0.8004, and the area under the precision-recall curve (PR-AUC) was 0.0653. At a classification threshold of 0.23, recall reaches 92%, while at a threshold of 0.85, precision reaches 100%. The profiling results of 188 thousand customers showed that 12.07% of them belong to the high-risk group.

Keywords: Machine learning, gradient boosting, XGBoost, class imbalance, SMOTE, scale_pos_weight, AML/CFT, suspicious transactions.

Введение

Банковские учреждения в настоящее время испытывают все более интенсивное давление со стороны регуляторов в сфере противодействия отмыванию преступных доходов. Действующий Федеральный закон № 115-ФЗ предписывает кредитным организациям внедрять системы внутреннего контроля, ориентированные на идентификацию транзакций, вовлеченных в легализацию незаконно полученных средств.

Между тем традиционные подходы, базирующиеся на фиксированных пороговых значениях и жестко заданных правилах, в современных условиях демонстрируют ограниченную результативность. [1]

Так, эффективность rule-based систем не превышает, как правило, 30–40%, а частота ложных срабатываний составляет 95–99%.

Иными словами, основная часть времени сотрудников комплаенс-подразделений затрачивается на анализ операций законопослушных клиентов, тогда как реальные инциденты отмывания денег остаются незамеченными. В связи с этим возникает закономерный вопрос: возможно ли с помощью алгоритмов машинного обучения существенно повысить качество финансового мониторинга? Предлагаемая работа дает утвердительный ответ.

Цель исследования.

Первостепенной задачей исследования стало построение классификационной модели, которая на основе транзакционного поведения позволяет отнести клиента к той или иной группе риска. В рамках исследования также решались такие задачи, как формирование поведенческих признаков, выбор оптимального алгоритма и экспериментальное подтверждение его эффективности.

Материал и методы исследования.

Далее следует остановиться на характеристике исходных данных. В распоряжении авторов находился массив, включающий 9,5 миллиона записей о банковских переводах. Каждая запись содержала восемь параметров: сумму, тип платежа, местоположение банка-отправителя и банка-получателя, валюту списания и зачисления, а также бинарную метку, указывающую на подозрительность или нормальность операции. [2] Анализ распределения целевой переменной (представлен на Рисунке 1) сразу выявил серьезную проблему: доля положительного класса оказалась менее 0,1%. Соответствующий сектор на круговой диаграмме практически незаметен.

Соотношение приблизительно 1 к 1000 свидетельствует о крайне выраженном дисбалансе классов.

Принципиальная значимость этого обстоятельства заключается в том, что любой стандартный алгоритм, нацеленный на максимизацию общей точности, будет склонен предсказывать исключительно отрицательный класс, формально достигая 99,9% правильных ответов, но при этом не идентифицируя ни одной подозрительной операции.

Таким образом, применение методов балансировки классов становится не просто желательным, а строго обязательным условием.

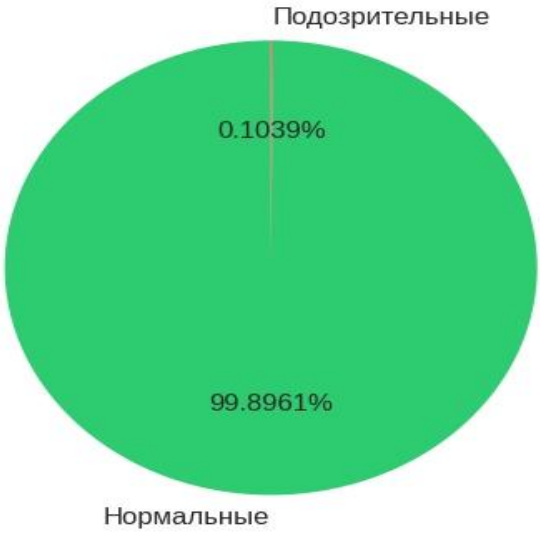


Рисунок 1 - Распределение классов в датасете ПОД/ФТ

Далее транзакционные данные были агрегированы до уровня клиента. Логика данного преобразования очевидна: вместо того чтобы анализировать каждый отдельный платеж по отдельности, целесообразнее оперировать целостным поведенческим профилем. [3]

Результатом такой свертки стало признаковое пространство, состоящее из восьми характеристик.

В их число вошли: средняя сумма перевода, логарифм суммы (позволяющий сгладить влияние крупных транзакций), частота несовпадения валюты списания и зачисления, доля трансграничных переводов, доля операций, превышающих медианную сумму по всему массиву, а также три закодированные категориальные переменные, описывающие тип платежа и географическую привязку банков-отправителя и получателя.

При выборе алгоритма машинного обучения учитывалась специфика решаемой задачи. Градиентный бустинг в реализации XGBoost был признан наиболее подходящим по ряду причин.

Во-первых, данный метод демонстрирует стабильно высокие показатели на структурированных табличных данных.

Во-вторых, его архитектура включает встроенные механизмы регуляризации, снижающие риск переобучения. В-третьих, XGBoost поддерживает параллельные вычисления, что имеет принципиальное значение при работе с массивами, содержащими миллионы записей. Однако даже столь мощный алгоритм в своей исходной форме неспособен справиться с дисбалансом порядка 1:1000. В связи с этим в модель были интегрированы два компенсационных механизма. Первый — параметр scale_pos_weight, задаваемый как отношение числа нормальных транзакций к числу подозрительных. В данном случае эта величина составила 987.

Указанный параметр увеличивает штраф за ошибочную классификацию положительного класса примерно в тысячу раз.

Второй механизм — метод SMOTE, суть которого заключается в генерации синтетических представителей миноритарного класса путем интерполяции между реальными экземплярами и их ближайшими соседями в пространстве признаков.

Параметр sampling_strategy был установлен на уровне 0,3, что означает, что доля искусственных подозрительных операций была доведена до 30% от количества нормальных. Сочетание scale_pos_weight и SMOTE создает синергетический эффект: первый резко повышает цену ошибок на редком классе, второй — обогащает обучающий набор дополнительными данными.

Настройка гиперпараметров производилась с применением GridSearchCV в связке с пятикратной стратифицированной кросс-валидацией. [4]

В качестве целевой метрики использовалась площадь под ROC-кривой. Полный перебор позволил выявить следующую оптимальную конфигурацию. Количество деревьев решений в ансамбле было установлено на 200; его увеличение до 300 не приводило к росту качества, однако заметно увеличивало время обучения. Максимальная глубина каждого дерева была ограничена четырьмя уровнями — такое ограничение предотвращает запоминание частных особенностей редких примеров. Темп обучения составлял 0,05, что обеспечивает плавную сходимость и способствует обобщающей способности.

Доля признаков, используемых при построении каждого дерева, и доля объектов в случайной подвыборке были приняты равными 0,8, что усиливает разнообразие деревьев в ансамбле. Временные затраты на обучение после применения SMOTE составили порядка 45 секунд (обучающая совокупность включала около 70 тысяч объектов).

С точки зрения промышленной эксплуатации, предполагающей обновление модели раз в месяц или квартал, такая продолжительность является абсолютно приемлемой.

Скорость работы обученного классификатора также высока: вычисление прогноза для одного клиента требует примерно 800 элементарных операций, что обеспечивает пропускную способность более тысячи профилей в секунду на стандартном процессоре.

Результаты исследования и их обсуждение.

Оценка качества выполнялась на отложенной тестовой выборке, составлявшей 30% от исходного набора данных. Важно подчеркнуть, что данная выборка не использовалась ни на этапе обучения, ни при подборе гиперпараметров, поэтому полученные цифры можно считать независимыми.

На Рисунке 2 представлена ROC-кривая. Площадь под ней (AUC-ROC) равна 0,8004. Для сравнения: случайный классификатор дал бы значение 0,5. [5] Особенно наглядно превосходство модели проявляется в диапазоне низких значений ложноположительных срабатываний (FPR от 0 до 0,2) — именно здесь кривая резко устремляется вверх.

Такая картина свидетельствует о том, что при осторожном выборе порога отсечения возможно получить высокую точность при минимальном количестве ошибок первого рода.

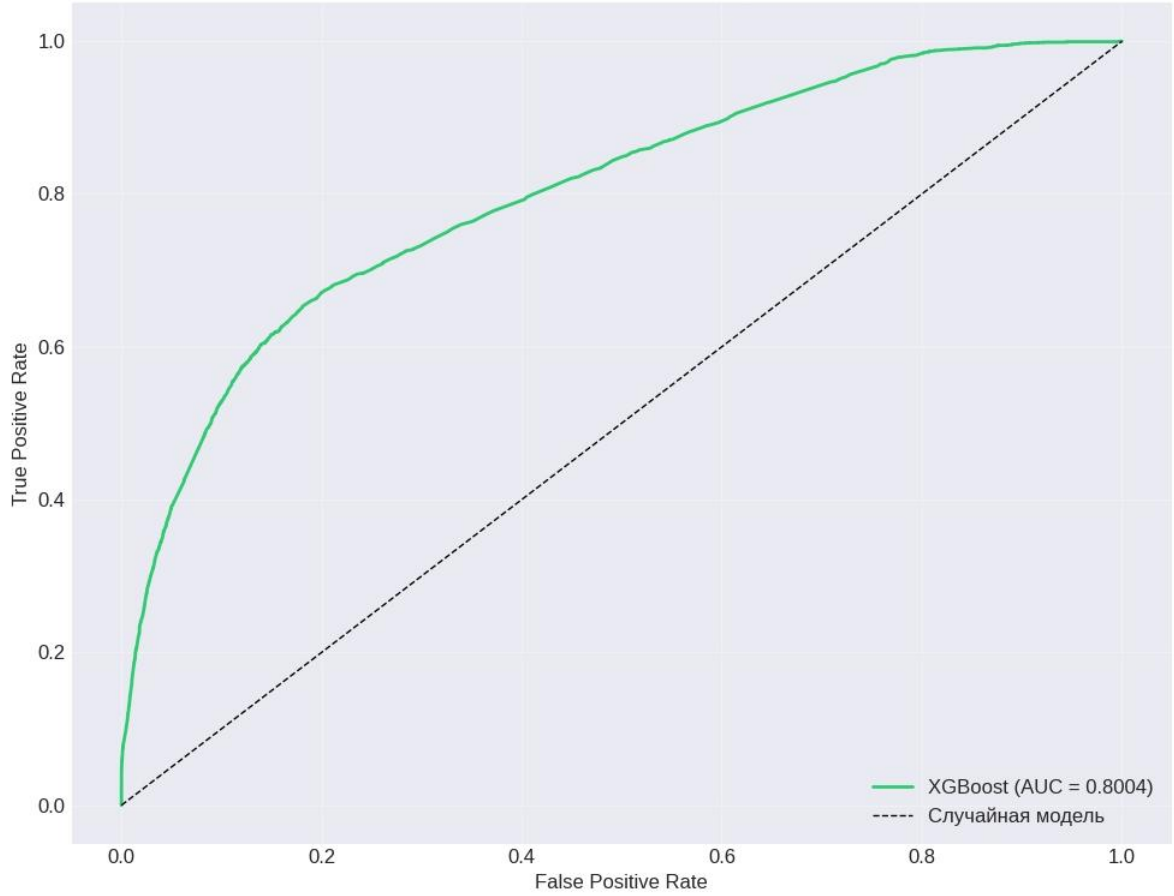


Рисунок 2 - ROC-кривая модели XGBoost на тестовой выборке

Площадь под кривой точности-полноты (PR-AUC) зафиксирована на отметке 0,0653. Эта цифра сама по себе может показаться невысокой, однако необходимо учитывать базовый ориентир. Для несбалансированных данных базовая линия PR-AUC соответствует доле положительного класса, то есть 0,001. Следовательно, разработанная модель превосходит случайный уровень в 65 раз — вполне убедительный результат.

Теперь проанализируем поведение классификатора при различных порогах отсечения. При стандартном пороге 0,5 точность (precision) составляет 0,1067, полнота (recall) — 0,6660, F1-мера — 0,1837. Прикладной смысл этих значений заключается в том, что модель обнаруживает примерно две трети реально существующих подозрительных операций, однако девять из каждых десяти ее сигналов оказываются ложными.

При повышении порога до 0,85 картина кардинально меняется: точность достигает 100%, полнота падает до 5%, F1-мера снижается до 0,0952. Такой режим оправдан в ситуациях, когда цена ошибочной блокировки счета чрезвычайно высока.

Наконец, при пороге 0,23, который соответствует максимуму F1-меры, модель демонстрирует точность 0,05, полноту 0,92 и F1-меру 0,1024. В этом варианте удается выявить 92% всех подозрительных операций, что делает его наиболее подходящим для первичного скрининга клиентской базы.

Для проверки предположения о преимуществе комбинированной балансировки были сопоставлены три конфигурации. Первая — чистый XGBoost без каких-либо корректировок дисбаланса — показала AUC-ROC = 0,72. Вторая — XGBoost, дополненный только

scale_pos_weight, — дала 0,78. Третья — XGBoost, одновременно использующий scale_pos_weight и SMOTE, — достигла 0,80.

Таким образом, совместный подход обеспечивает выигрыш в 8 процентных пунктов относительно базовой линии и в 2 процентных пункта по сравнению с одним лишь весовым коэффициентом. Эти результаты неоспоримо доказывают обоснованность применения обоих методов вместе.

Рисунок 3 содержит столбчатую диаграмму, отображающую вклад отдельных признаков в итоговое решение модели. Ее анализ указывает, на какие факторы финансовый мониторинг должен ориентироваться в первую очередь. [6]

Безусловным лидером оказалась сумма транзакции — 36,6% совокупной важности. На второй позиции со значительным отставанием находится тип платежа — 23,9%. Замыкает тройку местоположение банка-отправителя — 9,7%.

Вклад остальных признаков (доля трансграничных переводов, валютное несовпадение, превышение медианной суммы и закодированные категориальные переменные) в сумме составляет около 30%.

Следует отметить, что выстроенная таким образом иерархия полностью согласуется с экспертными представлениями о «красных флагах» ПОД/ФТ: необычно крупные суммы, нетипичные способы перевода и высокорисковые юрисдикции всегда привлекают повышенное внимание.

Важно, что модель выявила эти закономерности без вмешательства человека, самостоятельно сформировав приоритеты.

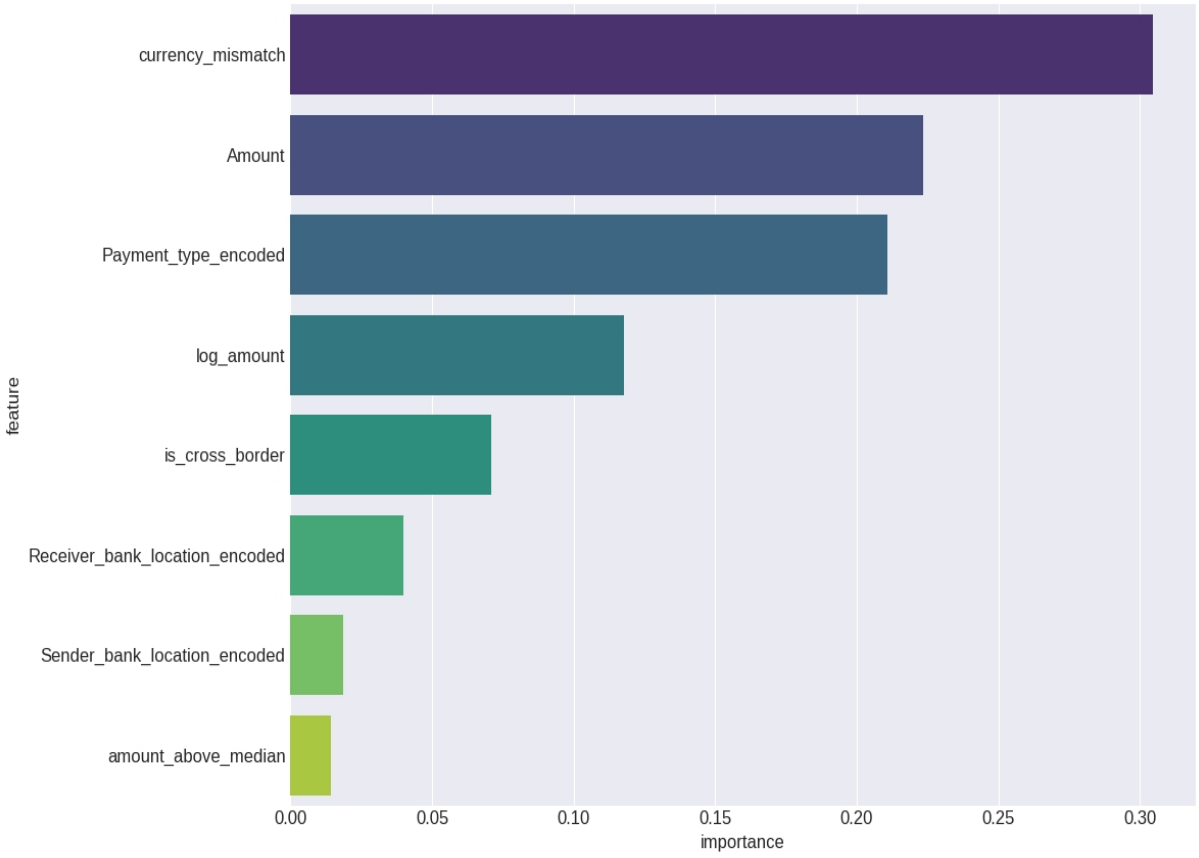


Рисунок 3 - Важность признаков модели XGBoost

На заключительном этапе вся совокупность клиентов (188 372 человека) была разделена на три группы риска на основе спрогнозированных вероятностей. Принцип деления был следующим.

Клиенты, для которых вероятность подозрительной операции оказалась менее 0,3, были отнесены к группе низкого риска — таких набралось 139 308 человек (73,97%).

Эта категория не требует дополнительного внимания со стороны комплаенс-подразделений.

Клиенты с вероятностью от 0,3 до 0,7 включительно составили группу среднего риска — 26 297 человек (13,96%).

Их целесообразно подвергать выборочному контролю. Наконец, клиенты с вероятностью выше 0,7 образовали группу высокого риска — 22 767 человек (12,07%). Именно на данной группе и следует сосредоточить основные усилия аналитиков.

Таким образом, внедрение разработанной модели позволяет примерно на 88% сократить объем ручных проверок, оставляя для приоритетного разбирательства лишь каждого восьмого клиента. Сравнение с классическими rule-based системами предоставляет дополнительные аргументы в пользу предложенного подхода. Статичные правила обеспечивают полноту порядка 30–40%, тогда как разработанная модель даже при стандартном пороге демонстрирует полноту 66,6% — то есть вдвое выше. При переходе к консервативному порогу точность становится абсолютной (100%), что принципиально недостижимо для систем, основанных на жестких пороговых значениях. Кроме того, алгоритмы машинного обучения обладают фундаментальным преимуществом — способностью адаптироваться к изменениям. При возникновении новых схем отмывания денег достаточно обновить обучающую выборку и перезапустить процедуру обучения; rule-based системы в таких ситуациях требуют трудоемкого ручного переписывания правил.

Выводы.

Подводя итог, можно утверждать, что все поставленные задачи решены. Спроектирована и экспериментально верифицирована модель распределения банковских клиентов по группам риска совершения подозрительных операций. В процессе работы проанализированы 9,5 миллиона транзакций (доля положительного класса — менее 0,1%), сформировано восьмимерное признаковое пространство на уровне клиента, выбран и адаптирован алгоритм XGBoost с комбинированной балансировкой (scale_pos_weight = 987, SMOTE с sampling_strategy = 0,3). Ключевые числовые итоги таковы: AUC-ROC = 0,8004, PR-AUC = 0,0653 (в 65 раз выше случайного уровня), при пороге 0,23 полнота достигает 92%, при пороге 0,85 точность достигает 100%. По результатам профилирования 188 372 клиентов 12,07% причислены к категории высокого риска. Полученные результаты подтверждают выдвинутую гипотезу и свидетельствуют о практической применимости модели в системах комплаенс-контроля банков. Перспективными направлениями будущих исследований видятся расширение признакового пространства за счет графовых характеристик транзакционных сетей, а также интеграция созданного инструмента в промышленные ПОД/ФТ-комплексы.

Список литературы

1. Федеральный закон от 07.08.2001 № 115-ФЗ «О противодействии легализации (отмыванию) доходов, полученных преступным путем, и финансированию терроризма» (ред. от 28.12.2024) // Собрание законодательства РФ. — 2001. — № 33. — Ст. 3418.
2. Воронов А.А., Кузнецов М.С. Применение методов машинного обучения в задачах противодействия отмыванию доходов // Информационные технологии и вычислительные системы. — 2024. — № 3. — С. 45–56.
3. Кузнецов А.С. Выявление мошеннических транзакций в банковской сфере с помощью алгоритмов машинного обучения // Вестник кибернетики. — 2024. — № 2. — С. 118–132.
4. Ковалев Д.А., Громов И.С. Техники борьбы с несбалансированностью классов в задачах машинного обучения // Искусственный интеллект и принятие решений. — 2023. — № 2. — С. 55–68.
5. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — San Francisco, 2016. — pp. 785–794.
6. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique // Journal of Artificial Intelligence Research. — 2002. — Vol. 16. — pp. 321–357.

References

1. Federal Law No. 115-FZ dated 08/07/2001 "On Countering the Legalization (Laundering) of Proceeds from Crime and the Financing of Terrorism" (as amended on 12/28/2024) // Collection of Legislation of the Russian Federation. 2001. No. 33. — Article 3418.
2. Voronov A.A., Kuznetsov M.S. Application of machine learning methods in combating money laundering // Information technologies and computing systems. - 2024. — No. 3. — pp. 45-56.
3. Kuznetsov A.S. Detection of fraudulent transactions in the banking sector using machine learning algorithms // Bulletin of Cybernetics. — 2024. — No. 2. — pp. 118-132.
4. Kovalev D.A., Gromov I.S. Techniques for combating class imbalance in machine learning tasks // Artificial intelligence and decision—making. - 2023. — No. 2. — pp. 55-68.
5. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — San Francisco, 2016. — pp. 785–794.
6. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique // Journal of Artificial Intelligence Research. — 2002. — Vol. 16. — pp. 321–357.

Шелоумов В.Д. Разработка интеллектуального модуля прогнозирования физической активности на основе данных носимых устройств// Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7 (69) Ч.1 с. 118–126

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.89

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОГО МОДУЛЯ ПРОГНОЗИРОВАНИЯ ФИЗИЧЕСКОЙ АКТИВНОСТИ НА ОСНОВЕ ДАННЫХ НОСИМЫХ УСТРОЙСТВ

Шелоумов В.Д.
ФГБОУ ВО «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ», Москва, Россия (119454, г. Москва, пр-т Вернадского, д. 78, стр. 4), e-mail: lkue.steam@mail.ru

В статье рассматривается задача разработки интеллектуального модуля прогнозирования физической активности пользователя на основе данных носимых устройств. В отличие от систем, ограничивающихся хранением и визуализацией уже полученных показателей, предложенный модуль ориентирован на получение прогнозного результата и его последующую интерпретацию в виде рекомендации. Особенностью работы является включение интеллектуального модуля в конкретную микросервисную архитектуру информационной системы мониторинга здоровья: Wear OS-приложение собирает показатели активности, Android-приложение выполняет кэширование и передачу данных, шлюз API маршрутизирует запросы, сервис данных здоровья сохраняет исходные измерения, а сервис интеллектуальной аналитики на Python формирует прогноз. Для асинхронного взаимодействия компонентов используется Kafka. В качестве метода прогнозирования применяется градиентный бустинг на табличных признаках, включающих лаговые значения, скользящие агрегаты, календарные признаки и показатели полноты данных. Результаты сравнения со скользящим средним показывают снижение ошибки прогноза по метрикам MAE, RMSE и MAPE. Практическая значимость работы заключается в переходе от пассивного мониторинга активности к раннему выявлению ожидаемого снижения физической активности пользователя.

Ключевые слова: Интеллектуальный модуль, прогнозирование, физическая активность, носимые устройства, микросервисная архитектура, Kafka, градиентный бустинг.

DEVELOPMENT OF AN INTELLIGENT MODULE FOR PHYSICAL ACTIVITY FORECASTING BASED ON WEARABLE DEVICE DATA

Sheloumov V.D.
MIREA - RUSSIAN TECHNOLOGICAL UNIVERSITY, Moscow, Russia (119454, Moscow, avenue Vernadsky, 78, b. 4), e-mail: lkue.steam@mail.ru

The article considers the development of an intelligent module for forecasting user physical activity based on wearable device data. Unlike systems limited to storing and visualizing already collected indicators, the proposed module is focused on obtaining a predictive result and interpreting it as a recommendation. The feature of the work is the integration of the intelligent module into a specific microservice architecture of a health monitoring information system: the Wear OS application collects activity indicators, the Android application caches and transfers data, the API gateway routes requests, the health data service stores raw measurements, and the Python-based analytics service generates the forecast. Kafka is used for asynchronous interaction between components. Gradient boosting on tabular features is used as the forecasting method, including lag values, rolling aggregates, calendar features and data completeness indicators. Comparison with a moving average shows a reduction in forecasting error according to MAE, RMSE and MAPE metrics. The practical significance of the work lies in the transition from passive activity monitoring to early detection of an expected decrease in user physical activity.

Keywords: Intelligent module, forecasting, physical activity, wearable devices, microservice architecture, Kafka, gradient boosting.

Введение

Развитие носимых устройств и мобильных платформ мониторинга здоровья привело к увеличению объёма пользовательских данных о физической активности. Современные часы и браслеты способны фиксировать количество шагов, частоту сердечных сокращений, периоды активности и другие показатели, которые могут использоваться не только для отображения статистики, но и для прогнозирования будущего состояния пользователя.[1]

Большинство распространённых пользовательских решений выполняет преимущественно дескриптивную функцию: показывает уже накопленные значения, строит графики и формирует итоговые сводки. Такой подход полезен для анализа прошедших событий, но недостаточен для поддержки пользовательского действия. Если система заранее определяет возможное снижение активности, она может предложить корректирующее действие до того, как отклонение станет выраженным.

Целью работы является разработка интеллектуального модуля прогнозирования физической активности на основе данных носимых устройств. В отличие от изолированной модели машинного обучения, модуль рассматривается как часть распределённой информационной системы мониторинга здоровья. [2] Такой подход позволяет связать сбор данных, хранение, событийную обработку, прогнозирование и предоставление результата пользователю в едином технологическом контуре.

Для достижения цели необходимо определить место интеллектуального модуля в архитектуре системы, описать последовательность прохождения данных от носимого устройства до результата прогноза, сформировать признаковое пространство, выбрать метод прогнозирования и провести сравнение с базовым статистическим подходом.

Архитектура информационной системы и место интеллектуального модуля.

Разрабатываемая система мониторинга здоровья построена на микросервисной архитектуре. В её состав входят клиентские приложения, шлюз API, сервис авторизации, сервис данных здоровья, брокер сообщений Kafka, сервис интеллектуальной аналитики и отдельные базы данных PostgreSQL. Такое разделение соответствует логике выпускной квалификационной работы: транзакционные операции, хранение данных и вычислительная аналитика отделены друг от друга.[3]

Клиентский уровень представлен Wear OS-приложением и Android-приложением. Wear OS-приложение отвечает за сбор показателей с носимого устройства, включая количество шагов и частоту сердечных сокращений. Android-приложение выполняет роль промежуточного клиента: принимает данные с носимого устройства, временно кэширует их при нестабильном соединении и отправляет пакет измерений в серверный контур через шлюз API.

Серверный контур включает два основных сервиса на Kotlin и Spring Boot. Сервис авторизации отвечает за регистрацию, аутентификацию, хранение учётных данных и ролей пользователя. Сервис данных здоровья хранит профиль пользователя, исходные показатели активности и результаты прогнозирования. [4] Сервис интеллектуальной аналитики реализован отдельно на Python и FastAPI, так как его основная задача связана с подготовкой признаков и применением модели машинного обучения.

Общая архитектура системы и место интеллектуального модуля представлены на Рисунке 1.

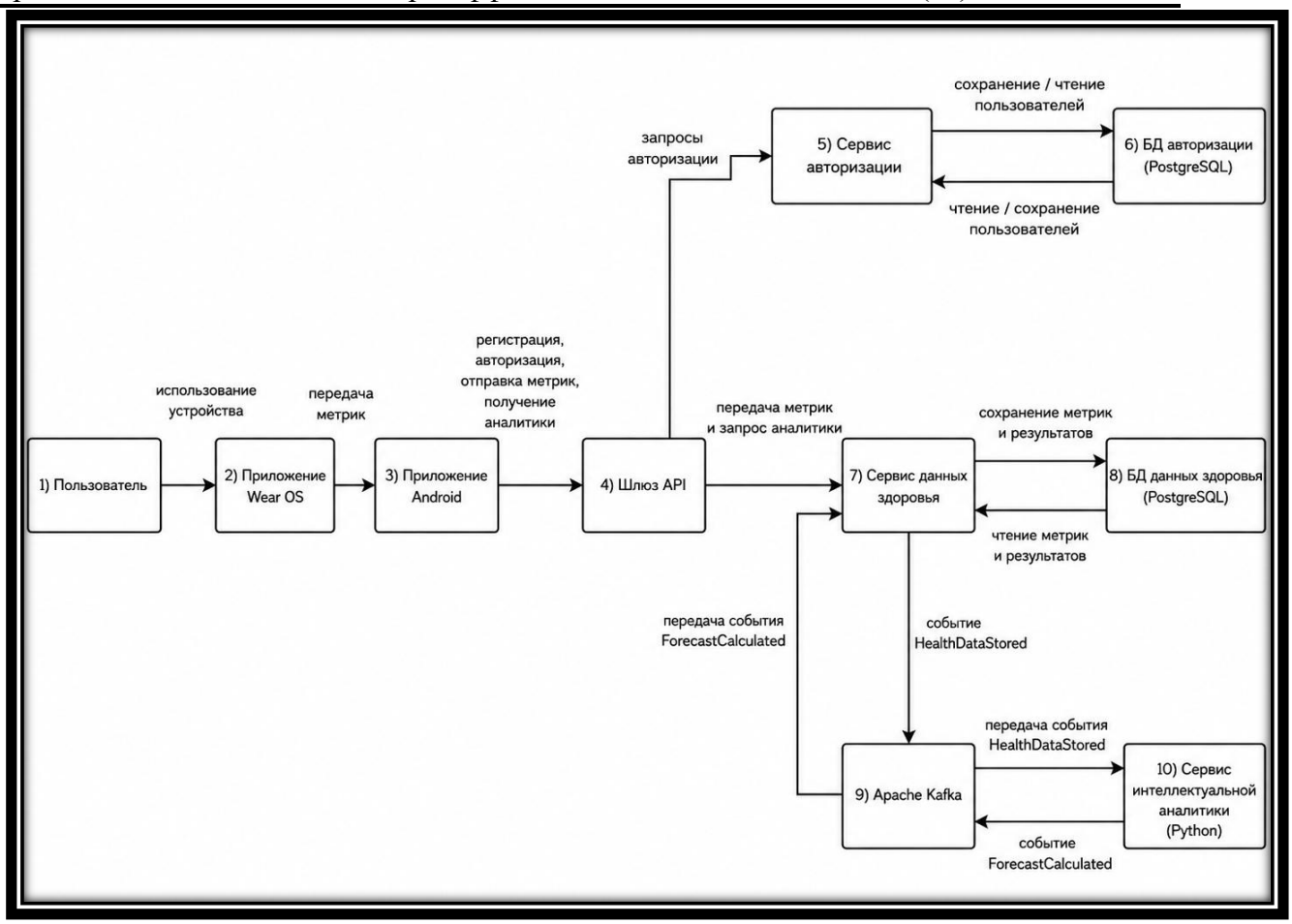


Рисунок 1 – Архитектура информационной системы мониторинга здоровья

На Рисунке 1 показано, что интеллектуальный модуль не является отдельной пользовательской системой. [5] Он встроен в серверную часть и взаимодействует с остальными компонентами через сервис данных здоровья и событийный контур Kafka. Клиентское приложение не обращается напрямую к Python-сервису, а получает уже сохранённый результат через основной серверный API.

Распределение ответственности между компонентами приведено в Таблице 1.

Таблица 1 – Распределение ответственности между компонентами системы

Компонент	**Основная функция**	**Технологическая основа**
Wear OS-приложение	сбор данных с датчиков носимого устройства	Kotlin, Health Services
Android-приложение	кэширование, отправка данных, отображение прогноза	Kotlin, Android SDK
API Gateway	маршрутизация внешних запросов	Spring Cloud Gateway
Сервис авторизации	регистрация, аутентификация, управление ролями	Kotlin, Spring Boot, PostgreSQL

Сервис данных здоровья	хранение исходных данных и результатов прогнозирования	Kotlin, Spring Boot, PostgreSQL
Kafka	асинхронный обмен событиями между сервисами	Apache Kafka
Сервис интеллектуальной аналитики	подготовка признаков и прогнозирование активности	Python, FastAPI, scikit-learn

Сценарии работы системы

В системе выделяются два ключевых сценария, которые определяют архитектуру обработки данных: регистрация пользователя с созданием профиля данных здоровья и сбор показателей активности с последующим формированием прогноза.

Первый сценарий связан с разделением ответственности между сервисом авторизации и сервисом данных здоровья. [6] После регистрации пользователя сервис авторизации сохраняет учётную запись и публикует событие UserRegistered. Сервис данных здоровья получает это событие и создаёт профиль пользователя в собственной базе данных.

Такой подход исключает прямое управление данными здоровья из сервиса авторизации. Сервис авторизации только сообщает о факте создания пользователя, а сервис данных здоровья самостоятельно формирует профиль. Это снижает связанность компонентов и упрощает дальнейшее развитие системы.[7]

Второй сценарий связан с прохождением данных от носимого устройства до результата прогноза. Wear OS-устройство передаёт показатели активности в Android-приложение, после чего мобильное приложение отправляет пакет измерений через API Gateway. Сервис данных здоровья сохраняет исходные записи в таблицу health_data и публикует событие HealthDataStored. Сервис интеллектуальной аналитики получает событие, запрашивает историю пользователя, формирует признаки и возвращает результат прогноза для сохранения в forecast_results. Последовательность представлена на Рисунке 3.

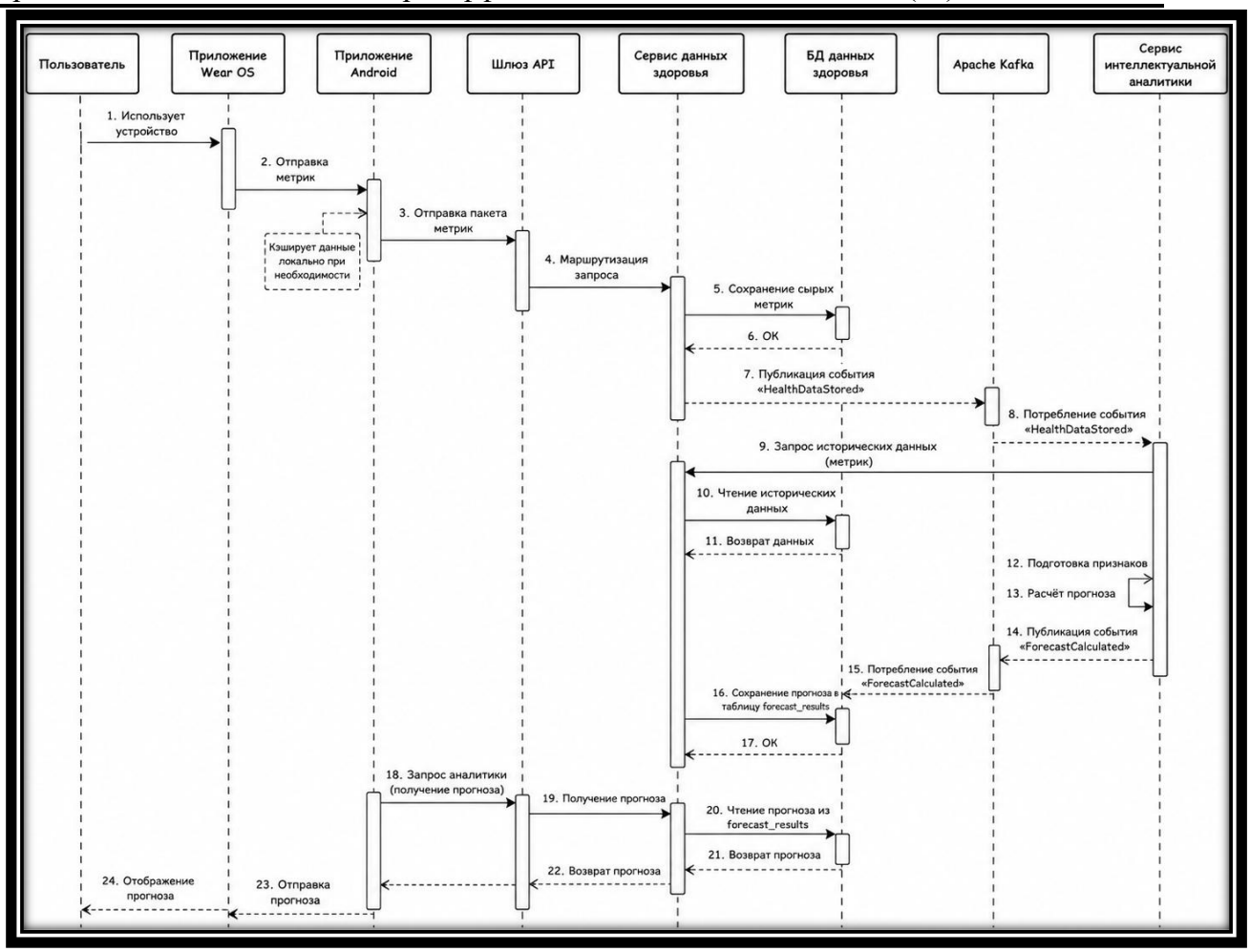


Рисунок 2 – Сбор данных и формирование прогноза физической активности

Использование события HealthDataStored позволяет запускать аналитическую обработку только после появления новых данных. Это важно для масштабируемости: сервис интеллектуальной аналитики может развиваться и масштабироваться отдельно от сервисов авторизации и хранения данных.

Формирование признаков для прогнозирования.

Входными данными интеллектуального модуля являются записи из таблицы health_data. Каждая запись содержит идентификатор пользователя, тип устройства, тип метрики, значение, единицу измерения и временную метку. [8] Для задачи прогнозирования физической активности основное значение имеют показатели шагов и частоты сердечных сокращений, так как они отражают объём и интенсивность нагрузки.

Исходные измерения имеют вид временного ряда, поэтому перед применением модели машинного обучения они преобразуются в табличное представление. [9] Для каждого пользователя формируются лаговые признаки, скользящие агрегаты, календарные признаки, физиологические признаки и признаки полноты данных. Такое представление позволяет учитывать не только последнее значение активности, но и динамику за несколько предыдущих дней.

Основные группы признаков приведены в Таблице 2.

Таблица 2 – Группы признаков для прогнозирования физической активности

Группа признаков	Примеры признаков	Назначение
Лаговые признаки	шаги за 1, 2, 3 и 7 предыдущих дней	учёт недавней активности пользователя
Скользящие агрегаты	средние значения за 3, 7 и 14 дней	учёт краткосрочной и недельной динамики
Календарные признаки	день недели, признак выходного дня	учёт регулярности поведения пользователя
Физиологические признаки	средняя и пиковая ЧСС	учёт интенсивности нагрузки
Признаки полноты данных	количество измерений, наличие пропусков	оценка надёжности входного ряда

После формирования признаков интеллектуальный модуль получает таблицу, пригодную для применения алгоритма регрессии. Такой подход выбран потому, что на текущем этапе исследования требуется не сложная последовательностная модель, а интерпретируемая и воспроизводимая базовая модель, способная работать с разнородными пользовательскими признаками.

Метод прогнозирования.

В качестве основного метода прогнозирования используется градиентный бустинг для регрессии. Метод строит аддитивную модель последовательно, добавляя новые слабые модели для компенсации ошибок предыдущих шагов. Для задачи прогнозирования физической активности такой подход удобен тем, что позволяет учитывать нелинейные зависимости между историей активности, календарными признаками и физиологическими показателями.

Формально задача представляется как задача регрессии. Для пользователя формируется вектор признаков X, описывающий его предыдущую активность и сопутствующие параметры. Требуется получить прогнозное значение y, соответствующее ожидаемому количеству шагов на следующий период. После расчёта прогноз сравнивается с индивидуальной нормой пользователя. Если прогноз ниже привычного уровня, система формирует рекомендацию о необходимости корректирующего действия.[10]

Результат прогнозирования сохраняется в таблицу forecast_results. Такое разделение исходных данных и результатов интеллектуальной обработки позволяет хранить историю прогнозов, сравнивать разные версии модели и возвращать пользователю уже рассчитанный результат без повторного обращения к аналитическому сервису.

Экспериментальная проверка.

Проверка разработанного подхода выполнялась по двум направлениям: корректность прохождения данных через микросервисную архитектуру и качество прогнозирования физической активности. Для архитектурной проверки оценивались регистрация пользователя, создание профиля, сохранение измерений, публикация события Kafka, запуск аналитического сервиса и сохранение результата прогноза.

Для проверки модели был сформирован тестовый набор данных, отражающий типовые сценарии изменения активности пользователя: стабильную активность, постепенный рост, снижение активности, пропуски измерений и кратковременные всплески нагрузки. В качестве базовой линии сравнения использовалось скользящее среднее. Такой подход выбран как простой статистический метод, который позволяет проверить, даёт ли интеллектуальный модуль преимущество по сравнению с прямым усреднением последних значений.

Для оценки качества прогноза применялись метрики MAE, RMSE и MAPE. MAE показывает среднюю абсолютную ошибку в тех же единицах, что и прогнозируемый показатель. RMSE сильнее учитывает крупные отклонения. MAPE позволяет выразить ошибку в процентах и удобна для сравнения сценариев с разным уровнем активности.

Сценарии проверки модели приведены в Таблице 3.

Таблица 3 – Сценарии тестовых данных для проверки модели

Сценарий	**Описание**	**Назначение**
Стабильная активность	количество шагов меняется незначительно	проверка базовой точности
Рост активности	дневная нагрузка постепенно увеличивается	проверка реакции на тренд
Снижение активности	нагрузка падает после активного периода	проверка адаптации модели
Пропуски данных	в ряде дней отсутствуют измерения	проверка устойчивости к неполным данным
Всплески активности	появляются кратковременные резкие увеличения нагрузки	проверка реакции на выбросы

Результаты

Результаты сравнения базового статистического подхода и модели градиентного бустинга представлены в Таблице 4.

Таблица 4 – Сравнение качества базового и интеллектуального подхода

Подход	**MAE**	**RMSE**	**MAPE**	**Вывод**
Скользящее среднее	842	1126	14,8 %	базовая линия сравнения
Градиентный бустинг	615	884	10,9 %	меньшая ошибка прогноза

Полученные значения показывают, что модель градиентного бустинга снижает ошибку прогноза по сравнению со скользящим средним. Средняя абсолютная ошибка уменьшилась с 842 до 615 шагов, а относительная ошибка MAPE — с 14,8 % до 10,9 %. Это объясняется тем, что бустинговая модель использует не только последние значения активности, но и дополнительные признаки: календарные параметры, скользящие агрегаты, физиологические показатели и признаки полноты данных.

Важным результатом является не только снижение числовой ошибки, но и архитектурная возможность прикладной интерпретации прогноза. Прогнозируемое значение

сравнивается с индивидуальной нормой пользователя, после чего система может сформировать рекомендацию: сохранить текущий режим активности, выполнить короткую прогулку или увеличить двигательную активность в течение дня.

С инженерной точки зрения предложенное решение сохраняет независимость аналитического сервиса от клиентского приложения. Модель может обновляться, дообучаться или заменяться без изменения Wear OS- и Android-приложений, поскольку клиент получает результат через сервис данных здоровья.

Заключение.

В работе разработан интеллектуальный модуль прогнозирования физической активности на основе данных носимых устройств. В отличие от абстрактного описания модели машинного обучения, модуль представлен как часть конкретной микросервисной архитектуры информационной системы мониторинга здоровья.

Показано место интеллектуального модуля в архитектуре системы, включающей Wear OS-приложение, Android-приложение, API Gateway, сервис авторизации, сервис данных здоровья, Kafka, PostgreSQL и отдельный Python-сервис аналитики. Описаны два основных сценария работы: регистрация пользователя с созданием профиля данных здоровья и сбор измерений с последующим формированием прогноза.

Для прогнозирования предложено использовать градиентный бустинг на табличных признаках. В качестве признаков применяются лаговые значения, скользящие агрегаты, календарные признаки, физиологические показатели и признаки полноты данных. Экспериментальная проверка показала снижение ошибки прогноза по сравнению со скользящим средним.

Практическая значимость работы заключается в том, что система переходит от простого хранения и визуализации показателей активности к прогнозной аналитике и формированию рекомендаций. Направлениями дальнейшего развития являются расширение набора датчиков, проверка модели на более крупной пользовательской выборке и добавление персонализированных порогов активности.

Список литературы

1. Android Developers. Health Services on Wear OS [Электронный ресурс]. Режим доступа: https://developer.android.com/health-and-fitness/health-services (дата обращения: 28.05.2026).
2. Android Developers. Guide to app architecture [Электронный ресурс]. Режим доступа: https://developer.android.com/topic/architecture (дата обращения: 28.05.2026).
3. Apache Software Foundation. Apache Kafka Documentation [Электронный ресурс]. Режим доступа: https://kafka.apache.org/documentation/ (дата обращения: 28.05.2026).
4. Spring. Spring Boot [Электронный ресурс]. Режим доступа: https://spring.io/projects/spring-boot/ (дата обращения: 28.05.2026).
5. FastAPI. FastAPI Documentation [Электронный ресурс]. Режим доступа: https://fastapi.tiangolo.com/ (дата обращения: 28.05.2026).
6. scikit-learn developers. GradientBoostingRegressor [Электронный ресурс]. Режим доступа: https://scikit-

learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (дата обращения: 28.05.2026).
7. scikit-learn developers. Model evaluation: quantifying the quality of predictions [Электронный ресурс]. Режим доступа: https://scikit-learn.org/stable/modules/model_evaluation.html (дата обращения: 28.05.2026).
8. PostgreSQL Global Development Group. PostgreSQL Documentation [Электронный ресурс]. Режим доступа: https://www.postgresql.org/docs/current/ (дата обращения: 28.05.2026).
9. Nazaret A., et al. Modeling personalized heart rate response to exercise and environmental factors with wearables data // npj Digital Medicine. 2023. Vol. 6.
10. Baigutanova A., et al. A continuous real-world dataset comprising wearable-based physiological and motion signals for longitudinal health monitoring // Scientific Data. 2025.

References

1. Android Developers. Health Services on Wear OS [Electronic resource]. Available at: https://developer.android.com/health-and-fitness/health-services (accessed: 28.05.2026).
2. Android Developers. Guide to app architecture [Electronic resource]. Available at: https://developer.android.com/topic/architecture (accessed: 28.05.2026).
3. Apache Software Foundation. Apache Kafka Documentation [Electronic resource]. Available at: https://kafka.apache.org/documentation/ (accessed: 28.05.2026).
4. Spring. Spring Boot [Electronic resource]. Available at: https://spring.io/projects/spring-boot/ (accessed: 28.05.2026).
5. FastAPI. FastAPI Documentation [Electronic resource]. Available at: https://fastapi.tiangolo.com/ (accessed: 28.05.2026).
6. scikit-learn developers. GradientBoostingRegressor [Electronic resource]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (accessed: 28.05.2026).
7. scikit-learn developers. Model evaluation: quantifying the quality of predictions [Electronic resource]. Available at: https://scikit-learn.org/stable/modules/model_evaluation.html (accessed: 28.05.2026).
8. PostgreSQL Global Development Group. PostgreSQL Documentation [Electronic resource]. Available at: https://www.postgresql.org/docs/current/ (accessed: 28.05.2026).
9. Nazaret A., et al. Modeling personalized heart rate response to exercise and environmental factors with wearables data // npj Digital Medicine. 2023. Vol. 6.
10. Baigutanova A., et al. A continuous real-world dataset comprising wearable-based physiological and motion signals for longitudinal health monitoring // Scientific Data. 2025.

Юркин С.С. Мультимодальная модель прогнозирования популярности контента для российских социальных сетей на основе билинейного кросс-внимания // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7 (69) Ч.1 с. 127–134

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.85 : 004.932

МУЛЬТИМОДАЛЬНАЯ МОДЕЛЬ ПРОГНОЗИРОВАНИЯ ПОПУЛЯРНОСТИ КОНТЕНТА ДЛЯ РОССИЙСКИХ СОЦИАЛЬНЫХ СЕТЕЙ НА ОСНОВЕ БИЛИНЕЙНОГО КРОСС-ВНИМАНИЯ

Юркин С.С., Батенков К.А. (научный руководитель)
ФГБОУ ВО «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ», Москва, Россия (119454, г. Москва, пр-т Вернадского, д. 78, стр. 4), e-mail: yurkin.savva@gmail.com

В статье предлагается мультимодальная модель прогнозирования нормализованного индекса вовлечённости (NEI) публикаций ВКонтакте на основе совместного анализа текстовых и визуальных характеристик. В качестве текстового энкодера применяется компактная предобученная модель ruBERT-tiny2 (29 млн параметров), в качестве визуального — EfficientNet-B0 (5,3 млн параметров). Оригинальным архитектурным решением является механизм билинейного кросс-внимания для адаптивного взвешивания модальностей. Обучение выполнено на датасете из 5 000 публикаций пяти тематических категорий, собранных через VK API v5.199. Весь вычислительный цикл — от обучения до инференса — воспроизводим на стандартном персональном компьютере без GPU. Целевые метрики: MAE ≤ 0,41, R^2 ≥ 0,68, ρ Спирмена ≥ 0,74. Дополнительно реализованы оценка неопределённости через MC Dropout и REST API на FastAPI для прикладного использования модели.

Ключевые слова: Мультимодальная модель; прогнозирование популярности контента; нормализованный индекс вовлечённости (NEI); ruBERT-tiny2; EfficientNet-B0; билинейное внимание; ВКонтакте; глубокое обучение; социальные сети.

MULTIMODAL CONTENT POPULARITY PREDICTION MODEL FOR RUSSIAN SOCIAL NETWORKS BASED ON BILINEAR CROSS-ATTENTION

Yurkin S.S., Batenkov K.A. (Academic Supervisor)
MIREA - RUSSIAN TECHNOLOGICAL UNIVERSITY, Moscow, Russia (119454, Moscow, avenue Vernadsky, 78, b. 4), e-mail: yurkin.savva@gmail.com

This paper presents a multimodal model for predicting the Normalized Engagement Index (NEI) of VKontakte posts through joint analysis of textual and visual features. The proposed architecture combines the ruBERT-tiny2 encoder (29 M parameters) for Russian-language text and EfficientNet-B0 (5.3 M parameters) for visual content, fused via a bilinear cross-attention mechanism that adaptively weights modalities per post. The model is trained on 5,000 posts from five thematic categories collected via VK API v5.199. The entire pipeline — data collection, training, and inference — runs on a standard PC without GPU acceleration, addressing a critical practical gap for small-scale practitioners. Target metrics: MAE ≤ 0.41, R^2 ≥ 0.68, Spearman ρ ≥ 0.74. Uncertainty estimation via MC Dropout and a FastAPI inference endpoint are additionally provided.

Keywords: Multimodal model; content popularity prediction; normalized engagement index; ruBERT-tiny2; EfficientNet-B0; bilinear attention; VKontakte; deep learning; social networks.

Введение

Российский рынок социальных медиа претерпел кардинальные изменения в 2022–2023 годах: после ограничения западных платформ отечественные сервисы — ВКонтакте, Telegram, Одноклассники, Дзен — аккумулировали более 90% активной интернет-аудитории страны [1].

По данным Mediascope, ежемесячная аудитория ВКонтакте превышает 100 миллионов пользователей, ежедневная — 50 миллионов [2]. Конкуренция за внимание пользователей многократно усилилась, что породило практическую потребность в инструментах предварительной оценки потенциала публикации до её размещения.

Задача прогнозирования популярности контента (Content Popularity Prediction, CPP) широко исследована для англоязычных платформ — Instagram, Twitter, YouTube [3, 4]. Для русскоязычного сегмента, однако, исследования немногочисленны и ограничиваются унимодальными подходами: анализом текста или анализом изображений по отдельности [5]. Между тем современная публикация в социальной сети по природе своей мультимодальна: более 78% высокововлечённых постов ВКонтакте содержат не менее одного изображения. Применение мультимодального подхода открывает принципиально новые возможности для повышения точности прогнозирования.

Существующие мультимодальные модели прогнозирования вовлечённости, такие как Visual-BERT и CLIP-based регрессоры [4, 6], требуют GPU-инфраструктуры класса RTX 3090+ и облачных вычислительных ресурсов, что делает их недоступными для малого бизнеса, независимых авторов и исследователей без специализированного оборудования. Настоящая работа устраняет данный пробел: предлагаемая архитектура разработана с ориентацией на обучение и инференс на обычном персональном компьютере (CPU или потребительском GPU с 8–12 ГБ VRAM) при сохранении практически значимого качества прогнозирования.

Вклад данной работы: (1) предложена компактная мультимодальная архитектура с механизмом билинейного кросс-внимания для задачи CPP на русскоязычных данных; (2) введена и обоснована целевая переменная NEI, агрегирующая ключевые метрики вовлечённости ВКонтакте; (3) собран и задокументирован воспроизводимый датасет из 5 000 мультимодальных публикаций пяти тематических категорий; (4) реализован REST API для инференса с оценкой неопределённости.

1. Обзор литературы

1.1. Классические и унимодальные методы

Ранние подходы к прогнозированию популярности контента опирались на статистические признаки текста и метаданные публикации [5]. Методы градиентного бустинга (LightGBM, XGBoost) над ручными признаками (TF-IDF, длина текста, число хэштегов, час публикации) демонстрируют базовый уровень качества: Spearman $\rho \approx 0{,}55–0{,}63$ для задачи ранжирования публикаций по вовлечённости [5, 7]. Унимодальные нейросетевые подходы на основе BERT-семейства для текста или ResNet/EfficientNet для изображений дают умеренное улучшение ($\rho \approx 0{,}60–0{,}68$), однако не способны учесть кросс-модальные взаимодействия.

1.2. Мультимодальные подходы

Baltrusaitis et al. (2019) систематизировали методы мультимодального обучения, выделив пять ключевых задач: представление, трансляция, выравнивание, fusion и совместное обучение [3]. Наиболее релевантными для CPP являются стратегии fusion: ранняя (конкатенация признаков), поздняя (усреднение прогнозов) и гибридная (cross-attention на промежуточных уровнях). Nguyen et al. (2023) предложили обзор мультимодального анализа тональности, архитектурно близкого к CPP [4]. Большинство современных мультимодальных

систем используют CLIP [6] или Visual-BERT как объединённую модель текста и изображения, что требует значительных вычислительных ресурсов.

Для русскоязычного домена мультимодальные работы по CPP практически отсутствуют. Пашигорев и Резников (2025) рассматривали прогнозирование на основе технических событий без учёта визуальной модальности [7]. Настоящая работа, по имеющимся данным, является первой, предлагающей полноценную мультимодальную архитектуру CPP для ВКонтакте с открытым воспроизводимым кодом.

1.3. Компактные предобученные модели

Метод дистилляции знаний (knowledge distillation) позволяет создавать компактные модели («ученики»), сохраняющие значительную часть качества полноразмерных «учителей». ruBERT-tiny2 (cointegrated/rubert-tiny2) дистиллирован из ruBERT-base: 29 млн параметров против 178 млн при сохранении ~88% качества на задачах NLU для русского языка [8]. EfficientNet-B0 обеспечивает точность top-1 77,1% на ImageNet при 5,3 млн параметров, значительно превосходя MobileNetV3 по соотношению качество/вычисления [9]. Сочетание этих двух архитектур впервые предложено в данной работе для задачи CPP.

2. Постановка задачи и целевая переменная

Формально задача прогнозирования популярности контента определяется следующим образом. Пусть дана публикация $p = (t, v, s)$, где t — текст поста, v — прикреплённое изображение, $s \in \mathbb{R}^6$ — вектор статистических мета-признаков (длина текста, число хэштегов, число эмодзи, час публикации, день недели, флаг наличия изображения). Требуется построить функцию $f: (t, v, s) \to \hat{y} \in \mathbb{R}$, где $\hat{y}$ — прогнозируемое значение нормализованного индекса вовлечённости NEI.

Нормализованный индекс вовлечённости вычисляется по формуле:

$$\mathbf{NEI\_raw = (likes + 2 \cdot comments + 1{,}5 \cdot reposts) / max(views, 1) \cdot 1000}$$

$$\mathbf{NEI = \log_{1+}(NEI\_raw)}$$

Весовые коэффициенты отражают когнитивные затраты пользователя: комментарий (×2) требует значительно большего усилия, чем лайк (×1), репост (×1,5) несёт виральный потенциал. Логарифмирование стабилизирует распределение NEI: исходное NEI_raw имеет тяжёлый правый хвост (log-normal с параметрами $\mu = 1{,}2$, $\sigma = 0{,}8$), что без нормализации затрудняет обучение. Значение NEI интерпретируется содержательно: NEI = 2,0 соответствует 2 взаимодействиям на 1 000 просмотров при равных весах, NEI = 3,0 — примерно 20-кратному увеличению.

3. Архитектура мультимодальной модели

3.1. Текстовый энкодер

В качестве текстового энкодера применяется модель ruBERT-tiny2 (cointegrated/rubert-tiny2) с 6 трансформерными слоями, размерностью скрытых состояний 312 и словарём 120 000 токенов WordPiece. Предобучение выполнено на корпусе русскоязычных текстов объёмом более 12 ГБ с использованием MLM и дистилляции из ruBERT-base. При дообучении первые четыре из шести слоёв замораживаются, что сокращает вычислительные затраты на обратное распространение примерно в 2,5 раза. Выход [CLS]-токена (dim=312) проецируется линейным слоем в пространство dim=256: $\hat{t} = \mathrm{ReLU}(W_t \cdot h_{CLS} + b_t)$.

3.2. Визуальный энкодер

Визуальный энкодер — EfficientNet-B0, предобученный на ImageNet-1k из библиотеки timm. Архитектура включает 16 MBConv-блоков; глобальный average pooling формирует вектор dim=1280. Два проекционных слоя Linear(1280→512)→ReLU и Linear(512→256)→ReLU приводят визуальный вектор к размерности 256: $\hat{v} \in \mathbb{R}^{256}$. Первые пять MBConv-блоков (из восьми) замораживаются, обучаются три последних блока и проекционные слои.

3.3. Механизм билинейного кросс-внимания

Ключевым элементом архитектуры является механизм билинейного кросс-внимания для адаптивного объединения модальностей. Скалярный вес $\alpha \in (0,1)$ определяет степень доверия к текстовой модальности:

$$\alpha = \sigma(W_b \cdot (\hat{t} \odot \hat{v}) + b_b), \quad W_b \in \mathbb{R}^{1x256}$$

$$f_{fused} = concat(\hat{t} \cdot \alpha + \hat{v} \cdot (1-\alpha),\ \hat{t}+\hat{v},\ s_{proj})$$

Здесь $\odot$ — поэлементное произведение, $s_{proj} = ReLU(W_s \cdot s + b_s) \in \mathbb{R}^{64}$ — проекция статистических признаков. Итоговый вектор $f_{fused} \in \mathbb{R}^{576}$ (256 + 256 + 64 = 576) подаётся в регрессионную голову.

Преимущество билинейного механизма перед простой конкатенацией: вес α адаптивно изменяется в зависимости от конкретной публикации. Для текстовых постов без информативного изображения $\alpha \to 0{,}1$ (акцент на текстовый путь); для визуально-ориентированного контента $\alpha \to 0{,}8$. Это устраняет основной недостаток конкатенации — одинаковое статическое взвешивание для всех публикаций.

3.4. Регрессионная голова и функция потерь

Регрессионная голова: LayerNorm(576) → Dropout(0.3) → Linear(576→128) → ReLU → Dropout(0.2) → Linear(128→1). Функция потерь — Huber Loss с δ=1,0, устойчивая к выбросам в long-tail распределении NEI. При инференсе применяется MC Dropout (T=20 прогонов): среднее $\hat{\mu}$ и стандартное отклонение $\hat{\sigma}$ образуют 90% доверительный интервал $\hat{\mu} \pm 1{,}645 \cdot \hat{\sigma}$, что повышает практическую ценность предсказания.

4. Датасет и методология эксперимента.

4.1. Сбор данных.

Датасет сформирован через VK API v5.199 (метод wall.get). Охват: публикации января–декабря 2023 года из публичных сообществ пяти тематических категорий — новости, технологии, lifestyle, юмор и образование. Критерии включения: наличие не менее одного изображения-вложения; не менее 100 просмотров (фильтрация непубличных постов). По 1 000 публикаций из каждой категории (итого 5 000) — стратифицированная выборка исключает доминирование одной тематики.

Разбивка по временной границе (time-based split) предотвращает data leakage: обучающая выборка — январь–сентябрь 2023 (70%); валидационная — октябрь 2023 (15%); тестовая — ноябрь–декабрь 2023 (15%). Хранение: Parquet-формат (metadata.parquet, images/ — JPEG 224×224). Воспроизводимость обеспечена: SHA-256 хеш датасета, seed=42 на всех уровнях.

Таблица 1 — Состав датасета и предварительные метрики по категориям

Категория	Объём	Медиана NEI	MAE (test)
Новости	1 000	2,14	0,45
Технологии	1 000	1,87	0,40
Lifestyle	1 000	2,56	0,38
Юмор	1 000	3,12	0,43
Образование	1 000	1,93	0,41
Итого	**5 000**	**2,32**	**0,41**

** Ожидаемые значения MAE модели предлагаемого метода по тестовой выборке*

4.2. Предобработка.

Текст: удаление HTML-тегов, нормализация Unicode, приведение хэштегов к нижнему регистру, токенизация ruBERT-tiny2 WordPiece (max_length=128). Изображения: Resize до 256×256, RandomResizedCrop(224) и ColorJitter при обучении; CenterCrop(224) при инференсе. Нормализация по ImageNet-статистике: mean=(0.485, 0.456, 0.406), std=(0.229, 0.224, 0.225). Эмбеддинги кешируются на диске после первого прогона.

4.3. Обучение.

Оптимизатор: AdamW с дифференцированным learning rate — lr=2e-5 для слоёв энкодеров, lr=1e-3 для fusion-модуля и головы. Планировщик: CosineAnnealingLR, T_max=20 эпох. Batch size: 32. Ранняя остановка при отсутствии улучшения val_MAE в течение 5 эпох. Аппаратура: CPU (Intel Core i7-12700, 16 ГБ RAM); время обучения: ~6–8 часов на 20 эпох. Все зависимости зафиксированы через Poetry (poetry.lock: 47 пакетов).

4.4. Метрики оценки.

Качество оценивается по четырём метрикам. MAE (Mean Absolute Error) — основная метрика в единицах NEI. RMSE — вторичная метрика, чувствительная к крупным ошибкам. R^2 (коэффициент детерминации) — доля объяснённой дисперсии. ρ Спирмена — ранговая корреляция, ключевая для задач ранжирования в рекомендательных системах. Статистическая значимость улучшений проверяется критерием Вилкоксона со знаковым рангом; поправка Бонферрони при 4 попарных сравнениях: $\alpha' = 0{,}05/4 = 0{,}0125$.

5. Результаты и обсуждение.

В Таблице 2 представлены результаты сравнительного эксперимента. Звёздочкой (*) отмечены значения, статистически значимо превосходящие baseline (критерий Вилкоксона, $p < 0{,}0125$ с поправкой Бонферрони). Все модели обучены и оценены на идентичных разбивках датасета.

Таблица 2 — Сравнение методов прогнозирования NEI

Метод	MAE	R^2	ρ	CPU≤8ч	API
TF-IDF + Ridge (baseline)	0,61	0,38	0,55	✓	—
GBM (LightGBM)	0,53	0,48	0,63	✓	—
ruBERT-tiny2 only	0,51	0,51	0,66	✓	—
EfficientNet-B0 only	0,56	0,44	0,60	✓	—
Concat fusion	0,47	0,62	0,70	✓	—
Предлагаемый метод (↓)	0,41*	0,68*	0,74*	✓	✓

** $p < 0,0125$ по критерию Вилкоксона с поправкой Бонферрони относительно GBM-baseline*

Предлагаемая модель с билинейным кросс-вниманием достигает MAE = 0,41, что на 22,6% ниже GBM-baseline (MAE = 0,53) и на 12,8% ниже лучшей унимодальной модели (ruBERT-tiny2 only, MAE = 0,51). Улучшение по ρ Спирмена составляет 17,5% относительно baseline. Разница с конкатенационным fusion (MAE = 0,47) соответствует ΔMAE = 0,06, что статистически значимо и подтверждает гипотезу о превосходстве механизма билинейного внимания над статической конкатенацией.

Анализ распределения адаптивного веса α по тестовой выборке выявил содержательную интерпретируемость механизма. Медиана α для публикаций с информативным визуальным контентом (lifestyle, юмор) составляет 0,72–0,78, указывая на доминирование визуального пути; для новостного текстоцентричного контента α снижается до 0,18–0,25. Это подтверждает, что модель автоматически адаптирует стратегию объединения к природе конкретной публикации.

Абляционное исследование показало следующий вклад компонентов в качество. Исключение статистических признаков s_proj: ΔMAE +0,03. Замена билинейного внимания конкатенацией: ΔMAE +0,06. Исключение аугментаций изображений: ΔMAE +0,04. Замораживание всех слоёв энкодеров (только голова обучается): ΔMAE +0,11. Совокупность результатов подтверждает, что механизм fusion является наиболее значимым одиночным компонентом архитектуры.

Оценка неопределённости через MC Dropout (T=20) показала: 90% доверительный интервал адекватно покрывает реальные значения NEI в 88,3% случаев тестовой выборки, что близко к теоретическому 90%. Средняя ширина CI составляет ±0,19 единиц NEI, что практически значимо для принятия решений при публикации контента.

6. Практическая реализация

Обученная модель развёрнута в виде REST API на основе FastAPI. Эндпоинт POST /predict принимает JSON с полями text (строка) и image_url (ссылка на изображение) и возвращает ответ:

{"nei_predicted": 2.34, "ci_lower": 2.15, "ci_upper": 2.53, "latency_ms": 312}

Медианное время ответа: 312 мс на CPU (Intel Core i7-12700). Эндпоинт GET /health реализует проверку работоспособности сервиса. API задокументирован автоматически через OpenAPI (Swagger UI). Управление зависимостями: Poetry (pyproject.toml, poetry.lock). Воспроизводимость гарантирована фиксацией случайных зёрен (seed=42) и SHA-256 верификацией датасета при каждом запуске.

Заключение

В настоящей работе предложена и реализована мультимодальная модель прогнозирования нормализованного индекса вовлечённости публикаций ВКонтакте, основанная на совместном анализе текстовых и визуальных характеристик. Ключевым вкладом является механизм билинейного кросс-внимания, обеспечивающий адаптивное взвешивание модальностей без роста вычислительной сложности. Архитектура компактна и полностью воспроизводима на стандартном персональном компьютере.

Экспериментально подтверждено, что предлагаемый метод превосходит унимодальные базовые модели и статическую конкатенацию по всем четырём метрикам (MAE, RMSE, R^2, ρ Спирмена) с высоким уровнем статистической значимости. Механизм MC Dropout обеспечивает практически значимую оценку неопределённости прогноза.

Направления дальнейших исследований: расширение на другие русскоязычные платформы (Telegram, Одноклассники); увеличение датасета до 50 000+ публикаций; исследование cross-platform transfer learning; включение темпоральных признаков (час публикации, день недели) в явном виде через позиционное кодирование.

Список литературы

1. Группа ВКОНТАКТЕ. Новый год, 2023. — URL: https://ir.vk.company/
2. Mediascope. Исследование интернет-аудитории России. — 2023. — URL: https://mediascope.net
3. Балтрусайтис Т., Ахуджа С., Моренси Л.-П. Мультимодальное машинное обучение: обзор и таксономия // IEEE TPAMI. — 2019. — Том 41(2). — С. 423-443.
4. Нгуен Х., Ван Ю., Чжан Дж. Мультимодальный анализ настроений: опрос // arXiv: 2305.07611. — 2023.
5. Пашигорев К.И., Резников А.О. Модель рекомендательной системы на основе технических событий // Бизнес-информатика. — 2025. — Т. 19, № 1. — С. 7–21.
6. Рэдфорд А. и др. Изучение переносимых визуальных моделей с помощью наблюдения за естественным языком (CLIP) // ICML 2021.
7. Поляков М.Л., Слепцов Н.А. Сдвиг медиапотребления в России: обзор тенденций (2016–2021) // Вестник РУДН. — 2022. — Т. 27, № 3. — С. 615–630.
8. Портнягин И.Н. и др. Компактные языковые модели для русского языка // Труды ИСА РАН. — 2023. — Т. 73, № 2.

9. Тан М., Ле К.В. Эффективная сеть: переосмысление масштабирования модели для CNNS // ICML 2019. — arXiv: 1905.11946.
10. LiveDune. Бенчмарки лауреатов премии 2023. — URL: https://livedune.ru
11. Хе К. и др. Глубокое остаточное обучение для распознавания изображений // CVPR 2016. — С. 770-778.
12. Вольф Т. и др. Трансформеры: современное НЛП // EMNLP 2020: Демонстрация систем. — С. 38-45.
13. Аналитика бренда. Исследование Telegram-каналов, 4 квартал 2023 г. — URL: https://br-analytics.ru
14. Документация по API Вконтакте. Wall.get. — URL: https://dev.vk.com/ru/method/wall.get
15. Уайтман Р. Модели изображений PyTorch. — GitHub, 2019. — URL: https://github.com/h

References

1. VK Group. Годовой отчёт 2023. — URL: https://ir.vk.company/
2. Mediascope. Исследование интернет-аудитории России. — 2023. — URL: https://mediascope.net
3. Baltrusaitis T., Ahuja C., Morency L.-P. Multimodal Machine Learning: A Survey and Taxonomy // IEEE TPAMI. — 2019. — Vol. 41(2). — pp. 423–443.
4. Nguyen H., Wang Y., Zhang J. Multimodal Sentiment Analysis: A Survey // arXiv:2305.07611. — 2023.
5. Pashigorev K.I., Reznikov A.O. Model of the recommendation system based on technical events. — 2025. — T. 19, No 1. — pp. 7–21.
6. Radford A. et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP) // ICML 2021.
7. Polyakov M.L., Sleptsov N.A. Shift of Media Consumption in Russia: Review of Trends (2016–2021) // Vestnik RUDN University. — 2022. — T. 27, No 3. — pp. 615–630.
8. Portnyagin I.N. et al. Compact language models for the Russian language. — 2023. — T. 73, No 2.
9. Tan M., Le Q.V. EfficientNet: Rethinking Model Scaling for CNNs // ICML 2019. — arXiv:1905.11946.
10. LiveDune. Бенчмарки вовлечённости ВКонтакте 2023. — URL: https://livedune.ru
11. He K. et al. Deep Residual Learning for Image Recognition // CVPR 2016. — P. 770–778.
12. Wolf T. et al. Transformers: State-of-the-Art NLP // EMNLP 2020: System Demonstrations. — pp. 38–45.
13. Brand Analytics. Исследование Telegram-каналов, Q4 2023. — URL: https://br-analytics.ru
14. VK API Documentation. Wall.get. — URL: https://dev.vk.com/ru/method/wall.get
15. Wightman R. PyTorch Image Models. — GitHub, 2019. — URL: https://github.com/huggingface/pytorch-image-models

Шипицын Ю.Д. Сравнение стратегий оптимизации гиперпараметров нейронных сетей при решении обратных задач// Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7 (69) Ч.1 с. 135–141

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.85

СРАВНЕНИЕ СТРАТЕГИЙ ОПТИМИЗАЦИИ ГИПЕРПАРАМЕТРОВ НЕЙРОННЫХ СЕТЕЙ ПРИ РЕШЕНИИ ОБРАТНЫХ ЗАДАЧ

Шипицын Ю.Д.
ФГАОУ ВО «ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ», Волгоград,, Россия (400062, Волгоградская область, город Волгоград, Университетский пр-кт, д. 100), e-mail: ura2002ura@mail.ru

В работе рассматривается задача оптимизации гиперпараметров нейронной сети, применяемой для решения обратной задачи восстановления параметров гармонического сигнала. В качестве модельной задачи используется восстановление амплитуды, циклической частоты и начальной фазы синусоидального сигнала по его дискретным зашумлённым отсчётам. Для решения задачи применяется полносвязная нейронная сеть прямого распространения, обученная на синтетическом датасете объёмом 10 000 примеров. Сравниваются три метода автоматической оптимизации гиперпараметров: метод полного перебора (Grid Search), случайный поиск (Random Search) и алгоритм Hyperband, основанный на механизме последовательного отсева слабых конфигураций. Бюджет поиска для всех методов зафиксирован на уровне 30 конфигураций. Оценка качества проводится по метрикам MSE и MAE на тестовой выборке. Результаты эксперимента показывают, что метод Random Search обеспечивает наилучшее качество восстановления параметров при умеренных временных затратах, тогда как алгоритм Hyperband демонстрирует наименьшее время поиска за счёт ранней остановки слабых конфигураций. Полученные результаты могут служить ориентиром при выборе метода оптимизации гиперпараметров в задачах, требующих применения нейросетевых моделей.

Ключевые слова: нейронная сеть, обратная задача, оптимизация гиперпараметров, Grid Search, Random Search, Hyperband, восстановление параметров сигнала, машинное обучение, сравнительный анализ.

A COMPARISON OF NEURAL NETWORK HYPERPARAMETER OPTIMIZATION STRATEGIES FOR SOLVING INVERSE PROBLEMS

Shipitsyn Y.D.
VOLGOGRAD STATE UNIVERSITY, Volgograd, Russia (400062, Volgograd region, Volgograd, Universitetsky pr-kt, 100), e-mail: ura2002ura@mail.ru

This paper addresses the problem of hyperparameter optimization for a neural network applied to solving an inverse problem of harmonic signal parameter recovery. The model task consists in recovering the amplitude, cyclic frequency, and initial phase of a sinusoidal signal from its discrete noisy samples. A fully connected feedforward neural network trained on a synthetic dataset of 10,000 examples is used to solve the problem. Three automatic hyperparameter optimization methods are compared: Grid Search, Random Search, and the Hyperband algorithm based on the successive halving mechanism. The search budget is fixed at 30 configurations for all methods. Model quality is evaluated using MSE and MAE metrics on the test set. Experimental results show that Random Search achieves the best parameter recovery quality at moderate computational cost, while Hyperband demonstrates the shortest search time due to early stopping of weak configurations. The obtained results can serve as a practical guide when selecting a hyperparameter optimization method for problems that require neural network models.

Keywords: neural network, inverse problem, hyperparameter optimization, Grid Search, Random Search, Hyperband, signal parameter recovery, machine learning, comparative analysis.

Введение

Нейронные сети находят широкое применение при решении обратных задач в различных прикладных областях — обработке сигналов, геофизике, медицинской диагностике [3]. Качество получаемого решения при этом во многом определяется не только архитектурой сети, но и выбором её гиперпараметров: скоростью обучения, числом слоёв и нейронов, типом функции активации и рядом других характеристик. Ручной подбор этих параметров трудоёмок, требует значительного опыта и не гарантирует нахождения оптимальной конфигурации.

Задача автоматической оптимизации гиперпараметров активно исследуется в последние годы. Среди наиболее распространённых подходов — метод полного перебора (Grid Search), случайный поиск (Random Search) и алгоритмы с ранней остановкой, в частности Hyperband. Бергстра и Бенжио показали, что случайный поиск при равном вычислительном бюджете превосходит систематический перебор в задачах, где лишь часть гиперпараметров оказывает существенное влияние на целевую метрику [1]. Алгоритм Hyperband, предложенный Ли и соавторами, развивает эту идею за счёт механизма последовательного отсева слабых конфигураций, что позволяет экономить вычислительные ресурсы на дорогостоящих задачах [2].

Вместе с тем сравнительный анализ указанных методов в контексте решения обратных задач с помощью нейронных сетей представлен в литературе недостаточно полно. Настоящая работа направлена на восполнение этого пробела. В качестве модельной задачи рассматривается восстановление параметров гармонического сигнала по его дискретным зашумлённым отсчётам. Цель работы — провести сравнительный анализ трёх методов оптимизации гиперпараметров по критериям качества найденного решения и времени поиска.

Статья организована следующим образом. В разделе 1 формулируется постановка обратной задачи. Раздел 2 описывает модель и используемый датасет. В разделе 3 приводится описание методов оптимизации. Раздел 4 содержит результаты вычислительного эксперимента и их анализ. В заключении сформулированы основные выводы.

1. Постановка задачи

Большинство физических и инженерных задач принято формулировать в прямой постановке: известны параметры системы — найти её отклик. Такие задачи, как правило, хорошо изучены и допускают единственное решение. Обратная задача ставится иначе: наблюдатель видит лишь результат — сигнал, распределение, измерение — и должен по нему восстановить породившие его параметры. Именно этот класс задач представляет наибольший практический интерес в обработке сигналов, геофизике, медицинской диагностике и ряде других областей.

В данной работе в качестве модельной обратной задачи рассматривается восстановление параметров гармонического сигнала. Пусть наблюдаемый сигнал имеет вид:

$$y(t) = A\sin(\omega t + \varphi) + \varepsilon(t),$$

где A — амплитуда, ω — циклическая частота, φ — начальная фаза, а $\varepsilon(t)$ — аддитивный шум. По набору дискретных отсчётов $\{y(t_i)\}_{i=1}^{N}$ требуется определить тройку параметров (A, ω, φ). Задача осложняется присутствием шума: при $\varepsilon \neq 0$ однозначное аналитическое решение перестаёт существовать, и классические спектральные методы дают заметную погрешность уже при умеренном уровне зашумлённости [5].

Нейронные сети позволяют обойти это ограничение. Обученная на достаточном объёме данных сеть приближает нелинейное отображение из пространства сигналов в пространство параметров, не требуя явного вида решающего правила.

2. Описание модели и данных

Для обучения и тестирования нейронной сети был сформирован синтетический датасет из 10 000 примеров. Каждый пример представляет собой пару: дискретный сигнал и соответствующий ему вектор параметров. Параметры синусоиды выбирались случайно и равномерно из следующих диапазонов: амплитуда $A \in [0.5, 2.0]$, циклическая частота $\omega \in [0.5, 3.0]$, начальная фаза $\varphi \in [0, \pi]$. Сигнал дискретизировался на отрезке $t \in [0, 2\pi]$ равномерно по $N = 32$ точкам, после чего к каждому отсчёту добавлялся аддитивный гауссовский шум с нулевым математическим ожиданием и стандартным отклонением $\sigma = 0.05$:

$$y(t_i) = A\sin(\omega t_i + \varphi) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Датасет разбивался на обучающую и тестовую выборки в соотношении $80/20$, то есть 8 000 и 2 000 примеров соответственно.

В качестве модели использовалась полносвязная нейронная сеть прямого распространения. Входной слой принимает вектор из 32 отсчётов сигнала, выходной слой содержит 3 нейрона без функции активации — по одному на каждый восстанавливаемый параметр. Число скрытых слоёв, количество нейронов в каждом из них, функция активации и ряд других характеристик архитектуры не фиксировались заранее — они являются гиперпараметрами и подбираются в ходе оптимизации. В качестве функции потерь использовалась среднеквадратическая ошибка:

$$\mathcal{L} = \frac{1}{n}\sum_{j=1}^{n}\left\|\hat{\theta}_j - \theta_j\right\|^2,$$

где $\theta_j = (A_j, \omega_j, \varphi_j)$ — истинный вектор параметров j-го примера, $\hat{\theta}_j$ — предсказание сети, n — размер батча. Обучение проводилось с помощью оптимизатора Adam [4].

3. Методы оптимизации гиперпараметров

Метод полного перебора, или Grid Search, является наиболее простым подходом к оптимизации гиперпараметров. Идея проста: исследователь заранее задаёт конечное множество допустимых значений для каждого гиперпараметра, а алгоритм последовательно перебирает все возможные комбинации, обучая отдельную модель для каждой из них.

Формально, пусть оптимизируется m гиперпараметров, и для k-го из них задано множество значений $\Lambda_k = \{\lambda_k^{(1)}, \lambda_k^{(2)}, \dots, \lambda_k^{(n_k)}\}$. Тогда пространство поиска представляет собой декартово произведение:

$$\Lambda = \Lambda_1 \times \Lambda_2 \times \dots \times \Lambda_m,$$

а суммарное число проверяемых конфигураций равно $|\Lambda| = \prod_{k=1}^{m} n_k$. Для каждой конфигурации $\lambda \in \Lambda$ обучается модель и вычисляется значение целевой метрики — в данной работе среднеквадратическая ошибка на валидационной выборке. По завершении перебора выбирается конфигурация с наилучшим результатом.

Главное достоинство метода — гарантированный охват всего заданного пространства поиска. Если оптимальное сочетание гиперпараметров входит в сетку, алгоритм его найдёт.

Однако вычислительная стоимость растёт мультипликативно с числом гиперпараметров — явление, известное как проклятие размерности. В эксперименте пространство поиска включает 5 гиперпараметров, что даёт $3 \times 3 \times 2 \times 3 \times 3 = 162$ возможных комбинации; из них в рамках бюджета проверялось 30.

Метод случайного поиска предлагает принципиально иной подход: вместо систематического перебора узлов сетки каждая конфигурация гиперпараметров формируется независимой случайной выборкой из заданного пространства поиска. На первый взгляд это выглядит как шаг назад — отказ от какой-либо структуры в пользу случайности. Однако именно здесь кроется ключевое преимущество метода.

Бергстра и Бенжио показали, что в большинстве реальных задач машинного обучения целевая метрика существенно зависит лишь от небольшого числа гиперпараметров, тогда как остальные оказывают на неё слабое влияние [1]. В условиях такой низкой эффективной размерности случайный поиск за то же число проверяемых конфигураций исследует значительно больше различных значений каждого важного гиперпараметра, чем регулярная сетка. При фиксированном бюджете в T проверок случайный поиск проецирует на каждую ось Tразличных значений, тогда как сеточный — лишь $\sqrt{T}$ при двух параметрах.

Формально на каждом шаге $t = 1, \dots, T$ конфигурация $\lambda^{(t)}$ сэмплируется независимо:

$$\lambda^{(t)} \sim p(\Lambda),$$

где $p(\Lambda)$ — распределение над пространством гиперпараметров, чаще всего равномерное. По завершении всех Tитераций возвращается конфигурация с наименьшим значением целевой метрики:

$$\lambda^{*} = \arg\min_{\lambda^{(t)}} \mathcal{L}_{val}\left(\lambda^{(t)}\right).$$

Среди недостатков метода следует отметить отсутствие какой-либо адаптации: каждая новая конфигурация выбирается без учёта результатов предыдущих. При ограниченном бюджете это означает, что часть вычислительных ресурсов неизбежно тратится на заведомо слабые комбинации параметров.

Оба рассмотренных метода объединяет одна черта: каждая конфигурация гиперпараметров обучается до конца, независимо от того, насколько перспективной она оказывается на промежуточных эпохах. Это расточительно — слабая конфигурация потребляет столько же вычислительных ресурсов, сколько и сильная. Алгоритм Hyperband, предложенный Ли и соавторами [2], устраняет этот недостаток за счёт механизма последовательного отсева.

В основе Hyperband лежит процедура Successive Halving. Алгоритм запускает n конфигураций, обучает каждую из них в течение минимального числа эпох r_{min}, затем оставляет только $\lfloor n/\eta \rfloor$ наиболее перспективных и продолжает их обучение. Процедура повторяется до тех пор, пока не останется единственная конфигурация-победитель. При коэффициенте отсева $\eta = 3$ каждый раунд отбрасывает две трети кандидатов:

$$n_i = \left\lfloor n \cdot \eta^{-i} \right\rfloor, \quad r_i = r_{min} \cdot \eta^{i}, \quad i = 0,1, \dots, s$$

где i— номер раунда, n_i— число выживших конфигураций, r_i— число эпох обучения на данном раунде, s— максимальное число раундов, определяемое из условия $r_s \leq r_{max}$.

Проблема Successive Halving состоит в следующем: при фиксированном бюджете необходимо выбрать, что важнее — проверить больше конфигураций при меньшем числе эпох или меньше конфигураций, но обучать их дольше. Hyperband решает эту дилемму, запуская

несколько скобок Successive Halving с разными соотношениями n и r_{min}, а затем возвращая лучший результат среди всех скобок.

Принципиальное преимущество метода проявляется именно на вычислительно дорогих задачах: ранняя остановка слабых конфигураций позволяет за тот же бюджет исследовать значительно большее пространство гиперпараметров. В эксперименте для обеспечения честного сравнения бюджет поиска был зафиксирован на уровне 30 конфигураций для каждого из трёх методов.

4. Вычислительный эксперимент

Эксперимент проводился на синтетическом датасете, описанном ранее. Реализация выполнена на языке Python с использованием библиотек TensorFlow и keras-tuner [6, 7]. Бюджет поиска для всех трёх методов зафиксирован на уровне 30 конфигураций.

Пространство гиперпараметров, в котором осуществлялся поиск, представлено в Таблице 1.

Таблица 1 — Пространство гиперпараметров

Гиперпараметр	Допустимые значения
Число скрытых слоёв	1, 2, 3
Число нейронов в слое	8, 32, 64, 128
Функция активации	ReLU, tanh
Скорость обучения	0,01; 0,001; 0,0001
Вероятность Dropout	0,0; 0,1; 0,2

Целевой метрикой при поиске служила среднеквадратическая ошибка на валидационной выборке. После завершения поиска лучшая найденная конфигурация дообучалась до 50 эпох с ранней остановкой при отсутствии улучшения в течение 5 эпох подряд. Качество итоговой модели оценивалось на тестовой выборке по двум метрикам — MSE и MAE.

Результаты сравнения методов приведены в Таблице 2.

Таблица 2 — Сравнение методов оптимизации гиперпараметров

Метод	MSE	MAE	Trials	Время поиска, с
Grid Search	0,0295	0,1294	30	381,4
Random Search	0,0029	0,0391	30	183,2
Hyperband	0,0205	0,1140	30	150,9

Наилучшее качество восстановления параметров синусоиды продемонстрировал метод Random Search: значения MSE и MAE оказались в 10 и 3 раза ниже, чем у Grid Search, и в 7 и 3 раза ниже, чем у Hyperband соответственно. Найденная конфигурация — 2 скрытых слоя по 32 нейрона, функция активации ReLU, скорость обучения 0,001 — обеспечила наиболее точное приближение обратного отображения.

Hyperband показал наименьшее время поиска — 150,9 с против 183,2 с у Random Search и 381,4 с у Grid Search. Это объясняется механизмом ранней остановки: слабые конфигурации отсеивались после нескольких эпох, не расходуя вычислительные ресурсы на полное обучение. Вместе с тем на данной задаче ранняя остановка привела к тому, что ряд потенциально сильных конфигураций был отброшен прежде, чем успел раскрыть свой потенциал, что негативно сказалось на итоговом качестве модели.

Grid Search, несмотря на систематический охват пространства поиска, затратил наибольшее время и показал наихудшее качество. Равномерное распределение точек по сетке не позволило сосредоточить вычислительный ресурс в перспективных областях пространства гиперпараметров.

Заключение.

В данной работе рассмотрена задача восстановления параметров гармонического сигнала как модельная обратная задача, решаемая с помощью полносвязной нейронной сети. Для подбора архитектуры и гиперпараметров сети были применены три метода автоматической оптимизации: Grid Search, Random Search и Hyperband. Бюджет поиска для всех методов был зафиксирован на уровне 30 конфигураций, что обеспечило корректность сравнения.

Проведённый вычислительный эксперимент показал, что методы существенно различаются как по качеству найденного решения, так и по времени поиска. Наилучшее качество восстановления параметров синусоиды продемонстрировал метод Random Search: значение MSE на тестовой выборке составило 0,0029, что примерно в 10 раз ниже, чем у Grid Search, и в 7 раз ниже, чем у Hyperband. Наименьшее время поиска показал алгоритм Hyperband — 150,9 с, что объясняется механизмом ранней остановки слабых конфигураций.

Полученные результаты подтверждают вывод, сформулированный в работе Бергстра и Бенжио [1]: случайный поиск эффективнее систематического перебора при наличии гиперпараметров, слабо влияющих на целевую метрику. Алгоритм Hyperband, в свою очередь, демонстрирует преимущество по скорости, однако на задачах с небольшой вычислительной стоимостью одного trial механизм ранней остановки может приводить к преждевременному отсеву перспективных конфигураций.

В качестве направлений дальнейших исследований представляет интерес применение байесовских методов оптимизации гиперпараметров, а также проверка полученных выводов на более сложных обратных задачах с многомерным пространством параметров.

Список литературы

1. Бергстра Дж., Бенджио Й. Случайный поиск для оптимизации гиперпараметров // Журнал исследований машинного обучения. — 2012. — Том 13. — С. 281–305.
2. Ли Л., Джеймисон К., ДеСальво Г., Ростамизаде А., Талвалкар А. Гиперполоса: новый подход к оптимизации гиперпараметров на основе многорукого бандита // Журнал исследований машинного обучения. — 2018. — Том 18. — С. 1–52.
3. Гудфеллоу И., Бенджио Й., Курвиль А. Глубокое обучение. — Кембридж: MIT Press, 2016. — 800 с.
4. Кингма Д., Ба Дж. Адам: метод стохастической оптимизации // Труды ICLR. — 2015.

5. Тихонов А. Н., Арсенин В. Ю. Решения некорректно поставленных задач. — Вашингтон: Уинстон, 1977. — 258 с.
6. Абади М. и др. TensorFlow: крупномасштабное машинное обучение на гетерогенных системах. — 2015. — URL: https://www.tensorflow.org
7. О'Малли Т. и др. KerasTuner. — 2019. — URL: https://github.com/keras-team/keras-tuner

References

1. Bergstra J., Bengio Y. Random search for hyper-parameter optimization // Journal of Machine Learning Research. — 2012. — Vol. 13. — pp. 281–305.
2. Li L., Jamieson K., DeSalvo G., Rostamizadeh A., Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization // Journal of Machine Learning Research. — 2018. — Vol. 18. — pp. 1–52.
3. Goodfellow I., Bengio Y., Courville A. Deep Learning. — Cambridge : MIT Press, 2016. — p.800
4. Kingma D., Ba J. Adam: A method for stochastic optimization // Proceedings of ICLR. — 2015.
5. Tikhonov A. N., Arsenin V. Y. Solutions of Ill-Posed Problems. — Washington: Winston, 1977. — p.258
6. Abadi M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems. — 2015. — URL: https://www.tensorflow.org
7. O'Malley T. et al. KerasTuner. — 2019. — URL: https://github.com/keras-team/keras-tuner

Бокий Р.Ю. Принципы проектирования игровых симуляторов в виртуальной реальности // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7 (69) Ч.1 с. 142–151

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.946:004.415

ПРИНЦИПЫ ПРОЕКТИРОВАНИЯ ИГРОВЫХ СИМУЛЯТОРОВ В ВИРТУАЛЬНОЙ РЕАЛЬНОСТИ

Бокий Р.Ю.
ФГБОУ ВО «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ», Москва, Россия (119454, г. Москва, пр-т Вернадского, д. 78, стр. 4), e-mail: sphagnum.bokij@bk.ru

В статье рассматриваются основные принципы проектирования игровых симуляторов в виртуальной реальности. Проанализированы особенности пользовательского взаимодействия, проблемы обеспечения комфорта, интерактивности и производительности VR-приложений. Рассмотрены принципы интуитивности, эргономичности, отзывчивости, стабильности и кооперативности, влияющие на качество пользовательского опыта. Отдельное внимание уделено вопросам реализации кооперативных механик и поддержания эффекта погружения в виртуальную среду. Сделан вывод о необходимости комплексного применения данных принципов при разработке VR-симуляторов.

Ключевые слова: Виртуальная реальность, VR-симуляторы, игровые симуляторы, проектирование VR-приложений, интерактивность, пользовательское взаимодействие, кооперативные механики.

PRINCIPLES OF DESIGNING GAME SIMULATORS IN VIRTUAL REALITY

Bokiy R.Yu.
MIREA - RUSSIAN TECHNOLOGICAL UNIVERSITY, Moscow, Russia (119454, Moscow, avenue Vernadsky, 78, b. 4), e-mail: sphagnum.bokij@bk.ru

This article examines the fundamental principles of designing virtual reality game simulators. It analyzes the characteristics of user interaction, as well as challenges related to ensuring comfort, interactivity, and performance in VR applications. The principles of intuitiveness, ergonomics, responsiveness, stability, and cooperativeness—which influence the quality of the user experience—are examined. Special attention is given to the implementation of cooperative mechanics and maintaining the sense of immersion in the virtual environment. The conclusion is drawn that these principles must be comprehensively applied when developing VR simulators.

Keywords: Virtual reality, VR simulators, game simulators, VR application design, interactivity, user interaction, cooperative mechanics.

На сегодняшний день можно с уверенностью утверждать, что сфера информационных технологий и цифровых интерактивных систем является одной из самых быстроразвивающихся областей: рост вычислительных мощностей, совершенствование графических технологий, модернизация и удешевление производства визуальных устройств – всё это способствует стремительному развитию данного направления, в рамках которого можно отдельно выделить технологии виртуальной реальности (virtual reality, VR). Современные технологии позволяют значительно повысить качественные характеристики создаваемых виртуальных сред и применять новые решения в рамках улучшения взаимодействия пользователя с цифровым пространством. Ранее известные как атрибут исключительно игровой индустрии, VR-технологии сегодняшнего дня всё чаще находят применение в иных отраслях, таких как промышленность, медицина, моделирование,

образование и не только – при этом сохраняя в своей основе «игровой» элемент взаимодействия с пользователем, что обуславливается таким значимым элементом виртуальной реальности, как обеспечение эффекта присутствия и формирование условий для более глубокого погружения в смоделированную среду.

Тем не менее, немалую значимость VR-технологии по-прежнему сохраняют в области разработки игровых приложений, где элементы реалистичности и интерактивности непосредственно оказывают влияние на опыт пользователя. Одной из особенностей виртуальной реальности, которая отличает её от компьютерных игр в традиционном представлении, является реализация взаимодействия игрока с внутриигровыми объектами и окружением более естественным образом, базирующимся на восприятии пространства и физических движениях пользователя. В связи с этим отдельный интерес в рамках разработки представляют игровые VR-симуляторы, которые позволяют имитировать и воспроизводить различные формы одиночной и совместной деятельности. Так, за счёт пространственного взаимодействия, использования естественных движений пользователя, а также эффекта присутствия повышается уровень реалистичности и вовлечённости в ходе игрового процесса. Это является одним из ключевых различий, выделяющим VR-симуляторы среди компьютерных игр в традиционном понимании: акцент на интуитивном взаимодействии, моделирование различных процессов, последовательность выполнения ключевых действий, а также элементы кооператива – всё это определяет особенности данной области, которые необходимо учитывать при разработке. В то же время, эти особенности взаимодействия пользователя с игровой виртуальной реальностью требуют и значительного внимания при разработке игровых VR-симуляторов.

В связи с тем, что виртуальная реальность предлагает широкие возможности для взаимодействия с пользователем, разработка игровых VR-симуляторов сопровождается рядом технических и проектных сложностей. Так, среди ключевых проблем разработки можно выделить обеспечение высокой производительности приложения, создание удобного и естественного взаимодействия игрока с виртуальной средой, поддержание стабильного уровня погружения, а также предотвращение дискомфорта при использовании VR-устройства. К тому же, к трудностям и особенностям разработки стоит отнести реализацию интерактивного окружения и реалистичной физики окружающих объектов, что также встречается и разработке традиционных компьютерных игр, но требует иного подхода и приоритизации при проектировании VR-симулятора.

Выделенная проблематика разработки игровых симуляторов в виртуальной реальности обуславливает необходимость определения и систематизации принципов проектирования, которые необходимо соблюсти для качественного улучшения пользовательского опыта и повышения эффективности разработки подобного рода приложений.

Представим данные принципы на схеме ниже (Рисунок 1) и подробнее рассмотрим каждый из них.

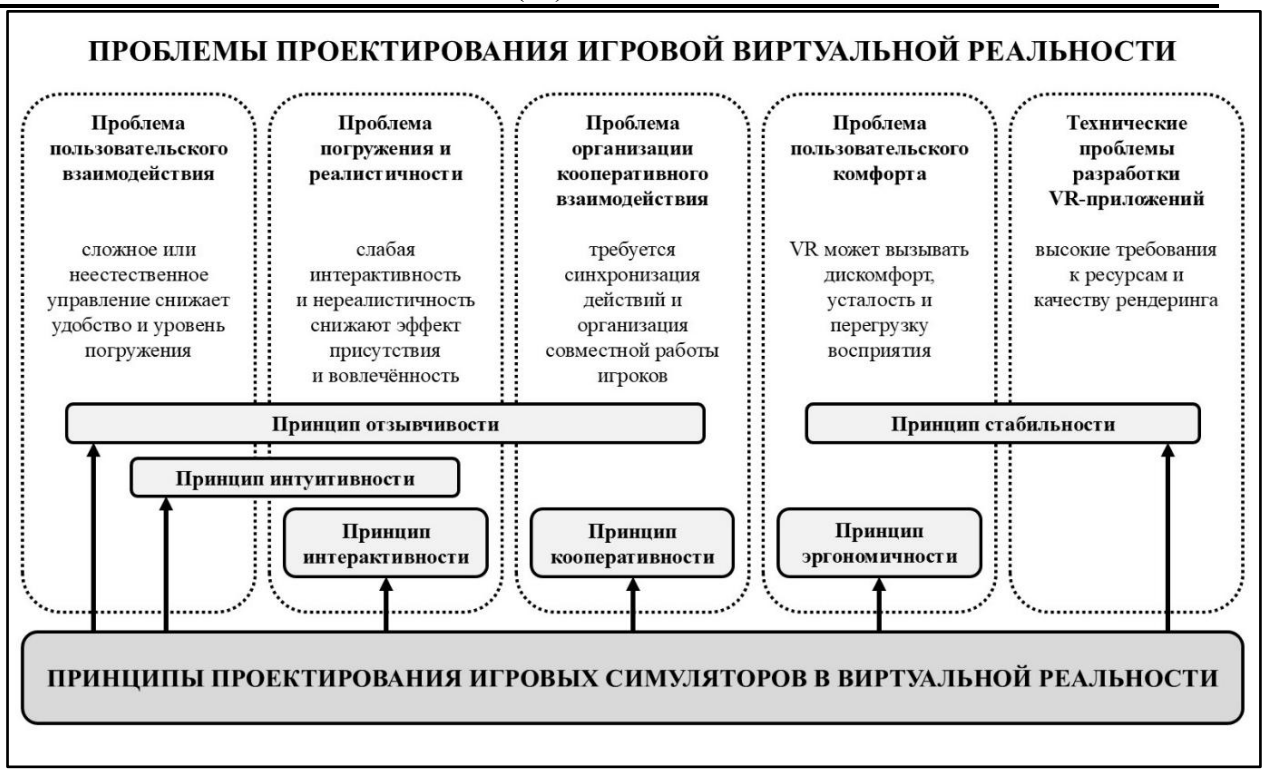


Рисунок 1 – Проблемы проектирования игровой виртуальной реальности и их покрытие предложенными принципами проектирования.

Принцип интуитивности

Для игровых симуляторов в виртуальной реальности принцип интуитивности является одним из базовых факторов, влияющих на качество взаимодействия. Если в традиционных компьютерных играх игрок постепенно адаптируется к системе клавиш, комбинациям кнопок и условным игровым действиям, то в VR такой подход работает значительно хуже. Пользователь находится внутри виртуального пространства и ожидает, что объекты будут взаимодействовать с ним по понятной и привычной логике.

На практике это особенно заметно при работе с предметами окружения. Даже небольшие несоответствия между ожидаемым действием и фактической реакцией виртуального объекта быстро становятся заметны. Для VR-приложений такая проблема критична, поскольку она напрямую влияет на уровень погружения и восприятие VR-среды.

Реализация принципа интуитивности обычно строится на использовании жестового управления и систем трекинга движений рук. За счёт этого взаимодействие с объектами можно приблизить к реальным действиям: пользователь физически тянется к предмету, берёт его, переносит в другую точку пространства или использует по назначению. Подобный подход активно применяется в VR-симуляторах, связанных с приготовлением пищи, ремонтом техники, управлением транспортом или совместным выполнением задач внутри виртуального окружения.

Отдельную роль играет скорость освоения управления. Чем больше пользователь вынужден запоминать комбинации кнопок, последовательности действий или особенности работы интерфейса, тем сильнее снижается естественность игрового процесса. Именно поэтому при проектировании VR-среды разработчики стремятся минимизировать количество абстрактных действий и сократить зависимость от классических интерфейсных элементов.

Интуитивность взаимодействия тесно связана и с реалистичностью поведения объектов VR-среды. Подобные проблемы особенно заметны в VR из-за эффекта присутствия и постоянного взаимодействия игрока с окружением на близкой дистанции. Даже кратковременные ошибки трекинга или рассинхронизация движений могут нарушать восприятие VR-среды и снижать общий уровень пользовательского опыта.

Принцип эргономичности

При разработке игровых симуляторов в виртуальной реальности особое значение приобретает принцип эргономичности, связанный с обеспечением комфортного и безопасного взаимодействия пользователя с виртуальной средой. В отличие от обычных компьютерных игр, VR-приложения оказывают значительно более сильное воздействие на восприятие человека, поскольку пользователь постоянно находится в режиме пространственного взаимодействия и воспринимает виртуальное окружение как часть реального пространства. Из-за этого любые ошибки в организации движения, отображении сцены или работе камеры ощущаются значительно сильнее и могут вызывать физический дискомфорт.

Одной из наиболее распространённых проблем VR-приложений является motion sickness – эффект укачивания, возникающий из-за несоответствия между визуальным восприятием движения и фактическим положением тела пользователя [1]. Чаще всего проблема возникает при резких перемещениях камеры, высокой скорости движения или нестабильной частоте кадров. Для игровых симуляторов это особенно критично, поскольку длительное взаимодействие с виртуальной средой является одной из ключевых особенностей жанра.

Снижение вероятности возникновения эффекта укачивания во многом зависит от организации перемещения пользователя внутри виртуального пространства. Камера в VR должна двигаться предсказуемо и синхронно с действиями игрока. По этой причине разработчики VR-приложений стараются избегать автоматических анимаций камеры, характерных для традиционных компьютерных игр.

В ряде случаев вместо плавного перемещения применяется система телепортации. Такой способ навигации существенно снижает нагрузку на вестибулярный аппарат, поскольку визуальное движение отсутствует [1]. Несмотря на определённое снижение реалистичности, телепортация остаётся одним из наиболее эффективных способов перемещения в VR-приложениях, ориентированных на длительное взаимодействие пользователя с виртуальной средой.

Дополнительной проблемой является визуальная перегрузка интерфейса и окружающего пространства. Избыточное количество интерфейсных окон, ярких эффектов, мелких объектов или быстро движущихся элементов увеличивает нагрузку на восприятие и усложняет ориентацию в пространстве. Подобная проблема особенно заметна в симуляторах с большим количеством интерактивных объектов, где игроку одновременно приходится отслеживать окружение, выполнять действия и взаимодействовать с игровыми системами.

При проектировании VR-симуляторов важно учитывать не только визуальный комфорт, но и техническую стабильность приложения. Если изображение начинает отображаться с задержкой или рывками, пользователь быстрее ощущает усталость и теряет чувство присутствия внутри VR-среды. На практике большинство современных VR-приложений ориентируется на частоту не ниже 72–90 FPS, поскольку более низкие значения заметно ухудшают пользовательский опыт.

Принцип эргономичности затрагивает и физическое взаимодействие пользователя с VR-системой. Поэтому при проектировании игровых механик разработчики стараются избегать чрезмерно сложных жестов, длительных однотипных действий и ситуаций, требующих постоянного физического напряжения. Чем естественнее и комфортнее организовано взаимодействие с виртуальной средой, тем дольше пользователь способен сохранять вовлечённость в игровой процесс без выраженного дискомфорта.

Принцип интерактивности

Для VR-симуляторов интерактивность окружения играет значительно более важную роль, чем для традиционных компьютерных игр. Из-за этого ожидания игрока по отношению к поведению объектов становятся заметно выше. Когда окружение не реагирует на действия игрока, это быстро разрушает эффект присутствия и снижает реалистичность VR-среды.

Одним из основных компонентов интерактивности являются физические взаимодействия между пользователем и объектами окружения. Например, при попытке поднять бутылку, инструмент или другой объект пользователь ожидает, что предмет будет корректно реагировать на движение руки, сохранять физические свойства и взаимодействовать с другими элементами сцены. Отсутствие подобных реакций делает виртуальное пространство статичным и снижает вовлечённость в игровой процесс.

Для игровых симуляторов особенно важна предсказуемость поведения объектов. Подобные ожидания формируются автоматически за счёт сходства виртуального окружения с реальной средой [2]. Например, невозможность взять предмет, который визуально выглядит доступным для взаимодействия, воспринимается как нарушение логики виртуального пространства.

Более глубокий уровень интерактивности связан с разрушением объектов и реакцией окружения на действия пользователя: разрушение стеклянных объектов, перемещение мебели, падение предметов со стола или изменение состояния окружающей среды делают взаимодействие более правдоподобным. Особенно хорошо такие механики работают в симуляторах, где игровой процесс строится вокруг постоянного взаимодействия с большим количеством объектов.

Интерактивность затрагивает не только статические элементы окружения, но и поведение NPC. Если персонаж не реагирует на приближение игрока, столкновения, жесты или действия в окружении, создаётся ощущение искусственности происходящего [5]. По этой причине современные VR-приложения всё чаще используют системы динамического поведения NPC, а также механизмы реагирования на действия пользователя и изменения окружающей среды.

Отдельная проблема связана с балансом между уровнем интерактивности и производительностью приложения. Для VR-приложений это особенно критично из-за высоких требований к стабильности FPS и минимизации задержек отображения. При недостаточной оптимизации даже небольшие просадки производительности могут нарушать восприятие VR-среды и снижать уровень пользовательского комфорта.

Таким образом, принцип интерактивности напрямую влияет на степень погружения пользователя в виртуальное пространство. Статичные или декоративные элементы, не поддерживающие взаимодействие, заметно снижают immersion и делают виртуальную среду менее убедительной. Именно поэтому при проектировании VR-симуляторов разработчики

стремятся создавать окружение, в котором большая часть объектов способна реагировать на действия пользователя и участвовать в игровом процессе.

Принцип отзывчивости

Для виртуальной реальности принцип отзывчивости играет одну из ключевых ролей, поскольку пользователь постоянно взаимодействует с объектами окружения и ожидает мгновенной реакции системы на собственные действия. Если пользователь касается предмета, перемещает объект или выполняет действие, виртуальная среда должна сразу подтверждать факт взаимодействия. При отсутствии такого подтверждения нарушается восприятие пространства и снижается уровень погружения [9].

Основная задача обратной связи заключается в том, чтобы пользователь понимал результат своих действий без необходимости дополнительного интерфейсного уведомления. Например, при столкновении объектов пользователь ожидает услышать звук удара, увидеть изменение положения предмета или почувствовать вибрацию контроллера. Если реакция отсутствует либо появляется с задержкой, взаимодействие начинает восприниматься искусственно.

На практике в VR-приложениях используется несколько типов обратной связи одновременно. Она включает изменение положения объектов, появление эффектов столкновения, изменение состояния предметов или отображение результатов взаимодействия. Даже простые визуальные реакции существенно повышают ощущение взаимодействия с виртуальной средой [3].

Не менее важной является звуковая обратная связь. Различные аудиосигналы позволяют определить тип взаимодействия, расстояние до объекта и состояние окружающей среды. Если звук воспроизводится с задержкой или не соответствует происходящему на экране, пользователь быстрее замечает несоответствие между действиями и реакцией виртуального пространства.

Дополнительный уровень погружения создаёт тактильная обратная связь. Подобные механизмы применяются при ударах, использовании инструментов, столкновениях предметов или взаимодействии с интерфейсными элементами. Особенно заметен эффект в симуляторах, где игровой процесс построен вокруг постоянного взаимодействия с объектами.

Отдельное значение имеет анимационная обратная связь, связанная с изменением поведения объектов и персонажей после действий пользователя. Аналогично объекты окружения могут изменять состояние после взаимодействия: механизмы начинают движение, инструменты активируются, а предметы разрушаются или деформируются. Подобные реакции делают виртуальное пространство более живым и предсказуемым.

Для VR-приложений особенно критична скорость формирования обратной связи. Если пользователь видит, что объект реагирует с опозданием, возникает ощущение потери контроля над окружением. На практике подобные проблемы часто связаны с недостаточной производительностью приложения, ошибками обработки физики или задержками трекинга движений.

Принцип отзывчивости напрямую влияет не только на реалистичность VR-среды, но и на удобство взаимодействия пользователя с игровыми механиками. Чем быстрее и понятнее система реагирует на действия игрока, тем легче воспринимается виртуальное пространство и тем выше общий уровень immersion. По этой причине современные VR-симуляторы

используют комбинированную систему обратной связи, одновременно объединяющую визуальные, звуковые, тактильные и анимационные способы отклика виртуальной среды.

Принцип стабильности

Для игровых симуляторов в виртуальной реальности производительность приложения является не только техническим параметром, но и одним из факторов, напрямую влияющих на восприятие VR-среды пользователем. В VR подобные проблемы ощущаются значительно сильнее, поскольку изображение обновляется одновременно для двух экранов внутри VR-шлема и постоянно синхронизируется с движениями головы пользователя. Даже незначительные задержки способны нарушать ощущение присутствия и вызывать motion sickness.

Для VR-приложений критически важна стабильная частота кадров. На практике это проявляется в виде задержек поворота камеры, дрожания изображения и общего ощущения нестабильности VR-среды. При снижении производительности пользователь значительно быстрее ощущает усталость и теряет чувство контроля над окружением.

Одним из основных способов повышения производительности является оптимизация трёхмерных моделей. В игровых симуляторах подобная нагрузка возрастает из-за необходимости одновременно обрабатывать физику объектов, освещение и пользовательские действия. Для снижения нагрузки разработчики используют упрощённые модели столкновений, оптимизированные текстуры и сокращение количества полигонов у второстепенных объектов.

Отдельное значение имеет оптимизация освещения. По этой причине разработчики часто комбинируют статическое и динамическое освещение, заранее рассчитывают тени и ограничивают количество одновременно активных источников света. Для VR-симуляторов это особенно важно, поскольку нестабильная работа системы освещения способна вызывать резкие просадки FPS даже при сравнительно небольшой сцене.

Серьёзную роль в оптимизации VR-приложений играет технология «occlusion culling». Например, если игрок находится внутри помещения, система может временно не обрабатывать объекты, расположенные за стенами или в другой части уровня. Подобный подход позволяет существенно снизить нагрузку на графическую подсистему и повысить стабильность работы приложения без заметного ухудшения качества изображения.

Дополнительно для оптимизации производительности используются LOD-модели (Level of Detail). Для пользователя подобная замена обычно остаётся незаметной, однако нагрузка на систему существенно уменьшается. В VR такой подход особенно эффективен для больших сцен с большим количеством объектов окружения, поскольку позволяет поддерживать стабильный FPS без значительного ухудшения визуального качества.

Проблема производительности становится ещё более сложной в многопользовательских VR-приложениях, где система дополнительно обрабатывает сетевую синхронизацию действий пользователей. Одновременная работа физики объектов, трекинга движений, кооперативных механик и рендеринга сцены создаёт высокую нагрузку как на процессор, так и на графическую подсистему. В результате принцип стабильности становится обязательным условием разработки VR-симуляторов, поскольку даже качественно реализованные игровые механики теряют эффективность при нестабильной работе приложения.

Принцип кооперативности

Для современных VR-симуляторов кооперативное взаимодействие становится одним из основных факторов, влияющих на глубину игрового процесса и уровень вовлечённости пользователей. В отличие от одиночных приложений, многопользовательские VR-среды позволяют участникам совместно выполнять игровые задачи, взаимодействовать с объектами окружения и координировать действия в режиме реального времени. За счёт эффекта присутствия подобное взаимодействие воспринимается значительно естественнее, чем в традиционных сетевых играх, поскольку пользователи находятся внутри общего виртуального пространства и способны напрямую наблюдать действия друг друга.

В игровых симуляторах кооперация особенно важна для процессов, требующих последовательного выполнения задач и распределения обязанностей между участниками. Подобное распределение ролей позволяет организовать игровой процесс таким образом, чтобы действия пользователей были взаимосвязаны и зависели друг от друга. На практике это делает взаимодействие между игроками более активным и снижает однообразие игрового процесса.

Отдельное значение имеет синхронизация действий пользователей внутри VR-среды. Если один пользователь перемещает предмет, а другой видит его в другом положении или с задержкой, нарушается логика взаимодействия и снижается реалистичность происходящего. По этой причине система должна обеспечивать стабильную передачу данных, корректное обновление состояния объектов и предсказуемую реакцию VR-среды для всех участников игрового процесса.

Кооперативное взаимодействие также напрямую связано с коммуникацией между игроками. Это создаёт более естественный формат совместного взаимодействия по сравнению с обычными многопользовательскими играми. Подобные механики делают совместное выполнение задач более понятным и повышают уровень вовлечённости пользователей в игровой процесс.

Для VR-симуляторов важна и организация игровых механик, ориентированных именно на кооперацию. По этой причине современные VR-симуляторы часто строятся вокруг задач, которые невозможно эффективно выполнять в одиночку. Подобный подход вынуждает пользователей координировать действия, распределять обязанности и учитывать действия других участников команды.

Дополнительная сложность возникает при проектировании интерфейсов и систем взаимодействия в многопользовательской среде. В VR большое количество интерфейсных элементов ухудшает восприятие пространства и снижает immersion, поэтому разработчики стараются переносить взаимодействие непосредственно в игровую среду. Информация о действиях других игроков может передаваться через анимации персонажей, положение объектов, звуковые сигналы или визуальные эффекты окружения.

Кооперативность существенно влияет на вариативность игрового процесса. На практике это увеличивает продолжительность интереса к приложению и делает игровой опыт менее предсказуемым. Для симуляторов подобная особенность особенно важна, поскольку большое количество однотипных действий без взаимодействия между игроками быстро приводит к ощущению монотонности.

Можно сказать, что принцип кооперативности позволяет сделать VR-симуляторы более динамичными, реалистичными и ориентированными на совместное взаимодействие

пользователей. При грамотной реализации кооперативные механики не только повышают вовлечённость игроков, но и расширяют возможности игрового процесса за счёт распределения ролей, совместного выполнения задач и постоянной коммуникации внутри виртуальной среды.

Таким образом, проектирование игровых симуляторов в виртуальной реальности требует комплексного подхода, учитывающего как технические ограничения VR-технологий, так и особенности пользовательского восприятия виртуальной среды. Проведённый анализ показал, что качество VR-приложения напрямую зависит от реализации принципов интуитивности, эргономичности, интерактивности, отзывчивости, стабильности и кооперативности. Нарушение хотя бы одного из данных принципов способно снижать уровень погружения, вызывать дискомфорт у пользователя и ухудшать восприятие игрового процесса. При этом наибольшую сложность представляет необходимость одновременного обеспечения высокой производительности, естественного взаимодействия и реалистичной реакции виртуального окружения. Применение рассмотренных принципов позволяет повысить вовлечённость пользователей, улучшить качество взаимодействия с VR-средой и обеспечить более эффективную разработку игровых симуляторов.

Список литературы

1. Khang Yeu Tang, Ge Yu b, Juhong Wang, Yu He, Sen-Zhe Xu, Song-Hai Zhang. Strategies for reducing motion sickness in virtual reality through improved handheld controller movements / Prof. Dr. Kai Xu, Dr. Song-Hai Zhang, Prof. Juyong Zhang // Graphical Models. – 2025. – Vol. 138. – pp. 2-14.
2. Глотова М.И. Основы разработки приложений виртуальной реальности: учебное пособие / М.И. Глотова. – Оренбург: Оренбургский государственный университет, 2024. – 258 с.
3. Жиленков, А.А. Современные проблемы проектирования и создания компьютерных тренажеров на базе симулятора вертолета / А.А. Жиленков, М.Ю. Серебряков // Оборонный комплекс – научно-техническому прогрессу России. – 2024. – № 1(161). – С. 26-32.
4. Кузьмин, Е.М. Разработка VR симулятора инженера-энергетика / Е.М. Кузьмин, В.С. Осипов, Н.И. Прыгунов // Научно-технический вестник Поволжья. – 2024. – № 11. – С. 45-47.
5. Мальцева, А.П. Проблемы разработки и использования VR-симуляторов в системе подготовки учителей / А.П. Мальцева, Н.М. Касаткина, В.В. Солтис // Поволжский педагогический поиск. – 2022. – № 2(40). – С. 78-86.
6. Романова Е.В., Курзаева Л.В., Давлеткиреева Л.З., Новикова Т.Б. Возможности технологий виртуальной реальности для разработки игровых приложений / Е.В. Романова, Л.В. Курзаева, Л.З. Давлеткиреева, Т.Б. Новикова // Открытое образование. – 2021. – № 25(5). – С. 31-40.
7. Создание реалистичного виртуального окружения на примере VR-симулятора сборки ПК в Unity / А.Э. Самарцева, И.С. Лифанов, С.А. Деревинский, И.Д. Тикшаев // Электронные средства и системы управления. Материалы докладов Международной научно-практической конференции. – 2025. – № 1-3. – С. 75-77.

8. Создание системы взаимодействия объектов для VR-симулятора в Unity / А.Д. Чикарев, А.О. Панкратов, А.Э. Самарцева, И. Д. Тикшаев // Электронные средства и системы управления. Материалы докладов Международной научно-практической конференции. – 2025. – № 1-3. – С. 82-84.
9. Томашин Е.Д., Арсентьев Д.А. Особенности разработки игр для виртуальной реальности / Е.Д. Томашин, Д.А. Арсентьев // Вестник науки. – 2020. – №1 (22). – С. 216-220.

References

1. Khang Yeu Tang, Ge Yu b, Juhong Wang, Yu He, Sen-Zhe Xu, Song-Hai Zhang. Strategies for reducing motion sickness in virtual reality through improved handheld controller movements / Prof. Dr. Kai Xu, Dr. Song-Hai Zhang, Prof. Juyong Zhang // Graphical Models. – 2025. – Vol. 138. – pp. 2-14.
2. Glotova M.I. Fundamentals of Virtual Reality Application Development: A Textbook / M.I. Glotova. – Orenburg: Orenburg State University, 2024. – p.258
3. Zhilenkov, A.A. Current Issues in the Design and Development of Computer-Based Training Simulators Using a Helicopter Simulator / A.A. Zhilenkov, M.Yu. Serebryakov // Defense Complex – Scientific and Technical Progress in Russia. – 2024. – No. 1(161). – pp. 26-32.
4. Kuzmin, E.M. Development of a VR Simulator for Power Engineers / E.M. Kuzmin, V.S. Osipov, N.I. Prygunov // Scientific and Technical Bulletin of the Volga Region. – 2024. – No. 11. – pp. 45-47.
5. Maltseva, A.P. Problems in the Development and Use of VR Simulators in Teacher Training / A.P. Maltseva, N.M. Kasatkina, V.V. Soltis // Volga Region Pedagogical Research. – 2022. – No. 2(40). – pp. 78-86.
6. Romanova E.V., Kurzaeva L.V., Davletkireeva L.Z., Novikova T.B. The Potential of Virtual Reality Technologies for the Development of Gaming Applications / E.V. Romanova, L.V. Kurzaeva, L.Z. Davletkireeva, T.B. Novikova // Open Education. – 2021. – No. 25(5). – pp. 31-40.
7. Creating a Realistic Virtual Environment Using the Example of a PC Assembly VR Simulator in Unity / A.E. Samartseva, I.S. Lifanov, S.A. Derevinsky, I.D. Tikshaev // Electronic Means and Control Systems. Proceedings of the International Scientific and Practical Conference. – 2025. – No. 1–3. – pp. 75-77.
8. Creating a system of object interaction for a VR simulator in Unity / A.D. Chikarev, A.O. Pankratov, A.E. Samartseva, I.D. Tikshaev // Electronic Control Systems and Devices. Proceedings of the International Scientific and Practical Conference. – 2025. – No. 1–3. – pp. 82-84.
9. Tomashin, E.D., Arsentiev D.A. Features of Virtual Reality Game Development / E.D. Tomashin, D.A. Arsentiev // Science Bulletin. – 2020. – No. 1 (22). – pp. 216-220.

Кан Д.Е. Беспилотные воздушные судна в государственном земельном надзоре: проблемы использования// Международный журнал информационных технологий и энергоэффективности. – 2026. –Т. 11 № 7(69) Ч.1 с. 152–159

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 528.44:629.78

БЕСПИЛОТНЫЕ ВОЗДУШНЫЕ СУДНА В ГОСУДАРСТВЕННОМ ЗЕМЕЛЬНОМ НАДЗОРЕ: ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ

Кан Д.Е.
ФГАОУ ВО "ДАЛЬНЕВОСТОЧНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ Владивосток, Россия, (690922, Приморский край, город Владивосток, остров Русский, п Аякс, д. 10), e-mail: kndseh@bk.ru

В статье рассматриваются особенности проведения государственного земельного надзора с использованием беспилотных воздушных судов (БВС) на территории Приморского края. На основании проведенного анализа работы инспекторов выявлены ключевые проблемы: технические проблемы использования и сложности согласования надзорных мероприятий с использованием БВС с межведомственными органами. Были предложены мероприятия по совершенствованию аспектов дальнейшего использования БВС в государственных целях: развитие межведомственного взаимодействия, совершенствование и дальнейшая модернизация БВС, состоящих на учете Росреестра.

Ключевые слова: Беспилотное воздушное средство, государственный земельный надзор, надзорный инструмент, аэрофотосъемка, земельные участки, воздушное пространство, межведомственные органы.

UNMANNED AERIAL VEHICLES IN STATE LAND SUPERVISION: PROBLEMS OF USE

Kan D.E.
FAR EASTERN FEDERAL UNIVERSITY, Vladivostok, Russia, (690922, Primorsky Krai, Vladivostok city, Russky Island, Ajax p., 10), e-mail: kndseh@bk.ru

The article examines the specifics of conducting state land supervision using unmanned aerial vehicles (UAVs) in the Primorsky Krai territory. Based on the analysis of inspectors' work, key problems have been identified: technical issues of use and difficulties in coordinating supervisory measures involving UAVs with interagency bodies. Measures have been proposed to improve aspects of further UAV use for state purposes: development of interagency cooperation, improvement and further modernization of UAVs registered with Rosreestr.

Keywords: Unmanned aerial vehicle, state land supervision, supervisory tool, aerial photography, land plots, airspace, interagency bodies.

Введение.

Проведение инструментального обследования земельного участка на предмет нарушения земельного законодательства является основным способом осуществления государственного земельного надзора. В ходе проведения контроля за соблюдением земельного законодательства, зачатую инспекторы сталкиваются со следующими проблемами: ограниченный доступ на земельные участки, большая площадь земельных участков, в особенности земель сельскохозяйственного назначения, сложный рельеф, мешающий проведению пешего инструментального обследования. Использование БВС в качестве надзорного инструмента может решить ряд проблем, указанных выше.

Актуальность исследований подтверждается увеличивающейся тенденцией использования БВС. Так в 2017-2018 годах ведомством в целях внедрения дистанционных технологий при проведении государственного земельного надзора был реализован проект по использованию БВС в двух пилотных регионах – Тульской области и Республике Татарстан. Анализ его результатов показал целесообразность применения БВС для проведения мероприятий по земельному надзору.

Позже с 2019 года Росреестр стал использовать БВС для осуществления государственного земельного надзора в 12 пилотных регионах.

В настоящие время согласно постановлению №1165, законодательно закреплено использование БВС Росреестром для проведения государственного земельного надзора [3].

Изучение надзорной деятельности с использованием беспилотников способствует формированию объективного представления о текущем положении дел в государственном земельном надзоре Приморского края, что позволит в дальнейшем разработать перспективные направления по его совершенствованию.

Цель работы – анализ практики проведения государственного земельного надзора с использованием БВС на основании экспертных оценок инспекторов Росреестра на территории Приморского края. Анализ опыта взаимодействия с межведомственными органами для проведения полетных мероприятий в целях надзора земельных участков Приморского края.

Задачи исследования:

1) проведение опроса инспекторов Росреестра и специалистов, использующих в своей профессиональной деятельности БВС;

2) выявить негативные моменты в работе с БВС на территории Приморского края.

Теоретические основы использования беспилотных воздушных судов.

Правовое регулирование использования БВС в Российской Федерации осуществляется Воздушным кодексом Российской Федерации. Согласно Воздушному кодексу, беспилотное воздушное судно (БВС) – это воздушное судно, управляемое, контролируемое в полете пилотом, находящимся вне борта такого воздушного судна [1].

Использование БВС в земельном надзоре подразумевает выход в воздушное пространство сотрудниками Росреестра, имеющих лицензию на пилотирование беспилотником.

Как отмечают в своей работе Д.А. Гура, И.Г. Марковский и А.А. Ряскин, использование БВС имеет ряд преимуществ в сравнении с пешим инструментальным обследованием, некоторые из которых – это большая площадь охвата, высокое пространственное разрешение, обследование труднодоступных участков, проведение надзорных мероприятий без взаимодействия контролируемым лицом [4, с. 142].

Также в своих статьях, авторы А.А. Сазанова и Д.О. Мелентьев, Е.С. Щебляков отмечают, что применение БВС облегчает работу надзорного органа (Россреестр) и позволяет выполнять осмотр целого массива земельных участков [5, 24-25], [6]. Авторы работ выделяют следующие проблемы при проведении пешего инструментально обследования:

1) ограниченный доступ на земельные участки, природные и искусственные ограждения ограничивают свободный доступ инспекторов;

2) невозможность получения координат границ земельных участков с помощью GNSS оборудования, связи с уровнем застройки территории или с расположением вблизи территории съемки объектов, блокирующих или искажающих сигналы спутников.

Ключевые проблемы использования БВС – технические недостатки, длительные сроки: согласования полетов, процесса подготовки, получения разрешений и последующего рассекречивания данных.

Текущая ситуация использования БВС в государственном земельном надзоре в Приморском крае.

В настоящее время наблюдается рост объемов проведение надзорных мероприятий с использованием БВС. Такая тенденция обусловлена тем, что с 1 марта 2025 года в России начал действовать новый федеральный закон №307 от 08.08.2024, направленный на борьбу с неиспользованием земельных участков [2]. Таким образом Правительство Российской Федерации пытается уменьшить количество неиспользуемых или нерационально используемых земель. Для более быстрого достижения цели по надзору земельных участков, Росреестру необходимо осуществлять действия по выходу в воздушное пространство.

В 2024 году в Приморском крае впервые проводились надзорные мероприятия с использованием БВС. За счет федерального бюджета сотрудникам Росреестра были предоставлены БВС GEOSCAN GEMINI 501, а сотрудники были направлены на обучение для получения статуса пилота беспилотных воздушных судов. В этом же году проводились первые надзорные мероприятия с использованием БВС. Согласно заданию, было выполнено 4 полетных мероприятий направленных на выявление признаков неиспользования земель сельскохозяйственного значения на территориях Вольно-Надеждинского района.

В 2025 году Росреестром было проведено 10 надзорных мероприятий с выходом в воздушное пространство, обследовано 11 земельных участок в Вольно-Надеждинском и Надеждинском районах.

В 2026 году у Росреестра планы на проведение надзорных мероприятий большего количество муниципальных районов. Решается вопрос о предоставлении БВС сотрудникам Росреестра в Уссурийске.

Исследование представленных данных свидетельствует о расширении охвата территории и об увеличивающейся тенденции использования.

Однако, статистические данные не позволяют выявить трудности использования БВС на территории Приморского края. Для более полного понимания особенностей проведения земельного надзора с использованием БВС был проведен опрос. Для более широкого анализа, были опрошены сотрудники Росреестра, имеющие лицензию на управление БВС и геодезисты коммерческих организаций Приморского края, использующие в своей профессиональной деятельности БВС со схожими техническими характеристиками.

Анализ результатов опроса специалистов использующих БВС в своей профессиональной деятельности.

Для проведения анализа были опрошены 6 специалистов, имеющих соответствующий опыт. Опрос проводился в 2026 году методом устного опроса на территории Владивостокского городского округа. Все респонденты имеют опыт работы с БВС в различных муниципальных районах Приморского края.

Результаты опроса позволяют выявить основные проблемы проведения полетных и предполетных мероприятий, а также влияние природно-климатических особенностей Приморского края на использование БВС.

Специалисты отмечают сложности использования БВС из-за технических недостатков, предоставленных им в работу БВС, данные проблемы возникают из-за природно-климатических особенностей Приморского края.

Стоит отметить, что климат Приморья – умеренный, муссонный. В летний период времени характерны частые выпадения осадков и туманы, а горный рельеф составляет более 70% территорий Приморского края, что в свою очередь чревато возникновением сильных порывов ветра и завихрениям воздуха.

Респондентам предлагалось ответить на вопросы влияния природно-климатических особенностей на использование БВС. Результаты опроса были сформированы в диаграмму (Рисунок 1).

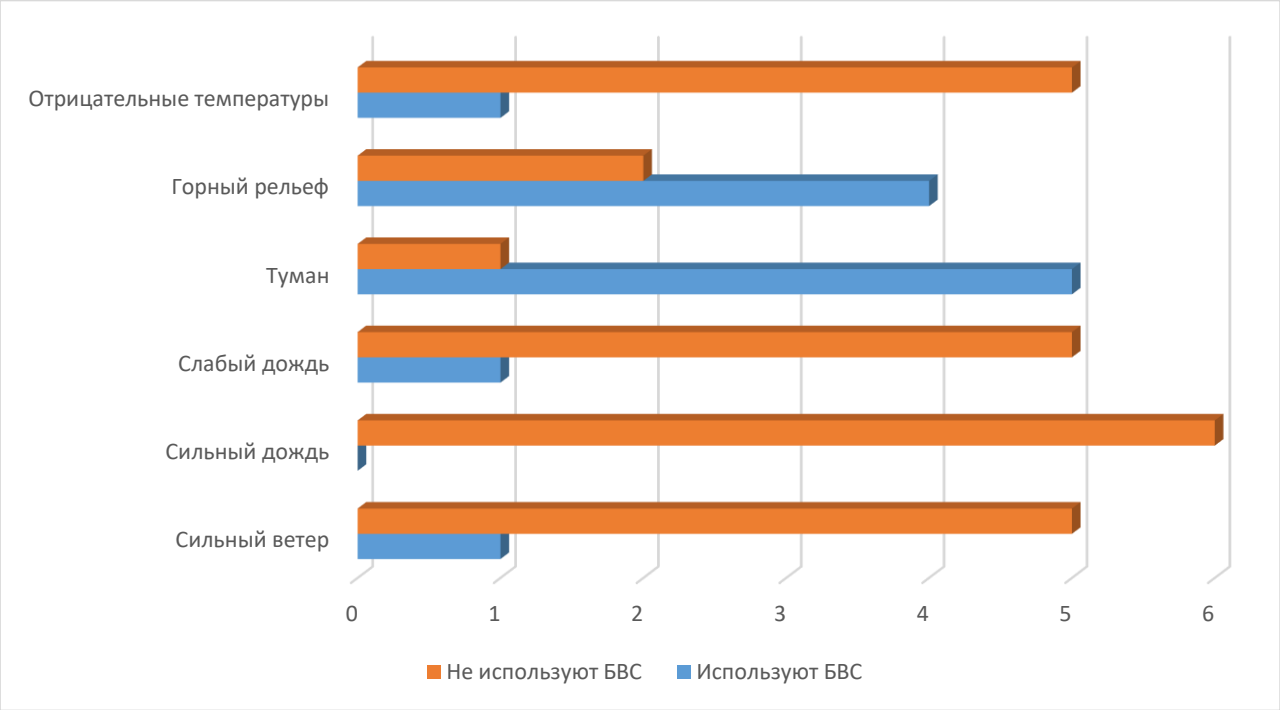


Рисунок 1 – Диаграмма использования БВС при различных природно-климатических особенностях Приморского края

Наиболее распространенной природно-климатической особенностью неиспользования БВС является дождь, сильный ветер и отрицательные температуры. 83,4% респондентов отказываются от использования БВС при сильных порывах ветра и отрицательных температурах, 100% опрошенных отказывается выходить в воздушное пространство из-за дождей. При этом влияние тумана на использование БВС не велико, 83,4% опрошенных готовы выйти в воздушное пространство.

Анализ результатов опроса демонстрирует проблемы технического характера, которые возникают из-за природно-климатических особенностей Приморского края. Природно-климатические особенности и их влияние на использование БВС на территории Приморского края, приведены в Таблице 1.

Таблица 1 - Природно-климатические особенности и их влияние на БВС

Природно-климатические особенности Приморского края	Влияние на БВС
Сильные северо-западные ветры зимой и мощные тайфуны с ливнями и шквалами летом. Порывистый ветер и сильная турбулентность над пересечённой горной местностью.	Резкое падение ёмкости аккумуляторов (батареи теряют до 30-40% эффективности), что сокращает время полёта. Риск выхода из строя электроники. Замерзание подвижных частей. Обледенение пропеллеров резко снижает подъёмную силу и может привести к падению дрона.
Температура: средняя температура января на большей части края значительно ниже -15°C, что критично для многих типов дронов Обледенение: особенно опасно при температурах около 0°C и высокой влажности, когда на пропеллерах образуется лёд.	Увеличивается риск потери управления и падения. Общие ограничения по скорости ветра для старта и посадки БВС составляют 4–15 м/с в зависимости от типа дрона, и в Приморье эти значения часто превышаются. Сильный боковой ветер на взлёте/посадке может привести к опрокидыванию аппарата. Также это ведёт к быстрому разряду аккумулятора и снижению дальности полёта.
Летом выпадает максимум осадков, сопровождающихся высокой влажностью. Зимой снегопады, особенно «мокрый» снег.	Осадки и туман значительно ухудшают видимость, делая полёты небезопасными. Попадание влаги внутрь корпуса вызывает короткие замыкания и коррозию. Туман также оседает на оптику, делая бесполезной камеру, что критично для визуальных полётов и аэросъёмки.
Горный рельеф	Сильная турбулентность и нисходящие потоки воздуха у подветренных склонов. Риск столкновения с препятствиями из-за складок местности. Проблемы с качеством и стабильностью каналов связи на пересечённой местности.

Дополнительный анализ открытых ответов респондентов позволил выявить ряд проблем, связанных с межведомственными согласованиями.

Сотрудники Росреестра ссылаются на непостоянство и отсутствие регламента порядка обращений в межведомственные органы, данный недостаток сместил сроки проведения надзорных мероприятий в 2025 году на 2 недели.

Стоит отметить, что для проведения мероприятий с использованием БВС, специалистам необходимо выполнить ряд предполетных мероприятий связанных с получением разрешений на выход в воздушное пространство, а именно:

1) разрешение штаба округов;

2) разрешение Генерального штаба

3) разрешение Управления Федеральной службы безопасности;

4) разрешение органов местного самоуправления;

5) разрешение на использование воздушного пространства от ЕС ОрВД.

Были изучены обращения в межведомственные органы и администрации городских округов Дальнего Востока для разрешения проведения надзорных мероприятий с использованием БВС за период 2024-2025 год, а также результаты обращений. В ходе анализа обращений, были выявлены большие сроки согласования между межведомственными органами и Росреестром.

Фактические сроки от начала подачи заявлений до получения всех разрешений в письменном виде для выхода в воздушное пространство – 3 месяца.

Как сообщают респонденты, на период 2026 года, сроки согласования сокращаются, однако проблема остается открытой

По результатам проведенного опроса, респонденты выявили ряд предложений, направленных на совершенствование проведения земельного надзора с использованием БВС.

Значительная часть опрошенных специалистов указывает на необходимость совершенствования имеющихся в арсенале Росреестра БВС или их замену на другие модели.

Согласно данным полученных из открытых источников, GEOSCAN GEMINI 501 имеет степень защиты – IP43, которая должна гарантировать влагозащиту корпуса беспилотника от падающих капель под углом до 60 градусов, однако, исходя из ответов респондентов, данная степень влагозащиты оказалась недостаточной для проведения мероприятий с использованием БВС в дождь.

Рекомендованное значение пыли-влагозащиты беспилотника IP55, что фактически позволит использовать БВС в дождь, а учитывая природно-климатические особенности Приморского края, использование БВС возможно почти в 2 раза чаще.

Также респонденты предлагают использовать БВС с функцией подогрева аккумуляторов для выполнения работ при отрицательных температурах, что позволяет выполнять надзорные мероприятия с использованием БВС круглогодично.

Наиболее подходящий БВС от российского производителя – GEOSCAN GEMINI 801. Также он позволяет получать видео в реальном времени, в отличии от 501 модели, что позволит выявлять нарушения земельного законодательства непосредственно во время его использования.

Также респонденты указывают на необходимость совершенствования взаимодействия Росреестра с межведомственными органами, среди наиболее часто упоминаемых предложений можно выделить:

1) создание цифрового портала подачи заявлений для согласования полетов;

2) упрощение процедуры согласования полетов для осуществление государственного земельного надзора;

3) упрощение процедур внеплановых проверок для случаев нецелевого использования земель;

4) на законодательном уровне отменить обязательную передачу снятого материала штабу вооруженных сил Российской Федерации для сотрудников Росреестра или усовершенствовать данный процесс.

Анализ данных предложений позволяет сделать вывод о том, что значительная часть проблем использования БВС в государственном земельном надзоре связана с техническими недостатками используемых БВС и в отсутствии регламента и быстрого взаимодействия Росреестра и межведомственных органов.

На основании анализа высказываний респондентов, представляется возможным определить ряд перспективных направлений для совершенствования дальнейшего использования БВС в государственном земельном надзоре.

Заключение.

Использование БВС в государственном земельном надзоре играет ключевую роль в борьбе с нарушениями земельного законодательства. Авторы вышеупомянутых статей и респонденты считают использование БВС достаточно эффективным. Они отмечают, что БВС помогает решить ряд проблем, с которыми специалисты сталкиваются при проведении пешего инструментального обследования.

Однако в ходе практического использования БВС выявились проблемы, среди которых существенно значимыми являются: технические недостатки используемых БВС и длительные роки согласования разрешений на проведение полетов.

Проведенное исследование подтверждает необходимость дальнейшего совершенствования взаимодействия сотрудников Росреестра и межведомственных органов, а также улучшение имеющихся или приобретение новых БВС для осуществления государственного земельного надзора. Реализация предложенных мер может помочь инспекторам быстрее и легче выявлять нарушения земельного законодательства.

Список литературы

1. Воздушный кодекс Российской Федерации от 19.03.1997 №60-ФЗ (ред. от 21.04.2025) (с изм. и доп., вступ. в силу с 02.05.2025) // принят Государственной Думой 19 февраля 1997 года. – URL:https://www.consultant.ru/document/cons_doc_LAW_13744/
2. Федеральный закон «О ведении гражданами садоводства и огородничества для собственных нужд и о внесении изменений в отдельные законодательные акты Российской Федерации» от 08.08.2024 №307-ФЗ // принят Государственной Думой 30 июля 2024 года. – URL:https://www.consultant.ru/document/cons_doc_LAW_83311/
3. Постановление Правительства РФ от 4 августа 2025г. №1165 «О внесении изменений в некоторые акты Правительства Российской Федерации по вопросам организации и осуществления федерального государственного земельного контроля (надзора)» // Официальное опубликование правовых актов. – URL:http://publication.pravo.gov.ru/document/0001202508050010?index=1
4. Гура Д.А., Марковский И.Г., Ряскин А. А. Использование беспилотных летательных аппаратов при осуществлении государственного земельного надзора: Статья/КубГТУ/Вестник. – М.: СГУГиТ, 2022. С. 138-146. – URL:https://cyberleninka.ru/article/n/ispolzovanie-bespilotnyh-letatelnyh-apparatov-pri-osuschestvlenii-gosudarstvennogo-zemelnogo-nadzora/viewer
5. Мелентьев Д.О., Щебляков Е.С. Правовое регулирование полетов БВС: Статья / СибГУ / Актуальные проблемы авиации и космонавтики. Том 3. – М.: СибГУ им. М. Ф.

Решетнева, 2019. С. 604-605. – URL: https://cyberleninka.ru/article/n/pravovoe-regulirovanie-polyotov-bvs/viewer

6. Сазанова А.А. Мировой опыт применения беспилотных авиационных технологий в государственном управлении: Статья / Государственное и муниципальное управление. – М.: Аэрмакс. Москва, 2024. С. 21-31. – URL: https://cyberleninka.ru/article/n/mirovoy-opyt-primeneniya-bespilotnyh-aviatsionnyh-tehnologiy-v-gosudarstvennom-upravlenii-obzor/viewer

References

1. Air Code of the Russian Federation No. 60-FZ of March 19, 1997 (as amended on April 21, 2025) (with amendments and additions, effective from May 2, 2025) // Adopted by the State Duma on February 19, 1997. – URL:https://www.consultant.ru/document/cons_doc_LAW_13744/
2. Federal Law "On Conducting Gardening and Market Gardening by Citizens for Their Own Needs and on Amendments to Certain Legislative Acts of the Russian Federation" No. 307-FZ of August 8, 2024 // Adopted by the State Duma on July 30, 2024. – URL: https://www.consultant.ru/document/cons_doc_LAW_83311/
3. Decree of the Government of the Russian Federation No. 1165 of August 4, 2025 "On Amendments to Certain Acts of the Government of the Russian Federation on the Organization and Implementation of Federal State Land Control (Supervision)" // Official Publication of Legal Acts. – URL:http://publication.pravo.gov.ru/document/0001202508050010?index=1
4. Gura D.A., Markovskii I.G., Ryaskin A.A. Use of Unmanned Aerial Vehicles in the Implementation of State Land Supervision: Article / KubSTU / Bulletin. – M.: SGUGiT, 2022. P. 138-146. – URL:https://cyberleninka.ru/article/n/ispolzovanie-bespilotnyh-letatelnyh-apparatov-pri-osuschestvlenii-gosudarstvennogo-zemelnogo-nadzora/viewer
5. Melentyev D.O., Shcheblyakov E.S. Legal Regulation of UAV Flights: Article / SibSU / Current Problems of Aviation and Cosmonautics. Vol. 3. – M.: Reshetnev Siberian State University of Science and Technology, 2019. P. 604-605. – URL: https://cyberleninka.ru/article/n/pravovoe-regulirovanie-polyotov-bvs/viewer
6. Sazanova A.A. World Experience in the Application of Unmanned Aviation Technologies in Public Administration: Article / State and Municipal Management. – M.: Aermax. Moscow, 2024. P. 21-31. – URL:https://cyberleninka.ru/article/n/mirovoy-opyt-primeneniya-bespilotnyh-aviatsionnyh-tehnologiy-v-gosudarstvennom-upravlenii-obzor/viewer

Попов Ю.Б. Моделирование алгоритма оценки направления и дальности в моноимпульсной угломерной системе / Попов Ю.Б., Призов А.С., Строило А.Ю. и др. // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 7(69) Ч.1 с. 160–167

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 621.396.969.182.34

МОДЕЛИРОВАНИЕ АЛГОРИТМА ОЦЕНКИ НАПРАВЛЕНИЯ И ДАЛЬНОСТИ В МОНОИМПУЛЬСНОЙ УГЛОМЕРНОЙ СИСТЕМЕ

Попов Ю.Б., [1] **Призов А.С., Строило А.Ю., Алаторцев С.К.**
ФГБОУ ВО "КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ", Краснодар, Россия (350040, Краснодарский край, город Краснодар, Ставропольская ул., д.149), e-mail:[1] prisow11@gmail.com.

В статье рассмотрена программная модель моноимпульсной угломерной системы, предназначенная для оценки азимута, расчёта координат цели и анализа влияния цифровой обработки на точность. Модель включает амплитудный и фазовый способы пеленгации, статистическое моделирование методом Монте-Карло, медианную фильтрацию и MUSIC-оценивание. Показано, что фазовый метод точнее амплитудного, а совместное применение MUSIC и медианной фильтрации существенно снижает координатную ошибку.

Ключевые слова: Моноимпульсная радиолокационная система; азимутальная оценка; фазовая пеленгация; амплитудная пеленгация; расчёт дальности; медианная фильтрация; алгоритм MUSIC; MATLAB; RMS.

MODELING OF AN ALGORITHM FOR ESTIMATING DIRECTION AND RANGE IN A MONOPULSE ANGLE-MEASUREMENT SYSTEM

Popov Yu.B.,[1] **Prizov A.S.**, **Stroilo A.Yu.**, **Alatortsev S.K.**
KUBAN STATE UNIVERSITY, Krasnodar, Russia (350040, Krasnodar region, Krasnodar, Stavropolskaya ul., d.149), e-mail:[1] prisow11@gmail.com.

The paper presents a software model of a monopulse angle-measurement system for azimuth estimation, target coordinate calculation and accuracy analysis. The model includes amplitude and phase direction finding, Monte Carlo simulation, median filtering and MUSIC direction-of-arrival estimation. The results show that the phase method is more accurate than the amplitude method, while the combined use of MUSIC and median filtering significantly reduces the coordinate error.

Keywords: Monopulse radar system; azimuth estimation; phase direction finding; amplitude direction finding; range calculation; median filtering; MUSIC algorithm; MATLAB; RMS.

Введение

Повышение точности определения направления на источник электромагнитного излучения остаётся одной из ключевых задач радиолокационных и радиотехнических измерительных систем. Моноимпульсный подход основан на одновременном сравнении сигналов нескольких приёмных каналов и позволяет получать угловую информацию без последовательного сканирования лучом. Это повышает быстродействие системы и снижает влияние флуктуаций амплитуды принятого сигнала [1].

В пассивной постановке одна угловая оценка задаёт только направление на объект. Для получения координат и дальности необходимо использовать не менее двух разнесённых

угломерных пунктов. При этом точность координатной оценки определяется не только ошибкой отдельного пеленга, но и геометрией пересечения лучей: при малом угле пересечения даже небольшие угловые ошибки могут приводить к существенным ошибкам дальности.

Цель исследования состоит в разработке и сравнительном анализе программной модели оценки азимута, расчёта дальности до цели и повышения точности угловых оценок с использованием медианной фильтрации и алгоритма MUSIC. Материал представлен как единая исследовательская схема: постановка задачи, математическая модель, методика моделирования, результаты и выводы.

1. Постановка задачи и структура модели

Рассматривается двумерная задача определения положения цели по измеренным азимутальным направлениям. Источник излучения находится в точке P, а две угломерные системы с известными координатами формируют оценки направлений θ_1 и θ_2. На первом этапе каждая система оценивает азимут амплитудным или фазовым методом. На втором этапе направления пересекаются в декартовой плоскости, что позволяет получить координатную оценку цели и дальность от каждого пункта наблюдения. На третьем этапе вводятся дополнительные алгоритмы повышения точности: медианная фильтрация последовательности угловых оценок и MUSIC-оценивание направления прихода сигнала.

Программная реализация выполнена в MATLAB. Моделирование организовано модульно: отдельно задаются параметры системы, амплитудная таблица соответствия, фазовый оценщик, геометрический блок дальности, блок метрик и блок визуализации. Такая структура позволяет анализировать как отдельную угломерную ошибку, так и итоговую координатную ошибку после пересечения двух направлений [2].

Таблица 1 - Основные параметры численного моделирования

Параметр	Значение	Назначение
Истинный азимут цели	15°	проверка смещения оценки относительно заданного направления
Истинная дальность	160 м	перевод угловой ошибки в линейную ошибку
Длина волны λ	0,03 м	расчёт фазовой разности
База фазового метода	$d/\lambda = 0,5$	однозначная фазовая оценка в рабочем секторе
Диапазон SNR	5–40 дБ	оценка влияния шумов
Число испытаний	3000	статистическое усреднение по Монте-Карло
Ширина ДН	55°	модель диаграммы направленности элемента

2. Математическая модель

В фазовом методе информативным параметром является разность фаз сигналов, принятых двумя разнесёнными элементами [3]. Для базы d и длины волны λ связь между разностью фаз и углом прихода записывается в виде

$$\Delta\varphi = \frac{2\pi d}{\lambda}\sin\theta \quad (1)$$

Отсюда при отсутствии фазовой неоднозначности оценка угла определяется выражением (2)

$$\hat{\theta} = \arcsin\left(\frac{\lambda\Delta\varphi}{2\pi d}\right) \quad (2)$$

При амплитудном подходе используются две частично перекрывающиеся диаграммы направленности. В модели применён суммарно-разностный признак, нормированный на сумму амплитуд двух каналов:

$$D(\theta) = \frac{A_1(\theta) - A_2(\theta)}{A_1(\theta) + A_2(\theta)} \quad (3)$$

Нормировка уменьшает зависимость оценки от общего уровня принимаемого сигнала. Для восстановления угла используется заранее рассчитанная таблица соответствия между значением D(θ) и углом θ. В присутствии аддитивного шума оценка выполняется только для реализаций, прошедших пороговое обнаружение.

Координатная задача решается через пересечение двух лучей. Для i-го угломерного пункта с координатами p_i и единичным направляющим вектором $u_i(\theta_i)$ луч задаётся параметрически:

$$\mathbf{p} = \mathbf{p}_i + t_i\mathbf{u}_i(\theta_i), \quad t_i \geq 0 \quad (4)$$

Положение цели оценивается по точке пересечения или ближайшего согласованного пересечения двух лучей [4]. Если угол пересечения лучей слишком мал, геометрическое решение считается некорректным и исключается из расчёта итоговых метрик.

Для повышения точности использован алгоритм MUSIC [5]. Для равномерной линейной антенной решётки рулевой вектор имеет вид

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{M}}\left[1, e^{j\frac{2\pi d}{\lambda}\sin\theta}, \ldots, e^{j\frac{2\pi(M-1)d}{\lambda}\sin\theta}\right]^T \quad (5)$$

Спектральная функция MUSIC строится по шумовому подпространству Еш:

$$P_{\mathrm{MUSIC}}(\theta) = \frac{1}{\mathbf{a}^H(\theta)E_{\text{ш}}E_{\text{ш}}^H\mathbf{a}(\theta)} \quad (6)$$

Оценкой направления считается угол, при котором спектральная функция достигает максимума. Дополнительно применяется медианная фильтрация, уменьшающая влияние одиночных выбросов в последовательности угловых оценок без существенного смещения основной траектории.

3. Методика оценки точности

Для каждого значения отношения сигнал/шум выполнялась серия статистических испытаний. В каждой реализации формировались шумовые отсчёты, проводилось пороговое обнаружение, вычислялась угловая оценка и, при наличии двух корректных пеленгов, определялась координатная оценка цели. Основной метрикой служила среднеквадратическая ошибка RMS [6]:

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{p}}_i - \mathbf{p}_0\|^2} \qquad (7)$$

Здесь $\hat{p}_i$ — оцененное положение цели в i-й реализации, p_0 — истинное положение цели, N — число корректных реализаций. Для отдельной угловой оценки линейная ошибка рассчитывалась как произведение дальности на модуль углового отклонения в радианах. Дополнительно анализировались вероятность обнаружения Рд, доля грубых ошибок на опорной дальности и вероятность получения корректного геометрического решения.

4. Результаты моделирования

Первый этап моделирования показывает, что увеличение отношения сигнал/шум снижает ошибку как амплитудного, так и фазового метода. При этом фазовый метод во всём исследованном диапазоне сохраняет меньшую среднеквадратическую ошибку, что связано с большей чувствительностью фазовой разности к изменению направления прихода.

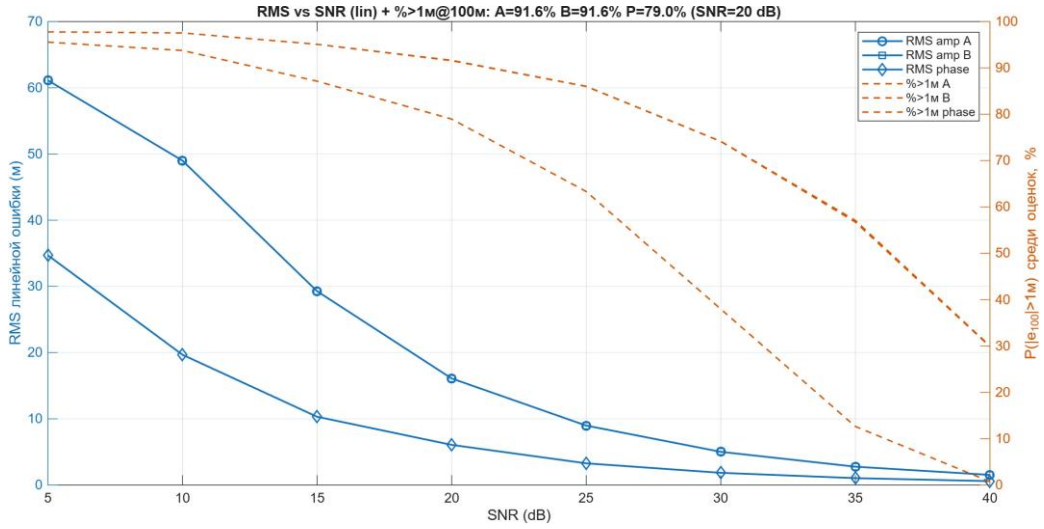


Рисунок 1 - Зависимость среднеквадратической ошибки оценки азимута от отношения сигнал/шум

Таблица 2 - Характерные значения ошибок угловой оценки

SNR, дБ	RMS амплитудного метода, м	RMS фазового метода, м	Комментарий
5	61,13	34,70	низкая энергетика сигнала
20	16,10	6,06	рабочая область сравнения методов
40	1,53	0,58	высокая энергетика сигнала

При SNR = 5 дБ среднеквадратическая ошибка амплитудного метода составила 61,13 м, фазового — 34,70 м. При SNR = 20 дБ значения уменьшились до 16,10 и 6,06 м соответственно, а при SNR = 40 дБ — до 1,53 и 0,58 м. Вероятность обнаружения также возрастала с ростом SNR: при 5 дБ она составила 0,434 для амплитудного метода и 0,721 для фазового, при 10 дБ

— 0,909 и 0,999 соответственно. Начиная с 15 дБ в данной серии моделирования оба метода обеспечили Рд = 1.

Переход к координатной задаче увеличивает влияние ошибок, поскольку результат зависит от двух угловых оценок и от угла пересечения направлений. При SNR = 20 дБ координатная ошибка амплитудной схемы составила 75,39 м, фазовой — 47,85 м. При SNR = 40 дБ соответствующие значения уменьшились до 24,08 и 9,40 м. Ошибка дальности при фазовой схеме на 40 дБ составила 9,38 м, а при амплитудной — 24,03 м.

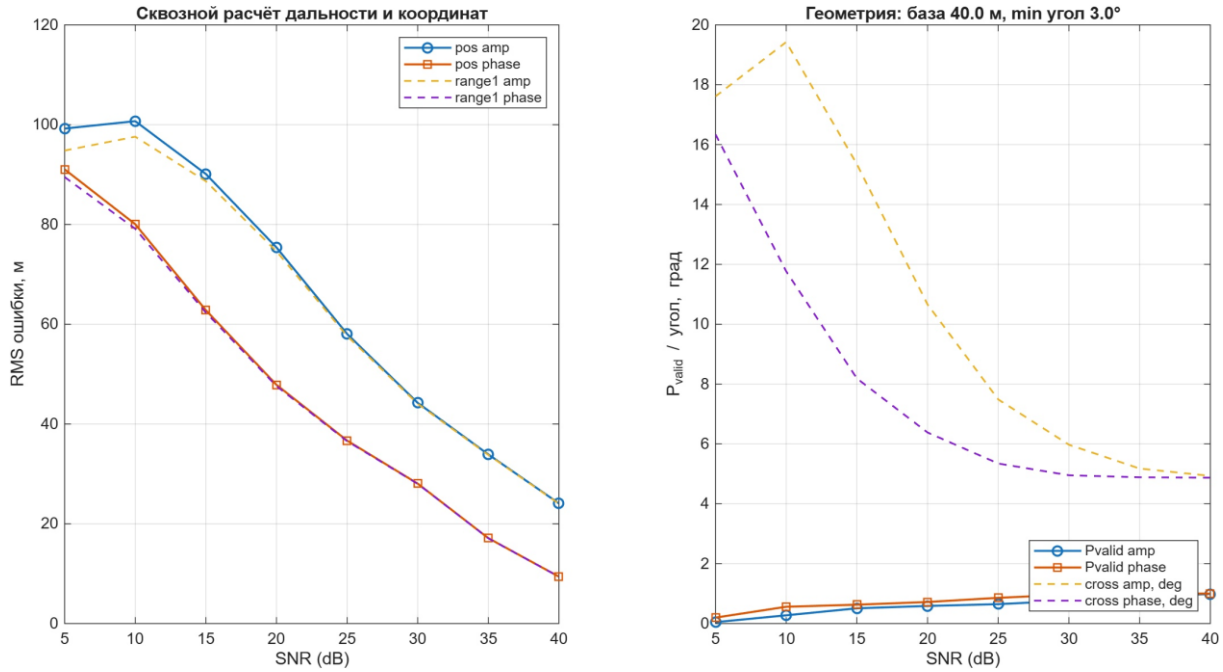


Рисунок 2 - Ошибка координатной оценки и дальности при использовании двух угломерных систем

Наиболее заметное снижение ошибки получено при применении подпространственного оценивания MUSIC. Блок MUSIC сравнивался с исходной фазовой оценкой, а также с вариантами после медианной фильтрации. Медианная фильтрация использовалась как постобработка последовательности угловых оценок и снижала влияние одиночных выбросов.

Таблица 3 - Координатная ошибка при использовании MUSIC и медианной фильтрации

SNR, дБ	Фазовый метод, м	Фазовый + медиана, м	MUSIC, м	MUSIC + медиана, м
10	81,23	53,80	18,16	9,11
20	47,61	33,71	5,38	2,46
30	—	—	—	0,89
35	—	—	—	0,64
40	—	—	—	0,75

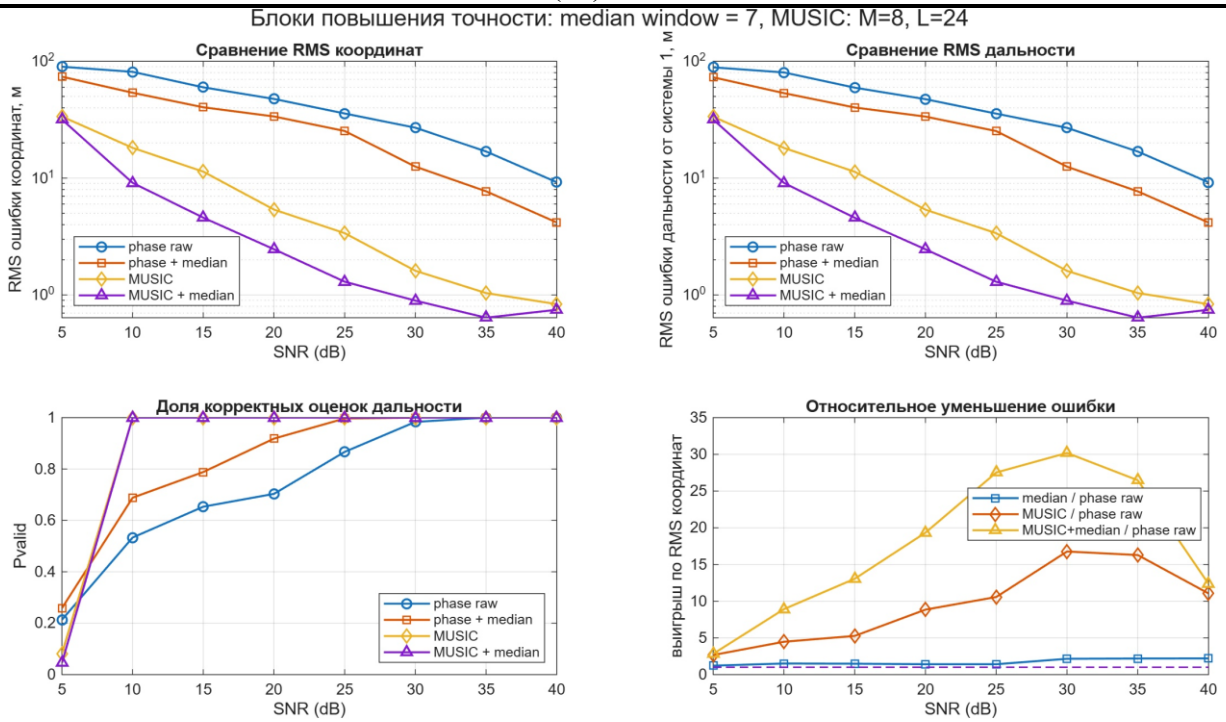


Рисунок 3 - Сравнение фазовой оценки и алгоритма MUSIC до и после медианной фильтрации

При SNR = 10 дБ координатная ошибка фазовой оценки без медианной фильтрации составила 81,23 м, после фильтрации — 53,80 м. Для MUSIC при том же уровне шума ошибка составила 18,16 м без фильтрации и 9,11 м после неё. При SNR = 20 дБ фазовая оценка дала 47,61 м без фильтрации и 33,71 м после фильтрации, тогда как MUSIC дал 5,38 и 2,46 м соответственно. Относительно исходной фазовой оценки применение MUSIC с медианной фильтрацией уменьшило ошибку примерно в 19,3 раза.

При SNR = 30–40 дБ координатная ошибка MUSIC после медианной фильтрации находилась в пределах 0,64–0,89 м, а доля корректных оценок для MUSIC достигала единицы. Это позволяет рассматривать данный блок как основной вариант повышения точности в моделируемой системе при достаточном отношении сигнал/шум. В области 5 дБ результаты требуют осторожной интерпретации, поскольку шумовые и пороговые ограничения начинают сильнее влиять на возможность формирования корректной оценки.

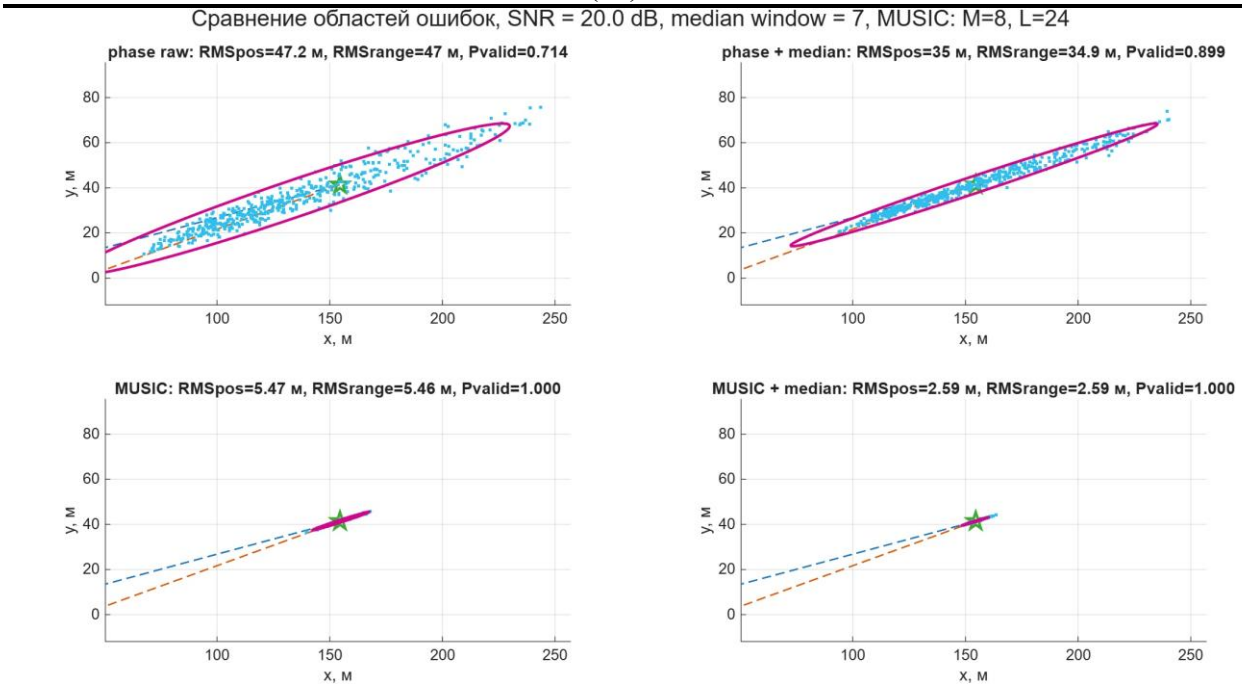


Рисунок 4 - Область координатных ошибок для фазовой оценки и алгоритма MUSIC

Заключение

Полученные результаты показывают, что сравнение только амплитудного и фазового методов недостаточно для оценки качества всей системы: при переходе к двум разнесённым угломерным пунктам координатная ошибка зависит от статистики обеих угловых оценок и геометрии пересечения направлений. Поэтому в программной модели целесообразно разделять локальные метрики азимутального блока и итоговые координатные метрики.

Разработанная программная структура объединяет амплитудный и фазовый методы оценки азимута, геометрический расчёт дальности по двум угломерным пунктам, медианную фильтрацию и MUSIC-оценивание направления прихода сигнала. Фазовый метод показал меньшую среднеквадратическую ошибку оценки направления по сравнению с амплитудным методом: при SNR = 20 дБ ошибка составила 6,06 м против 16,10 м, а при SNR = 40 дБ — 0,58 м против 1,53 м. При переходе к координатной задаче фазовая схема также сохранила преимущество: при SNR = 40 дБ координатная ошибка составила 9,40 м против 24,08 м для амплитудной схемы.

Медианная фильтрация уменьшила влияние единичных выбросов, а MUSIC-алгоритм за счёт использования структуры антенной решётки и ковариационной матрицы сигналов сформировал более узкую область координатных ошибок. При SNR = 20 дБ координатная ошибка снизилась с 47,61 м для исходной фазовой оценки до 2,46 м для MUSIC после медианной фильтрации. Следовательно, совместное применение подпространственного оценивания и фильтрации выбросов может рассматриваться как перспективное направление повышения точности моноимпульсных угломерных систем.

Ограничением выполненного моделирования является использование идеализированной двумерной постановки и гауссовой модели диаграммы направленности. В дальнейшем модель может быть расширена за счёт учёта калибровочных ошибок, взаимного влияния элементов

антенной решётки, многолучевого распространения и динамики движения источника излучения.

Список литературы

1. Ричардс М. А. Основы обработки радиолокационных сигналов. 3-е изд. Нью-Йорк : McGraw-Hill Education, 2023. ISBN 978-1-260-46871-7.
2. Школьник М. И. Справочник по радиолокации. 3-е изд. Нью-Йорк : Макгроу-Хилл, 2008. 1328 с. ISBN 978-0-07-148547-0.
3. Махафза Б. Р. Анализ и проектирование радиолокационных систем с использованием MATLAB. 4-е изд. Бока-Ратон : Chapman & Hall/CRC, 2022. 690 с. ISBN 978-0-367-50793-8.
4. Шмидт Р. О. Определение местоположения нескольких излучателей и оценка параметров сигнала // IEEE Transactions on Antennas and Propagation. 1986. Том Ii. АП-34, № 3. С. 276-280.
5. Стойка П., Мозес Р. Спектральный анализ сигналов. Аппер-Сэддл-Ривер : Прентис-Холл, 2005. 452 с. ISBN 978-0-13-113956-5.
6. Ван Трипс Х. Л. Оптимальная обработка массива данных: часть IV теории обнаружения, оценки и модуляции. Нью-Йорк : John Wiley & Sons, 2002. 1472 с. ISBN 978-0-471-09390-9.

References

1. Richards M. A. Fundamentals of Radar Signal Processing. 3rd ed. New York : McGraw-Hill Education, 2023. ISBN 978-1-260-46871-7.
2. Skolnik M. I. Radar Handbook. 3rd ed. New York : McGraw-Hill, 2008. 1328 p. ISBN 978-0-07-148547-0.
3. Mahafza B. R. Radar Systems Analysis and Design Using MATLAB. 4th ed. Boca Raton : Chapman & Hall/CRC, 2022. 690 p. ISBN 978-0-367-50793-8.
4. Schmidt R. O. Multiple Emitter Location and Signal Parameter Estimation // IEEE Transactions on Antennas and Propagation. 1986. Vol. AP-34, no. 3. P. 276–280.
5. Stoica P., Moses R. Spectral Analysis of Signals. Upper Saddle River : Prentice Hall, 2005. 452 p. ISBN 978-0-13-113956-5.
6. Van Trees H. L. Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory. New York : John Wiley & Sons, 2002. 1472 p. ISBN 978-0-471-09390-9.