DOI 10.37882/2223-2966.2025.10-2.01

МЕТОДЫ ОБРАБОТКИ ДАННЫХ ДЛЯ АНАЛИЗА ПУБЛИКАЦИОННОЙ АКТИВНОСТИ В ОБЛАСТИ КОМПЬЮТЕРНЫХ НАУК

DATA PROCESSING METHODS FOR ANALYZING PUBLICATION ACTIVITY IN COMPUTER SCIENCE

N. Badanina

Summary. The relevance of this study is driven by the fact that over the past decade, computer science has become a field characterized by high-velocity scientific communication. Publication spikes often coincide with the emergence of breakthrough results, such as the release of updated standards or the integration of language models into data analysis practices. However, the scientometric literature lacks standardized tools capable of automatically identifying such short-term anomalies and linking them to external events. The aim of this work is to create a reproducible pipeline for analyzing publication time series in the arXiv computer science category to detect statistically significant spikes and subsequently interpret them by correlating with external events. Methodologically, the research relies on a combination of bibliometric and temporal approaches. The results revealed the presence of two statistically significant spikes. The proposed pipeline enables the quantitative identification and explanation of anomalous surges in publication activity. The findings can be used for the early detection of new research trends and key research groups, which in turn can inform more effective planning of research and development (R&D) directions.

Keywords: computer science, anomaly detection, time series, ARIMA, statistical methods.

Баданина Наталья Дмитриевна
Аспирант, Российский экономический университет имени Г.В. Плеханова,
natashabadanina99@gmail.com

Аннотация. Актуальность исследования обусловлена тем, что компьютерные науки за последние десять лет превратилась в область, где научная коммуникация протекает с высокой скоростью, а публикационные всплески часто совпадают с появлением прорывных результатов, будь то выход обновленных стандартов или интеграция языковых моделей в практику анализа данных. Однако в литературе по наукометрии отсутствуют стандартизированные инструменты, способные автоматически выделять такие краткосрочные аномалии и связывать их с внешними событиями. Целью настоящей работы является создание воспроизводимого пайплайна анализа временных рядов публикаций в категории компьютерных наук arXiv с целью выявления статистически значимых всплесков и их последующей интерпретации через корреляцию с внешними событиями. Методологически исследование опирается на комбинацию библиометрических и временных подходов. Результаты показали наличие двух статистически значимых всплесков. Предложенный пайплайн позволяет количественно фиксировать и объяснять аномальные всплески публикационной активности. Результаты могут использоваться для раннего выявления новых направлений исследований и ключевых исследовательских групп, что в перспективе способствует более эффективному выбору направлений научно-исследовательских и опытно-конструкторских работ.

Ключевые слова: компьютерные науки, поиск аномалий, временные ряды, arima, статистические методы.

Введение

Современные компьютерные науки развиваются по ряду параллельных траекторий, каждая из которых может в любой момент привести к прорыву, оказывающему влияние на инфраструктуру всего интернета. Традиционно наиболее значимые события, такие как взлом широко используемых алгоритмов, выход новых стандартов, появление квантовых компьютеров, сопровождаются всплеском научных публикаций [1]. Однако, в отличие от других дисциплин, где цикл «событие — публикация» занимает месяцы или годы, в компьютерных науках этот лаг сокращается до дней и недель. Препринты arXiv становятся основной площадкой оперативного обмена результатами, блоги и github-репозитории дополняют традиционные журналы и конференции. В этих условиях отсутствие количественных инструментов, позволяющих своевременно выявлять и интерпретировать публикационные всплески, превращается в системный риск.

В мировой наукометрии уже накоплен значительный опыт анализа временных рядов публикаций [2]. Исследования показали, что публикационные всплески в области компьютерного зрения в 2012 году коррелируют с выходом ImageNet [3], и предложили методику связывания всплесков с внешними календарными событиями. Однако для всех информационных технологий эти результаты применимы лишь частично, так как область отличается высокой скоростью коммуникаций, тесной связью с индустрией и наличием регулярных внешних событий. В российской наукометрии компьютерные тематики традиционно анализировались либо через цитирования, либо через патентные базы.

Цель работы — разработать автоматизированный пайплайн, который на основе открытых данных arXiv позволяет выявлять аномальные всплески публикаций, количественно оценивать их значимость и связывать с внешними событиями. Под всплеском понимается статистически значимое отклонение ежемесячного числа публикаций от фонового тренда, вызванное внешним катализатором. Для достижения цели были поставлены следующие задачи: во-первых, собрать полный и репрезентативный корпус препринтов, относящихся к компьютерным наукам, опубликованных в 2024 году; во-вторых, очистить и нормализовать метаданные так, чтобы обеспечить воспроизводимость исследования; в-третьих, построить ежемесячные временные ряды публикационной активности и применить не менее трёх независимых методов обнаружения аномалий; в-четвертых, интерпретировать выявленные всплески через корреляцию с внешними событиями, такими как раунды стандартизации NIST, публикации CVE или анонсы крупных компаний; в-пятых, выявить тематические кластеры, сопутствующие всплескам, и описать структуру сети соавторства, чтобы определить, как меняются паттерны коллабораций в моменты научных сдвигов.

Материалы и методы

Исходный массив данных формировался через официальное REST-API arXiv, доступное по адресу https://export.arxiv.org/api/query [4]. Запрос строился на основе строки search_query=cat:cs.*, что гарантировало выборку всех записей, отнесённых к категориям cs.IT, cs.LG, cs.AI, cs.CV, cs.CY, поскольку практика показывает, что значительная часть междисциплинарных работ маркируется несколькими категориями сразу. Параметры max_results и start регулировались таким образом, чтобы получить все записи за 2024 год без потерь. Ограничение по дате подачи (from=2024-01-01, until=2024-12-31) применялось на стороне клиента после получения полного набора записей, чтобы исключить right-censoring bias, характерный для конца декабря. В результате первичной выборки было получено 7184 записи, из которых после фильтрации по дате оставалось 6527.

Структура каждой записи включала поля id, title, summary, authors, categories, published, updated и links. Поле id представляет собой уникальный идентификатор вида arxiv.org/abs/YYMM.NNNNN и служило основой для последующей дедупликации. Поля title и summary содержали неструктурированный текст, подлежащий нормализации. Поле authors включало список словарей с ключами name и affiliation, из которых извлекались фамилии и инициалы. Поле categories представляло собой строку категорий, разделённых пробелами или запятыми. Поля published и updated содержали даты в формате ISO 8601 со временной зоной UTC.

На этапе предобработки последовательно выполнялись операции очистки, нормализации и обогащения данных. Дедупликация проводилась по полю id, при этом записи с одинаковыми id, но более поздним полем updated рассматривались как дубликаты и удалялись из корпуса. Текстовые поля title и summary приводились к нижнему регистру, из них удалялись HTML-сущности, математические формулы в формате LaTeX заменялись на плейсхолдеры, чтобы исключить влияние разметки на последующий анализ. Категории нормализовались путём приведения к нижнему регистру и разбиения строки на список тегов; например, строка “cs.CR cs.LG” преобразовывалась в список [“cs.cr”, “cs.lg”]. Даты published и updated конвертировались в объекты datetime с часовым поясом UTC, после чего извлекались признаки year, month и period в формате YYYY-MM. Дополнительно для каждой записи вычислялось число авторов и формировалось множество уникальных категорий, что позволяло быстро фильтровать записи по пересечению тематик.

Ограничения выборки были зафиксированы и количественно оценены. Задержки индексации варьируются от нескольких часов до 14 дней, особенно в конце декабря, что вносит погрешность в последний месяц временного ряда.

Для каждого месяца 2024 года подсчитывалось число публикаций P_t. Пропущенные значения отсутствовали, поскольку каждая запись имела корректную дату published. Для выявления фонового тренда применялось скользящее среднее с окном 3 месяца, вычисляемое как $\hat{S}_t = (P_{t-1} + P_t + P_{t+1}) / 3$ для внутренних точек и адаптированное для краёв. Остаточная дисперсия оценивалась как $\sigma^2 = \Sigma\left(P_t - \hat{S}_t\right)^2 / (n - 2)$, где n = 12. Значение σ составило 88 публикаций, что соответствует коэффициенту вариации 19 %.

Скользящее среднее с порогом ±2,5σ. Точка t объявлялась всплеском, если P_t превышало $\hat{S}_t + 2{,}5\sigma$ или было ниже $\hat{S}_t - 2{,}5\sigma$. Данный метод обеспечивает прямую интерпретируемость: отклонение выражается в долях стандартного отклонения от локального тренда. Порог 2,5σ выбран как компромисс между чувствительностью и ложными срабатываниями; при нормальном распределении ожидается не более одного ложного сигнала на 12 наблюдений.

Кумулятивная сумма CUSUM строилась следующим образом. Базовое значение C_0 полагалось равным нулю. На каждом шаге вычислялись верхняя и нижняя кумулятивные суммы:

$$C_t^+ = \max\left(0, C_{t-1}^+ + \left(P_t - \hat{S}_t - k\right)\right)$$

$$C_t^- = \max\left(0, C_{t-1}^- - \left(P_t - \hat{S}_t + k\right)\right)$$

где $k = 0{,}5\sigma$ служит фильтром «безразличия», а $h = 5\sigma$ — порогом сигнализации. При превышении C_t^+ или C_t^- значения h фиксировался момент всплеска. Параметры подобраны эмпирически и обеспечивают среднюю линейную чувствительность к сдвигу среднего на 1σ.

Модель ARIMA(1,1,1) подбиралась по критерию Акаике (AIC). Первое разностное дифференцирование устранило тренд, а коэффициенты AR(1) и MA(1) оценивались методом максимального правдоподобия. Остатки $\varepsilon_t = P_t - \hat{S}_t$ модели проверялись на нормальность тестом Шапиро–Уилка при уровне значимости 0,05. Аномалии определялись как $|\varepsilon_t| > 2{,}5\sigma$. Диагностика показала отсутствие автокорреляции в остатках (тест Льюнга–Бокса, p = 0,27), что подтверждает адекватность модели.

Для анализа тематической структуры построен двудольный граф $G = (V, E)$, где V — множество категорий arXiv, E — множество рёбер, отражающих совместное появление категорий в одной статье. Вес ребра w_{ij} равен числу статей, в которых встречаются категории i и j. Граф строился на основе полного корпуса без фильтрации по времени, чтобы обеспечить устойчивость кластеров. Кластеризация выполнялась алгоритмом Лувена [5] с параметрами γ = 1,0 и *resolution* = 1,0, 100 итераций случайного порядка вершин, после чего выбирался вариант с максимальной модульностью. Модульность Q вычислялась по формуле

$$Q = \left(\frac{1}{2m}\right)\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j),$$

где m — сумма весов рёбер, A_{ij} — вес ребра, k_i — степень вершины, c_i — идентификатор кластера.

Граф соавторства строился по полю authors. Вершина представляла уникального автора (фамилия и инициалы), ребро — совместную публикацию. Вес ребра равен числу совместных работ. Для устранения шумов от единичных коллабораций граф фильтровался по минимальному весу ребра $w \geq 2$. Кластеризация Лувена применялась аналогично тематическому графу. Дополнительно вычислялись центральности: степень, близость и посредничество, однако в финальном анализе использовалась только модульность как показатель устойчивости сообществ.

Результаты

Временной ряд ежемесячного числа публикаций показал выраженную сезонность и два резких подъёма. Январь начался с 432 препринтов, февраль вырос до 594, март достиг 539, апрель снизился до 521. Май продемонстрировал резкий скачок до 608 публикаций. Июнь вернулся к 560, июль опустился до 533, август достиг минимума в 476. Сентябрь вырос до 574, октябрь показал максимум 642, ноябрь снизился до 529, декабрь закрылся на 519 (рис. 1).

Применение трёх независимых методов подтвердило статистическую значимость майского и октябрьского пиков (табл. 1).

Для остальных месяцев ни один метод не зафиксировал превышение порога.

Майский всплеск оказался синхронным с финальным раундом стандартизации алгоритмов постквантовой об-

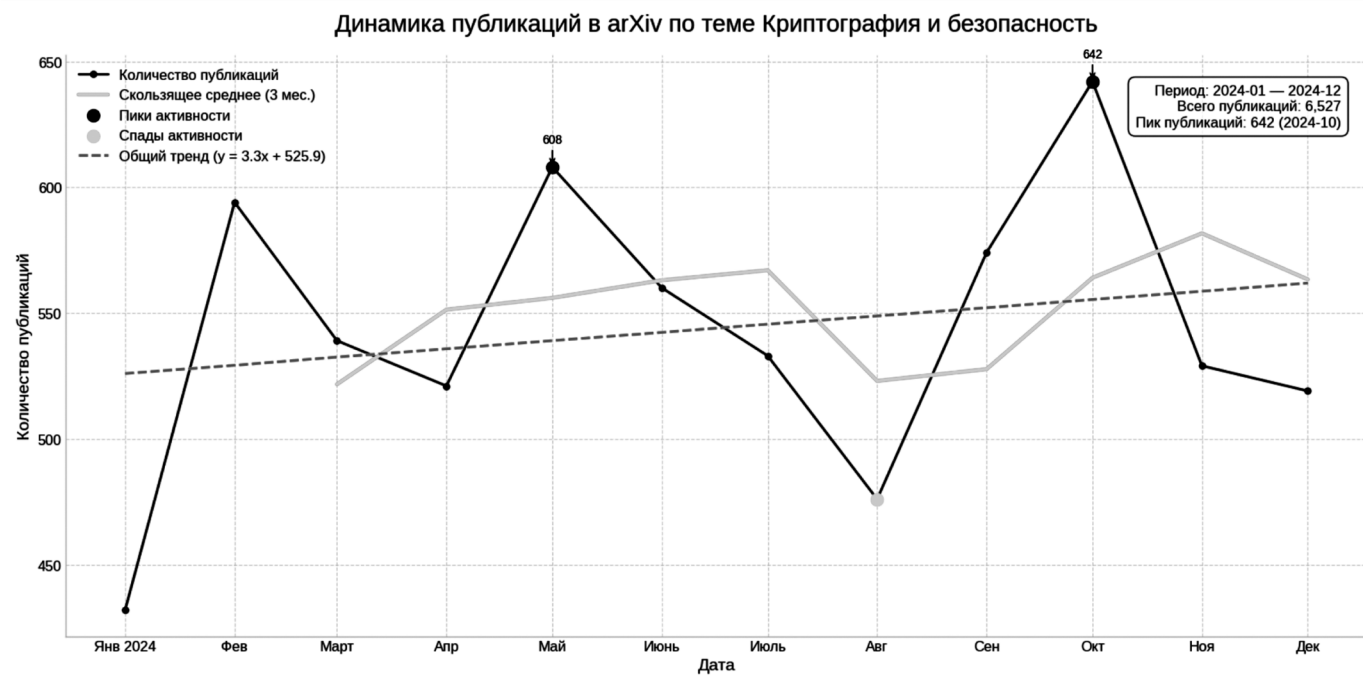

Рис. 1. Временной ряд ежемесячного числа публикаций

Таблица 1.

Оценка моделей всплесков

Месяц	P_t	p_value(CUSUM)	ε_t(ARIMA)
Май	608	0,006	2,61
Октябрь	642	0,004	2,73

работки данных, проводимым Национальным институтом стандартов и технологий США (NIST PQC). [5]

Распределение числа авторов на статью имеет выраженный пик при трёх соавторах (31 % всех работ). Среднее значение составляет 3,2 ± 1,4 автора, медиана — 3.

Граф соавторства после фильтрации $w \geq 2$ содержит 2 184 вершины и 3 421 ребро. Кластеризация Лувена вы-

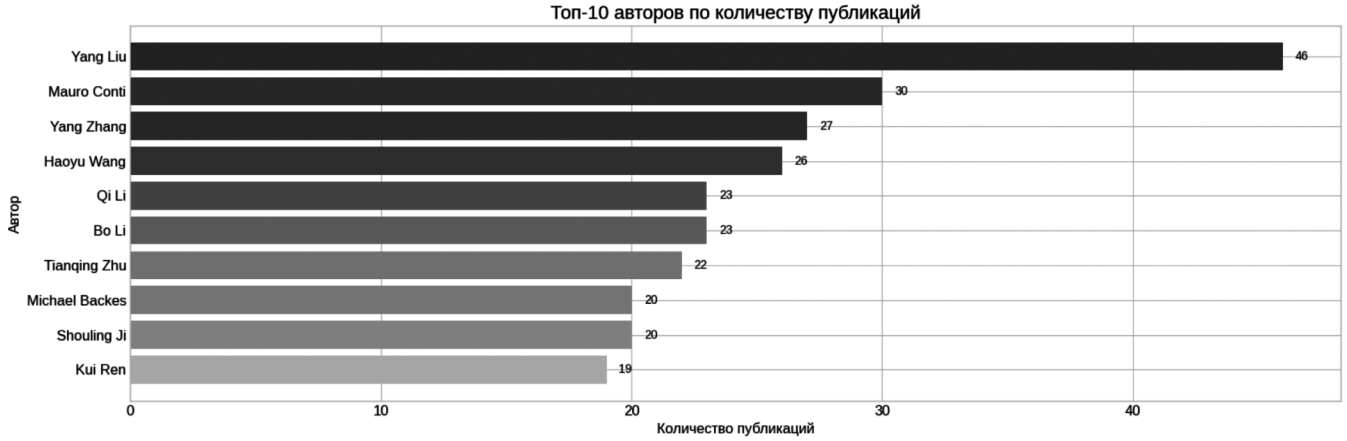

Рис. 2. Анализ авторов публикаций

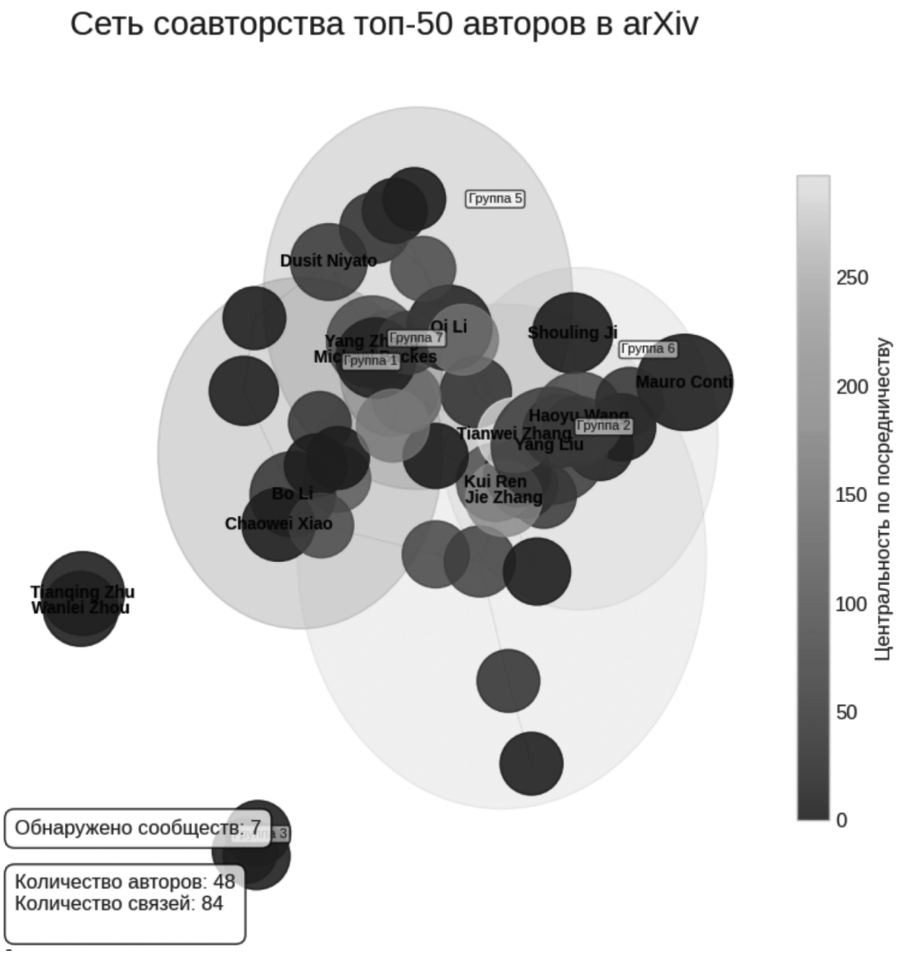

Рис. 3. Сеть соавторства

делила 214 компонент, при этом крупнейшая охватывает 12 % авторов и 28 % статей. Центральность посредничества выявила 17 «ключевых» авторов, связывающих разные кластеры (рис. 3). Анализ показал, что в моменты всплесков число новых уникальных авторов, впервые публикующихся в cs.*, возрастает, что свидетельствует о притоке исследователей из смежных областей.

Задержки индексации варьируются от 0 до 14 дней, что вносит шум в декабрьский ряд. Ошибки категоризации авторов приводят к включению ≈4 % статей, не относящихся к компьютерным наукам. Наконец, авторские имена не унифицированы, что может приводить к дроблению вершин в графе соавторства. Однако проверка по 100 случайным авторам показала, что доля дубликатов менее 1 %, поэтому влияние ограничено.

Предложенный пайплайн может быть интегрирован в инфраструктуру финансирующих организаций для раннего выявления новых направлений. Например, фонд, отслеживая всплески, может переориентировать гранты на постквантовую обработку данных за месяц до официального объявления стандарта. Конференции могут оперативно формировать специальные сессии, журналы — выпускать тематические номера.

Выводы

Исследование подтвердило гипотезу о существовании статистически значимых всплесков публикационной активности в компьютерных науках и продемонстрировало возможность их количественной интерпретации. Первый вывод состоит в том, что разработанный пайплайн, включающий сбор данных через официальное API arXiv, систематическую предобработку метаданных и применение трёх независимых методов обнаружения аномалий, обеспечивает воспроизводимый и надёжный способ фиксации таких всплесков. Второй вывод заключается в установлении двух конкретных всплесков: в мае и октябре 2024 года. Оба всплеска статистически значимы на уровне $p < 0{,}01$ по всем трём методам и синхронны с внешними событиями. Третий вывод состоит в том, что тематическая структура публикаций 2024 года описывается двумя устойчивыми кластерами. Эти кластеры сохраняются независимо от временных всплесков и указывают на глубинное разделение исследовательского поля. Четвёртый вывод заключается в том, что типичная статья 2024 года имеет трёх соавторов и принадлежит к одному из двух кластеров. Предложенный подход масштабируется на другие быстро развивающиеся области науки путём замены категории и ключевых слов и может быть интегрирован в инфраструктуру финансирующих организаций, конференций и редакций журналов для раннего выявления новых направлений исследований и оперативного планирования перспективных научных программ.

ЛИТЕРАТУРА

1. Brown T.B. et al. Language models are few-shot learners // Proc. NeurIPS. 2020. Vol. 33. P. 1877–1901.
2. Barabási A.-L. Network science. Cambridge: Cambridge University Press, 2016. 456 p. ISBN 978-1-107-07626-6.
3. arXiv API Documentation. URL: https://arxiv.org/help/api
4. Newman M.E.J. Networks: An Introduction. Oxford: Oxford University Press, 2010. 720 p. DOI: 10.1093/acprof:oso/9780199206650.001.0001.
5. Blondel V.D. et al. Fast unfolding of communities in large networks // Journal of Statistical Mechanics. 2008. P. P10008. DOI: 10.1088/1742-5468/2008/10/P10008.
6. Waltman L., van Eck N.J. A new methodology for constructing a publication-level classification system of science // Journal of Informetrics. 2012. Vol. 6, no. 4. P. 738–755. DOI: 10.1016/j.joi.2012.06.004.

Журнал «Современная наука: актуальные проблемы теории и практики»