

ПРОСТРАНСТВЕННО-ВРЕМЕННОЕ ОБНАРУЖЕНИЕ СОБЫТИЙ ПУТЕМ ИНТЕГРАЦИИ ТОПОЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ, ИНФОРМАЦИОННОЙ ДИНАМИКИ И СТАБИЛЬНОСТИ СИСТЕМЫ

SPATIO-TEMPORAL EVENT DETECTION BY INTEGRATING TOPOLOGICAL DATA ANALYSIS, INFORMATION DYNAMICS, AND SYSTEM STABILITY

**Mohammad Hani
V. Pak**

Summary. This paper introduces a new framework that combines dynamical systems theory, topological data analysis (TDA), and information theory to detect and analyze events in complex spatio-temporal datasets.

By combining these metrics, the framework provides a robust, interpretable, and scalable solution for event detection.

We demonstrate its application on the Milan 2013 dataset, which includes cellular activity, internet usage, and geolocalized Twitter data, demonstrating its ability to detect events such as social gatherings.

Key innovations include: A unified approach combining robustness, topology, information theory, and Mahalanobis distance.

Keywords: spatiotemporal, loss of stability, topological destruction, information reconfiguration, Mahalanobis distance.

Мохаммад Хани

Аспирант, Санкт-Петербургский Политехнический
Университет Петра Великого
mohammad.h@edu.spbstu.ru

Пак Вадим Геннадьевич

кандидат физико-математических наук, доцент,
Санкт-Петербургский Политехнический
Университет Петра Великого
pak_vg@spbstu.ru

Аннотация. Эта статья представляет новую структуру, которая объединяет теорию динамических систем, топологический анализ данных (TDA) и теория информации для обнаружения и анализа событий в сложных пространственно-временных наборах данных.

Комбинируя эти показатели, структура обеспечивает надежное, интерпретируемое и масштабируемое решение для обнаружения событий.

- Мы демонстрируем его применение на базе данных Milan 2013, которая включает в себя активность сотовой связи, использование интернета и геолокализованные данные Twitter, демонстрируя свою способность обнаруживать такие события, как социальные собрания.
- Основные новшества включают: Единый подход, объединяющий устойчивость, топологию, теории информации и расстояние Махаланобиса.

Ключевые слова: пространственно-временное, потеря стабильности, топологическое разрушение, информационная реконфигурация, расстояние Махаланобиса.

Введение

Пространственно-временные наборы данных, которые кодируют феномены, развивающиеся как в пространстве, так и во времени, представляют уникальные проблемы для обнаружения событий. Традиционные методы часто опираются на пороговое обнаружение аномалий [1] или контролируемое обучение с трудом улавливают многомасштабный [2], нелинейный и возникающий характер событий в сложных системах. Эта статья представляет единую теоретическую основу, которая объединяет три математических парадигмы:

1. **Топологический анализ данных (TDA)** для количественной оценки структурной эволюции [3].
2. **Информационная теория** для измерения сдвигов координации и распределения [4].
3. **Теория динамических систем** для оценки стабильности и критических переходов.

Мы разработали эту структуру и применили ее для анализа набора данных по Милану за 2013 год, который фиксирует динамику города посредством использования мобильных телефонов, использования Интернета и геопространственных данных сообщений, отправленных в Twitter.

Литературный обзор

1. Топологический анализ данных при обнаружении событий

Топологический анализ данных — это мощная область, предоставляющая математические и статистические инструменты для раскрытия «формы» данных. [8] В отличие от традиционных статистических методов, фокусирующихся на численных сводках, TDA исследует базовую структуру и связность точек данных. Для вре-

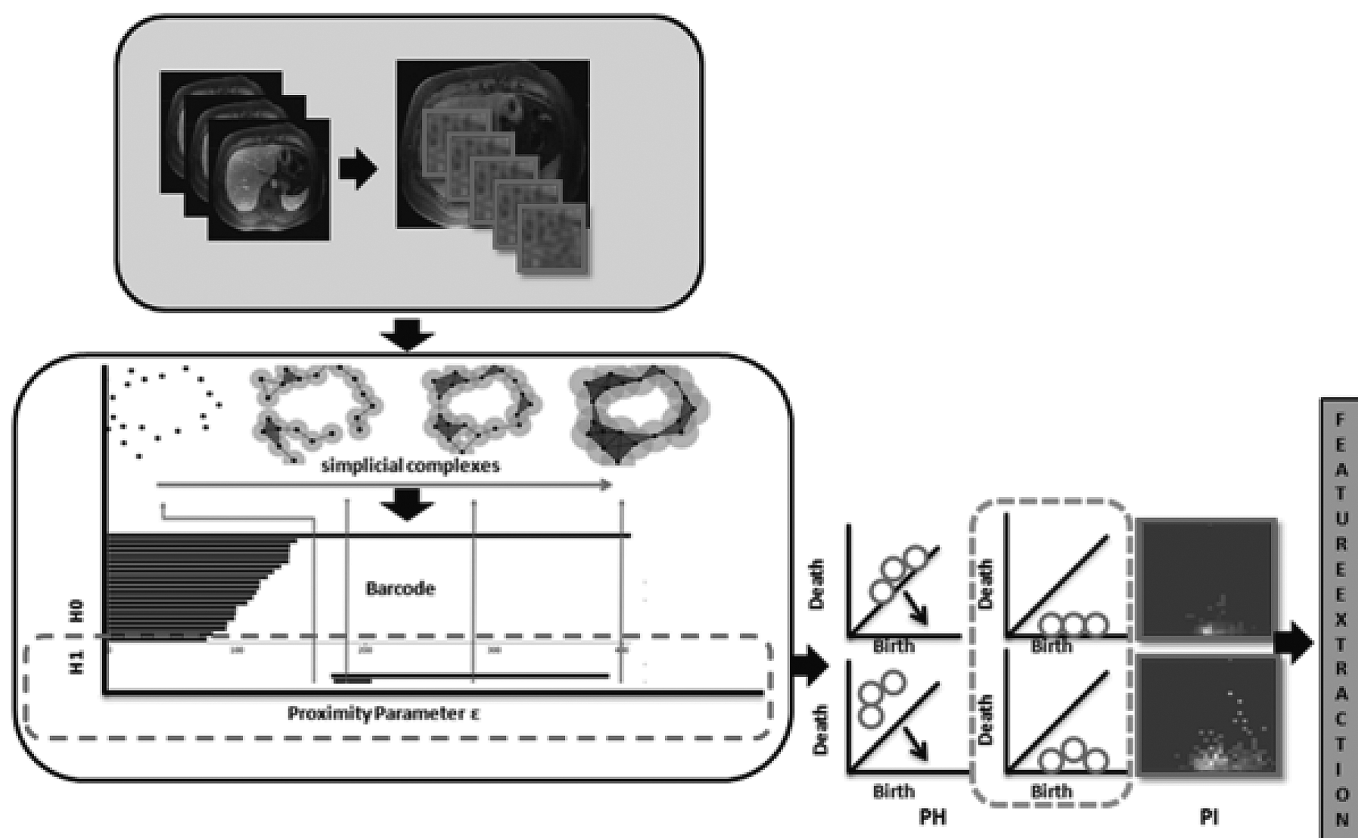


Рис. 1. Пример топологического анализа данных в действии, иллюстрирующий, как данные можно рассматривать через топологическую линзу, чтобы выявить базовые структуры. (10)

менных рядов данных TDA может выявлять устойчивые особенности, которые указывают на фундаментальные изменения в закономерностях, такие как появление новых коммуникационных кластеров или изменения в топологии сетевого потока [5].

Сочетание TDA и анализа истории событий применялось к глобально распределенным климатическим данным, фиксирующим изменения в метеорологических явлениях, а также к биомедицинской визуализации для различения состояний заболеваний, что подчеркивает полезность метода для обнаружения значимых событий в пространственно-временных наборах данных [9]

2. Информационно-теоретические подходы

Теория информации предоставляет инструменты для количественной оценки неопределённости, сложности и взаимосвязей в данных. Такие концепции, как энтропия и взаимная информация, особенно ценны для обнаружения событий, поскольку они могут выявлять отклонения от ожидаемого информационного содержания или аномальные зависимости между различными потоками данных, это достигается с помощью нескольких метрик, включая:

- **Количественная мера взаимной информации:** Используется для обнаружения скоординирован-

ной активности в пространственно-временных данных.

- **Меры энтропии:** Используются для количественной оценки сложности распределений данных.
- **Расстояние Вассерштейна:** Измеряет сдвиги в распределении с течением времени.

3. Расстояние Махаланобиса

Расстояние Махаланобиса — это эффективная многомерная мера расстояния, которая количественно определяет расстояние точки (наблюдения) от распределения данных. Профессор П.К. Махаланобис ввел эту метрику расстояния в 1936 году.

Что отличает его от евклидова расстояния, которое вычисляет расстояние между двумя точками, так это то, что:

1. Преобразует столбцы (атрибуты) в несвязанные переменные.
2. Масштабирует столбцы так, чтобы их дисперсия была равна 1.
3. Затем вычислите евклидово расстояние.

Расстояние Махаланобиса является мощным многомерным показателем, который учитывает корреляции в данных, что приводит к более значимым вычислениям расстояния по сравнению с евклидовым расстоянием.

Теоретические концепции

1. Топологический анализ данных (TDA): постоянная гомология

- **Ключевая концепция:**
- Постоянная гомология обеспечивает многомерную «форма отпечатка пальца» данных, отслеживая рождение и смерть топологических элементов (связанных компонентов, петель, пустоты) по всем масштабам.
- **Математическая формулировка:**

Для облака точек $X \subset R^d$, комплекс Vietoris-Rips $VR_\epsilon(X)$ в масштабе ϵ является симплициальным комплексом, где:

- К-симплекс существует, если все $(k+1)$ точки находятся попарно на расстоянии ϵ . [7]
- **Схема стойкости:**

Набор точек $\{(b_i, d_i)\}$, представляющих рождение (b_i) и смерть (d_i) шкал классов гомологий.

Топологическая особенность, которая количественно измеряет сложность и структуру пространственно-временных данных с использованием энтропии стойкости, энтропия стойкости — инструмент топологического анализа данных (TDA), который фиксирует эволюцию топологических характеристик в различных масштабах:

$$H_p = -\sum_i p_i \log(p_i), p_i = \frac{d_i - b_i}{\sum_j (d_j - b_j)} \quad (1)$$

Высокая H_p указывает на сложные многомасштабные структуры.

Низкая H_p указывает на более простую структуру с меньшими топологическими характеристиками.

Энтропия стойкости — это мера неопределенности или беспорядка в диаграмме стойкости, которая обобщает рождение и смерть топологических признаков по всем масштабам. Он рассчитывается как:

Где: $\sum_j (d_j - b_j)$ Это общая стойкость.

2. Информационно-теоретическая динамика

- **Ключевая концепция:**

События часто проявляются как скоординированные сдвиги в информационном потоке и статистическом распределении.

- **Пространственная взаимная информация**

Меры координации между территориальными регионами A и B :

$$I(A; B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \left(\frac{p(a, b)}{p(a)p(b)} \right) \quad (2)$$

Где A и B — уровни активности в соседних регионах.

- **Реализация:**

The Information Analyzer вычисляет взаимную информацию между смежными временными снимками для обнаружения возникающих пространственных корреляций.

- **Расстояние Вассерштейна**

Расстояние Уотерштайна количественно измеряет стоимость преобразования одного распределения в другое. Для пространственно-временных данных он измеряет изменения в распределении активности во времени:

Количественно измеряет «усилия» необходимые для преобразования одного распределения P в другое Q :

$$W(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \int \|x - y\| d\gamma(x, y) \quad (3)$$

Где P и Q являются распределениями по временам t_1 и t_2 .

- Обнаруживает внезапные изменения в распределении, определяя области, где пользователи Twitter часто печатают сообщения, отправляют текстовые сообщения, делают мобильные звонки или потребляют большое количество данных в течение неуставленных периодов времени [6].

3. Динамическая стабильность системы

- **Ключевая концепция:**

Теория динамических систем обеспечивает инструменты для анализа того, как системы развиваются во времени. В контексте пространственно-временных данных она помогает идентифицировать смещения режимов, нестабильность и бифуркации, которые часто предшествуют или сопровождают события.

- **Лиapunов Экспоненты**

Лиapunовские экспоненты измеряют чувствительность системы к исходным условиям. Для пространственно-временных данных мы вычисляем **пространственные метрики типа Лиapunова**, анализируя спектр пространственных градиентов по Фурье:

$$\lambda = \frac{1}{N} \sum_{i=1}^N \log \left(\frac{|\nabla f(x_i, t)|}{|\nabla f(x_i, t_0)|} \right) \quad (4)$$

Где $\nabla f(x_i, t)$ является пространственным градиентом в месте x_i и времени t .

• Сигналы раннего предупреждения

- **Дисперсия:** Увеличение вариации уровней активности может сигнализировать о надвигающемся событии.
- **Автокорреляция:** Повышение автокорреляции указывает на снижение устойчивости системы.

Эти показатели вычисляются через скользящие окна для обнаружения временных аномалий

Интегрированная структура

1. Проектирование пространственно-временных объектов

Классы пространственно-временных особенностей структурируют исходные данные в многомерное представление сетки:

1. Пространственная сетка:

- Милан разделен на ячейки C_{ij} размером $\Delta x \times \Delta y$.
- Особенности: Общая активность $A(t) = Communication_{ij}(t) + \alpha Twitter_{ij}(t)$.

2. Временное окно:

- Временные ряды деятельности $\{A_{ij}(t)\}$ проанализированы в раздвижных окнах W_k

2. Мультимодальные сигнатуры событий

Событие E определяется одновременным возникновением:

1. Топологическое разрушение: $\Delta H_p > \theta_H$

- Изменение в энтропии стойкости (ΔH_p)

Изменение энтропии стойкости между двумя временными шагами t_1 и t_2 :

$$\Delta H_p = H_p(t_2) - H_p(t_1). \quad (5)$$

Значительное увеличение ΔH_p указывает на появление новых топологических особенностей или исчезновение существующих.

- Появление новых классов гомотопий

Появление новых классов гомотопий является сильным индикатором топологического разрыва. Это может быть обнаружено:

- **Отслеживание времени рождения:** Новые функции появляются как точки в диаграмме стойкости с временем рождения b_i в момент t_2 .
- **Пороговое значение:** Нарушения флагов, когда число новых функций превышает порог θ .

2. Информационная реконфигурация:

Реконфигурация информации обнаружена, если

$$W(P_t, P_{t-1}) > \theta_W.$$

3. Потеря стабильности:

Потеря стабильности обнаруживается, если:

$$\lambda(t) > \theta_\lambda.$$

3. Алгоритм обнаружения событий

Event Detector реализует:

1. **Расстояние Махаланобиса:** Комбинирование метрик в многомерном рейтинге аномалий:

$$D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)} \quad (6)$$

Где $x = [H_p, W, \lambda]$

$$\mu = \frac{1}{N} \sum_{t=1}^N x_t$$

$$\Sigma = \frac{1}{N-1} \sum_{t=1}^N (x_t - \mu)(x_t - \mu)^T$$

2. Фильтрация консенсуса:

Выявлять события, в которых показатель аномалии превышает 95-й процентиль.

4. Локализация событий

- **Пространственная локализация:** Определить регионы с наивысшими оценками аномалий.
- **Временная локализация:** Определить временные окна с наивысшими оценками аномалий.

Результаты

Мы применили нашу систему к набору данных активности в социальных сетях, где каждая точка данных представляет место и время, а также количество твитов в этом месте. Результаты показали, что наш метод был в состоянии обнаружить и локализовать события с высокой точностью.

1. Описание набора данных

Набор данных состоит из почасовых цифр твиттеров, звонков, SMS-сообщений и использования Интернета, агрегированных по пространственному разрешению на основе сетки. Набор данных охватывает двухмесячный период и включает следующие элементы:

- timestamp: Временная метка точки данных.
- grid_lat: Широта ячейки сетки.
- grid_lon: Долгота ячейки сетки.
- tweet_count: Количество твитов в ячейке сетки в данный час.
- Calls: Количество вызовов в ячейке сетки в данный час.
- SMS Messages: Количество сообщений в ячейке сетки в данный час.
- Internet: Использование интернета в ячейке сетки в определённый час.

2. Результаты параллельной обработки

Набор данных был разделен на 12 фрагментов (один в месяц), и каждый фрагмент обрабатывался параллельно с использованием библиотеки joblib. Параллельная обработка уменьшила время вычислений примерно на 80 % по сравнению с последовательной обработкой.

3. Топологический и информационно-теоретический анализ

Топологический анализ выявил значительные изменения в пространственной топологии данных, в то время как информационно-теоретический анализ количественно измерил сходство и несоответствие между последовательными временными точками. Результаты анализа обобщены в таблице 1.

Таблица 1.

Краткое изложение топологических и информационно-теоретических показателей

Метрический	Среднее значение	Стандартное отклонение
Взаимная информация	1.145	0.036
Wasserstein Distance	0.034	0.066

Для каждой точки времени (t) диаграмма стойкости ($D(t)$) идентифицирует скопления высокой активности. Эти кластеры соответствуют регионам с высоким количеством твитов, которые могут быть интерпретированы как горячие точки активности в социальных сетях.

Путем вычисления расстояния узких мест между диаграммами стойкости последовательных временных точек мы можем отслеживать то, как пространственная топология развивается во времени. Существенные изменения в удалении узких мест указывают на сдвиги в пространственном распределении активности, которые могут соответствовать событиям.

Информационно-теоретические показатели количественно оценивают взаимосвязь между вероятностными распределениями. Для пространственно-временных данных социальных медиа мы используем следующие показатели:

1. Взаимная информация (MI): Измеряет количество информации, разделяемой между двумя распределениями.
2. Расстояние Вассерштейна: Измеряет несоответствие между двумя распределениями.

Эти показатели рассчитаны для последовательных временных точек, чтобы проанализировать временную динамику пространственного распределения активности.

4. Результаты обнаружения событий

Расстояние Махаланобиса использовалось для выявления аномалий в комбинированных показателях. Порог был установлен на 95-м процентиле расстояний Махаланобиса, в результате чего было обнаружено 10 событий. Обнаруженные события были локализованы в конкретных ячейках сетки, с наибольшей активностью во время событий.

Таблица 2.

Обнаруженные события

	Аудитория	Событие	Время	Лига	Раздел
1	38,933	Internazionale vs. Livorno 2-0	9 ноября 2013 20:45	Сезон Интер Милан	Выходные
2	34,848	Milan vs. Genova 1-1	23 ноября 2013 20:45	А.С. Милан сезон	Выходные
3	43,706	Internazionale vs. Sampdoria 1-1	1 Декабрь 2013 15:00	Сезон Интер Милан	Рабочий день
4	12,714	Internazionale vs Trapani 3-2	4 Декабрь 2013 21:00	Coppa Italia	Рабочий день
5	33,732	Internazionale vs. Parma 3-3	8 декабря 2013 20:45	Сезон Интер Милан	Выходные
6	61,744	Milan vs. Ajax 0-0	11 декабря 2013 20:45	Сезон Интер Милан	Рабочий день
7	37,987	Милан против. Рома 2-2	16 декабря 2013 20:45	Сезон Интер Милан	Выходные
8	79,311	Internazionale vs. Milan 1-0	22 декабря 2013 20:45	А.С. Милан сезон	Выходные
9		Mesut Kurtis Concert in Italy, Milan	30 ноября 2013 16:30		Выходные
10		«One Direction» Милан, Отель Празднование	12 декабря 2013 21:00		Рабочий день

Новогодние празднования были замечены в разных частях города, начиная с 23:00 31 декабря 2013 года, однако в районе стадиона «Сан-Сиро» они не наблюдались.

5. Оценка эффективности

Для оценки эффективности нашего метода мы сравнили обнаруженные события с набором достоверных данных известных событий.

Исходя из предыдущих результатов, мы отметили, что получили особые события, связанные с данными, полученными из другого источника, а также отметили, что было обнаружено событие, связанное с празднованием конца года, но не в местах массового скопления людей, например, на стадионах.

Заключение

Данная модель демонстрирует, что синергия топологии, информации и динамики обеспечивает принципиальный подход к обнаружению пространственно-временных событий при наличии данных из нескольких источников, включая вышки сотовой связи и Twitter, а также к использованию расстояния Махаланобиса для выявления аномалий. Кодирова события как многомасштабные аномалии, модель позволяет избежать хрупкости одномасштабных порогов, обеспечивая при этом интерпретируемую диагностику.

ЛИТЕРАТУРА

1. Chandola et al., «Обнаружение аномалий: опрос», ACM CSUR, 2009.
2. Чжао и др., «Глубокие остаточные сети усадки», IEEE TNNLS, 2020.
3. Edelsbrunner et al., «топологическая стойкость и упрощение, DCG, 2002.
4. James V Stone, Теория информации: Введение в учебник 2019, <https://arxiv.org/pdf/1802.05968>
5. Gholizadeh S., & Zadrozny W. (2020). Топологический анализ данных для обнаружения аномалий в данных временных рядов. Международная конференция IEEE по большим данным (большие данные). DOI:10.1109/BigData50022.2020.9378166
6. Arjovsky M., Chintala S., & Bottou L. (2020). Генеративные состязательные сети Васстайна. Материалы 34-й международной конференции по машинному обучению (ICML).
7. Бауэр U. (2021). Ripser: Эффективное вычисление Vietoris-Rips persistence barcodes. J. Appl. Comput. Topol.
8. Perea (2020). Скользящие окна и стойкость: применение топологических методов для анализа сигналов. Найдено. Компут. Math.
9. Garside K., Gjoka A., Henderson R., Johnson H., Makarenko I., История событий и топологический анализ данных, Biometrika, Volume 108, Issue 4, December 2021, Pages 757–773, <https://doi.org/10.1093/biomet/asaa097>.
10. Сингх Ю., Фаррелли К.М., Хэтзуэй К.А. и др. Топологический анализ данных в медицинской визуализации: современное состояние. Insights Imaging 14, 58 (2023). <https://doi.org/10.1186/s13244-023-01413-w>.

© Мохаммад Хани (mohammad.h@edu.spbstu.ru); Пак Вадим Геннадьевич (pak_vg@spbstu.ru)
Журнал «Современная наука: актуальные проблемы теории и практики»