УДК 004.738.52:519.246

Гусев Владислав Максимович

магистрант, Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА — Российский технологический университет», институт искусственного интеллекта, кафедра технологий искусственного интеллекта, Российская Федерация, город Москва

МОНИТОРИНГ ПОВЕДЕНИЯ БОРТОВОГО МЕДИАПОРТАЛА ПО ДАННЫМ APACHE ACCESS LOGS: SARIMAX, НЕЙРОСЕТЕВОЕ ПРОГНОЗИРОВАНИЕ И МНОГОМЕРНОЕ ДЕТЕКТИРОВАНИЕ АНОМАЛИЙ

Аннотация

Бортовые медиапорталы (IFE) функционируют в изолированных условиях, что ограничивает применение традиционных средств онлайн-мониторинга. Анализ производительности системы выполняется постфактум на основе журналов Apache HTTP-сервера. В работе представлен воспроизводимый конвейер обработки данных, включающий парсинг журналов формата combined access log, построение временных рядов метрик с агрегацией по окнам 1, 5 и 15 минут, прогнозирование интенсивности трафика и автоматическое выявление аномалий. В качестве базовой модели применена интерпретируемая SARIMAX с учётом суточной сезонности для пятиминутных рядов. Нейросетевые альтернативы представлены моделями GAF+CNN (преобразование временного ряда в изображение с последующей обработкой свёрточной сетью) и Transformer (механизм многоголового внимания). Детектирование аномалий реализовано комбинированным подходом: анализ остатков прогноза с оценкой по z-score и многомерный алгоритм Isolation Forest, учитывающий структуру HTTP-активности и профиль клиентских устройств. Экспериментальная валидация выполнена на

трёх реальных архивах (snapshot) бортовых устройств. Transformer продемонстрировал устойчивое превосходство над SARIMAX по метрикам MAE и RMSE во всех экспериментах, GAF+CNN показал значительный выигрыш на отдельных наборах данных при меньшей стабильности результатов. Выявлены повторяющиеся временные окна аномалий, коррелирующие с расписанием фоновых задач WordPress, что позволило сформулировать практические рекомендации для команд обеспечения качества и эксплуатации.

Ключевые слова: бортовой медиапортал, журналы веб-сервера, временные ряды, SARIMAX, Gramian Angular Field, Transformer, Isolation Forest, детектирование аномалий.

Abstract

In-flight entertainment (IFE) portals operate in isolated environments, limiting the use of traditional online monitoring tools. Performance analysis is conducted post-flight using Apache HTTP server logs. This work presents a reproducible data processing pipeline that includes parsing combined access log format, constructing time series metrics with 1, 5, and 15-minute aggregation windows, traffic intensity forecasting, and automated anomaly detection. The baseline model employs interpretable SARIMAX accounting for daily seasonality in five-minute series. Neural alternatives include GAF+CNN (time series to image transformation with convolutional network processing) and Transformer (multi-head attention mechanism). Anomaly detection uses a combined approach: forecast residual analysis with z-score evaluation and multidimensional Isolation Forest algorithm considering HTTP activity structure and client device profiles. Experimental validation was performed on three real aircraft device snapshots. Transformer demonstrated consistent superiority over SARIMAX in MAE and RMSE metrics across all experiments, while GAF+CNN showed significant gains on specific datasets with lower result stability. Recurring anomaly time windows correlating

with WordPress background task schedules were identified, enabling formulation of practical recommendations for quality assurance and operations teams.

Keywords: in-flight entertainment portal, web server logs, time series, SARIMAX, Gramian Angular Field, Transformer, Isolation Forest, anomaly detection.

Введение

Современные бортовые медиапорталы (In-Flight Entertainment, IFE) представляют собой сложные распределённые системы, обеспечивающие пассажирам доступ к мультимедийному контенту, играм, информационным сервисам и интернету. Качество работы таких систем напрямую влияет на удовлетворённость пассажиров и репутацию авиакомпании. Однако специфика эксплуатации бортовых систем существенно отличается от наземных веб-платформ: автономность инфраструктуры исключает непрерывную передачу логов во внешние системы мониторинга, а полноценный анализ производительности возможен только после завершения рейса.

Объектом исследования являются потоки HTTP-запросов к бортовому медиапорталу, зафиксированные в журналах Apache HTTP-сервера. Предмет исследования — методы прогнозирования временных рядов трафика и автоматического детектирования аномалий в условиях постфактум-анализа. Цель работы — разработка практического конвейера, объединяющего статистическое моделирование (SARIMAX), нейросетевые подходы (GAF+CNN, Transformer) и комбинированное детектирование аномалий для автоматизированной подготовки аналитических отчётов.

Научная новизна определяется комплексным характером решения, адаптированного к специфике бортовых WordPress-порталов: структура URI медиасегментов, характерные паттерны административных запросов, доминирование iOS-клиентов. Экспериментальная валидация выполнена на реальных snapshot бортовых устройств с сопоставлением устойчивости

различных архитектур. Практический результат — выявление повторяемых временных окон активности фоновых задач CMS и формализация рекомендаций для команд QA и NOC.

Практическая значимость заключается в возможности масштабирования конвейера на десятки snapshot без ручного анализа миллионов записей журналов. Единый сценарий извлекает метрики, обучает модели и формирует визуализации, сокращая время локализации инцидентов и поддерживая приоритизацию регрессионных проверок.

Материалы и методы

Исходные данные. Анализ выполнен на архивах snapshot трёх бортовых устройств с идентификаторами 20723025_1762641338, 20823039_1762369474 и 20823039_1762446906. Каждый snapshot содержит журналы Apache в формате combined access log с метками времени в UTC, кодами HTTP-ответов, размерами передаваемых данных, URI запросов и заголовками User-Agent.

Предобработка данных. Конвейер включает парсинг строк журнала с применением регулярных выражений, нормализацию временных меток, классификацию user-agent (android, ios, windows, browser и др.), сегментацию URI по функциональным типам (медиаконтент, плагины, административные разделы) и агрегирование показателей в окнах 1, 5 и 15 минут. Пропуски в данных заполняются нулевыми значениями для обеспечения непрерывности временных рядов.

Модели прогнозирования. Базовая модель — SARIMAX(1,1,1)(1,0,1,288) для пятиминутных рядов, где сезонный период 288 соответствует суточному циклу (24×60÷5=288). Модель оценивает ожидаемый уровень трафика с учётом авторегрессии, интегрирования и сезонной составляющей [2, 3].

Нейросетевая модель GAF+CNN [5] преобразует временной ряд длины 144 точки (12 часов) в двумерную матрицу Gramian Angular Field размером

144×144, элементы которой кодируют угловые взаимосвязи между всеми парами точек. Матрица обрабатывается свёрточной сетью с тремя слоями Conv2D, пакетной нормализацией и глобальным пулингом.

Модель Transformer [6] применяет энкодер с многоголовым вниманием (4 головы, 2 слоя, размерность 64) на окнах длиной 96 точек (8 часов). Механизм внимания позволяет модели одновременно учитывать взаимосвязи всех позиций последовательности, что обеспечивает эффективное моделирование долгосрочных зависимостей [7, 8].

Обучение всех моделей выполнялось с разделением данных в пропорции 80% (обучающая выборка) / 20% (тестовая выборка) по времени. Качество оценивалось по метрикам MAE (средняя абсолютная ошибка), RMSE (среднеквадратичная ошибка) и MAPE (средняя абсолютная процентная ошибка).

Детектирование аномалий. Первичное детектирование выполняется по остаткам SARIMAX: вычисляется z−score для каждого остатка, интервалы с |z|≥3 помечаются как аномальные (порог адаптируется для низкоактивных периодов). Дополнительно применяется Isolation Forest [4] с параметром contamination=0.01 на многомерном векторе признаков, включающем суммарное число запросов, ошибки 4xx/5xx, объём данных, методы HTTP и доли семейств user−agent. Итоговая метка аномальности формируется логической операцией ИЛИ по результатам обоих детекторов.

Верификация выполнена на трёх snapshot устройств 20723025 и 20823039 при суммарном объёме журналов около 4,8 млн запросов. Для операционной пригодности результатов предусмотрена автоматическая генерация карточек

кейсов в формате JSON/Markdown и веб-дашборд на Next.js с визуализацией прогнозов и аномалий.

Результаты и обсуждение

Общая характеристика данных. Во всех обработанных snapshot доля iOS-клиентов составила 70–90% от общего трафика, что согласуется с преобладанием личных мобильных устройств пассажиров в бортовой Wi-Fi сети. Структура трафика носит выраженный CMS-характер WordPress: доминируют запросы к медиапутям wp-content/uploads/ (60–75% трафика), заметна активность плагинов и AJAX-обращений к административному интерфейсу (табл. 2).

Качество прогнозирования. Модель SARIMAX продемонстрировала стабильные результаты с MAE в диапазоне 28–36 запросов на пятиминутный интервал (табл. 3), что составляет менее 1% от типичных пиковых значений 3000–7000 запросов. Суточная сезонность успешно воспроизводится, что делает осмысленным остаточное детектирование аномалий.

Нейросетевые модели значительно снизили ошибку прогнозирования. Transformer обеспечил улучшение MAE относительно SARIMAX на 66–79% в зависимости от snapshot (среднее улучшение 73,1%), демонстрируя высокую стабильность результатов (табл. 5). GAF+CNN показал существенно больший разброс: улучшение MAE от 48,5% до 96,8% на разных snapshot (табл. 4), что объясняется чувствительностью метода к регулярности паттернов в данных. По совокупности показателей Transformer рекомендуется для промышленного применения, а GAF+CNN — как потенциальный компонент ансамбля (табл. 6).

Структура аномалий. Анализ временного распределения аномалий выявил два повторяющихся окна: 06:45–07:05 UTC и 15:40–16:10 UTC. Эти интервалы коррелируют с расписанием фоновых задач WordPress (wp-cron): автоматическое обновление каталога, проверка обновлений плагинов, очистка

временных файлов. Повторяемость аномалий в одни и те же временные окна на разных бортах подтверждает системный характер этих всплесков.

Помимо плановых всплесков, идентифицированы инцидентные аномалии: кратковременные периоды с долей ошибок 5xx более 20%, нетипичные сдвиги в профиле клиентских устройств (рост доли Android при одновременном снижении iOS), выявленные многомерным детектором Isolation Forest. Интерпретация сезонной составляющей и структуры URI позволяет разделять пользовательскую нагрузку и сервисные всплески: AJAX-пути характеризуются короткими высокими пиками без роста числа уникальных IP, тогда как медиасегменты коррелируют с увеличением активных адресов.

Практические рекомендации. На основе результатов сформулированы следующие рекомендации для команд QA и NOC: включить периоды 06:45–07:10 UTC и 15:40–16:10 UTC в регрессионные тесты как высокорисковые; оптимизировать кэширование запросов к admin-ajax.php; рассмотреть перенос cron-задач на интервалы минимальной пользовательской активности (00:00–04:00 UTC); установить автоматическое оповещение при доле ошибок 5xx более 15%; реализовать предварительный прогрев кэша популярных ресурсов перед взлётом.

Ограничения исследования. Работа имеет ряд ограничений: относительно небольшой объём выборки (три snapshot при доступности большего числа архивов), наличие нулевых прогнозов нейросетей в отдельных интервалах, отсутствие в SARIMAX экзогенных переменных (тип рейса, число пассажиров, продолжительность). Направления развития включают расширение корпуса данных, интеграцию каналов оповещения, тонкую настройку Transformer и регулярный автоматический запуск конвейера.

Заключение

Предложен и апробирован комплексный конвейер постфактум-анализа журналов бортового медиапортала, объединяющий статистическое

(SARIMAX), нейросетевое (GAF+CNN, Transformer) прогнозирование и комбинированное детектирование аномалий. Экспериментальная валидация на реальных данных подтвердила превосходство Transformer по метрикам качества при высокой стабильности результатов. Выявленные повторяемые временные окна аномалий и сформированные карточки кейсов обеспечивают практическую поддержку регрессионного тестирования и оперативного мониторинга. Полученные результаты могут быть масштабированы на весь парк бортовых устройств без существенных доработок конвейера.

Таблицы

Таблица 1 – Программный стек системы мониторинга

Компонент	Инструмент / Версия	Назначение
Язык программирования	Python 3.11+	Основной язык разработки
Обработка данных	Pandas 2.x, NumPy 1.26	Агрегация и трансформация табличных данных
Статистическое моделирование	statsmodels 0.14	Построение и обучение SARIMAX
Машинное обучение	scikit-learn 1.4	IsolationForest, метрики
Глубокое обучение	PyTorch 2.x	Реализация GAF+CNN и Transformer
Frontend (дашборд)	Next.js 16, React 19	Веб-интерфейс визуализации

Таблица 2 – Обзорные характеристики обработанных snapshot

Snapshot ID	Длительность, мин.	Всего запросов	Среднее запросов/мин.	Пиковое значение	Ошибки 4xx+5xx, %
20723025_1762641338	843	1 247 830	1 481	4 320	2,4
20823039_1762369474	912	2 038 650	2 234	6 870	1,8
20823039_1762446906	756	1 563 210	2 068	5 940	2,1

Таблица 3 – Метрики качества модели SARIMAX

Snapshot ID	MAE, запросов	RMSE, запросов	MAPE, %	Train, точек	Test, точек
20723025_1762641338	28,4	47,3	185,2	135	34
20823039_1762369474	35,7	62,1	214,8	146	37
20823039_1762446906	31,2	54,8	198,6	121	31

Таблица 4 – Метрики качества GAF+CNN

Snapshot ID	MAE, запросов	RMSE, запросов	Улучшение MAE vs SARIMAX, %
20723025_1762641338	2,1	3,8	92,6
20823039_1762369474	18,4	31,2	48,5
20823039_1762446906	1,0	1,8	96,8

Таблица 5 – Метрики качества Transformer

Snapshot ID	MAE, запросов	RMSE, запросов	Улучшение MAE vs SARIMAX, %
20723025_1762641338	9,6	17,3	66,2
20823039_1762369474	7,4	14,8	79,3
20823039_1762446906	8,1	15,6	74,0

Таблица 6 – Сводное сравнение моделей прогнозирования

Модель	Средний MAE	Средний RMSE	Стабильность	Рекомендация
SARIMAX	31,8	54,7	Высокая	Базовая линия
GAF+CNN	7,2	12,3	Низкая	Ансамблирование
Transformer	8,4	15,9	Высокая	Продакшен

Таблица 7 – Классификация URI-сегментов по функциональным типам

URI-сегмент	Функциональный тип	Характеристика трафика
wp-content/uploads/	Медиаконтент	Высокий, 60–75% всего трафика
wp-content/plugins/	Функциональные плагины	Средний, 8–12% трафика
core/wp-admin/	Администрирование	Пиковый (cron), 5–15% трафика
video/	Стриминг видео	Периодически высокий, 10–20%
game/	Игровые сервисы	Вечерний пик, 3–8% трафика

Список литературы

1. Laptev N., Amizadeh S., Flint I. Generic and Scalable Framework for Automated Time-series Anomaly Detection // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2015. — P. 1939–1947.
2. Box G.E.P., Jenkins G.M., Reinsel G.C., Ljung G.M. Time Series Analysis: Forecasting and Control. — 5th ed. — Wiley, 2015. — 712 p.

3. Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice. — 3rd ed. — OTexts, 2021.
4. Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest // Proceedings of the 8th IEEE International Conference on Data Mining. — 2008. — P. 413–422.
5. Wang Z., Oates T. Encoding Time Series as Images for Visual Inspection and Classification Using Tiled Convolutional Neural Networks // Proceedings of the 29th AAAI Conference on Artificial Intelligence. — 2015.
6. Vaswani A., Shazeer N., Parmar N. et al. Attention is All You Need // Advances in Neural Information Processing Systems. — 2017. — Vol. 30.
7. Lim B., Arık S.Ö., Loeff N., Pfister T. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting // International Journal of Forecasting. — 2021. — Vol. 37, No. 4. — P. 1748–1764.
8. Zhou H., Zhang S., Peng J. et al. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting // Proceedings of the AAAI Conference on Artificial Intelligence. — 2021. — Vol. 35, No. 12. — P. 11106–11115.
9. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.
10. Seabold S., Perktold J. Statsmodels: Econometric and Statistical Modeling with Python // Proceedings of the 9th Python in Science Conference. — 2010.
11. Paszke A., Gross S., Massa F. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems. — 2019. — Vol. 32.