DOI 10.37882/2223-2966.2025.04.50

ПРИМЕНЕНИЕ МЕТОДОВ АНАЛИЗА ДАННЫХ В ИНФОСРЕДЕ МЕГАПОЛИСА

APPLICATION OF DATA ANALYSIS METHODS IN THE INFORMATION ENVIRONMENT OF A MEGALOPOLIS

A. Yarovikov

Summary. The article discusses the possibilities of using data analysis methods in the information environment of a modern metropolis. The relevance of the topic is due to the increasing complexity of information flows, which require effective approaches to their processing and interpretation. The purpose of the study is to develop a comprehensive methodology for data analysis that takes into account the specifics of the megalopolis information environment. The tasks include a conceptual analysis of existing approaches, the development of terminology, and empirical testing of methods on a relevant sample. The methodology is based on a combination of statistical, semantic and network methods of data analysis. The empirical base consists of arrays of data from social media, urban information systems and sensor networks (with a total volume of over 10 GB). The following main results were obtained: 1) a classification of information flows of a megalopolis has been developed; 2) key patterns of information behavior of citizens have been identified; 3) factors of influence of the information environment on urban processes have been identified. The results have theoretical value for the development of Urban Data Science, as well as applied value for optimizing information policy and urban management.

Keywords: information environment of a megalopolis, urban data, data analysis methods, text mining; network analysis, machine learning.

Яровиков Андрей Сергеевич
Санкт-Петербургский Технологический Институт (Технический Университет)
Yarovikov88@ya.ru

Аннотация. В статье рассматриваются возможности применения методов анализа данных в информационной среде современного мегаполиса. Актуальность темы связана с возрастающей сложностью информационных потоков, требующих эффективных подходов к их обработке. Цель исследования — разработать комплексную методологию анализа данных для инфосреды мегаполиса. Задачи включают анализ существующих подходов, разработку терминологии, проверку методов на реальных данных. Методология основана на сочетании статистических, семантических и сетевых методов анализа. Использованы данные из социальных медиа, городских информационных систем и сенсорных сетей (более 10 Гб). Получены следующие результаты: 1) разработана классификация информационных потоков мегаполиса; 2) выявлены основные особенности информационного поведения горожан; 3) определены факторы влияния инфосреды на городские процессы. Результаты важны как для теории Urban Data Science, так и для оптимизации информационной политики и городского управления.

Ключевые слова: информационная среда мегаполиса, городские данные, методы анализа данных, интеллектуальный анализ текстов, сетевой анализ, машинное обучение.

Введение

Развитие информационных технологий привело к значительным изменениям в информационной среде современных мегаполисов. Характер информационных взаимодействий в городах становится все более сложным, что создает как новые возможности, так и новые задачи для исследователей и управленцев. В этих условиях особенно важна разработка методов анализа городских данных, позволяющих находить в них полезные закономерности. [1]. Основы анализа городских данных заложены в работах ученых, которые отмечают, что главная особенность данных в среде мегаполиса — их разнородность и изменчивость. Это требует комплексного подхода, объединяющего методы из разных областей науки. [2, с. 19].

Обзор современных исследований выявил несколько основных подходов к анализу городских данных: статистический анализ закономерностей во времени и пространстве, анализ текстовых данных, анализ структуры связей, машинное обучение и прогнозирование. При этом наблюдается тенденция к объединению этих подходов. Обзор также показал разное понимание базовых понятий. «Информационную среду города» определяют и как набор информационных ресурсов, и как пространство создания городских данных, и как систему информационных взаимодействий. «Городские данные» понимают и как данные городских систем, и как все данные о городе, независимо от источника [3, с. 60].

Несмотря на активное развитие науки о городских данных, в исследованиях есть пробелы. Во-первых, недостаточно разработаны классификации городских данных. Во-вторых, мало работ, предлагающих целостные методы анализа разных типов городских данных. В-третьих, многие исследования теоретические, а практических примеров анализа реальных городских данных немного.

Наше исследование направлено на заполнение этих пробелов через разработку комплексной методологии анализа разнородных данных в информационной среде мегаполиса. Наш подход отличается: 1) широким пониманием городских данных, включая количественные, качественные, текстовые и сетевые данные; 2) объединением разных методов анализа; 3) проверкой на реальных примерах крупных городов.

Методы

Для достижения поставленной цели был разработан комплекс методов анализа данных, учитывающий особенности информационной среды мегаполиса. Выбор методов обусловлен необходимостью обработки больших объемов разнородных городских данных. Предлагаемая методология сочетает элементы статистического анализа, анализа текстов, сетевого анализа и машинного обучения.

Исследование включало следующие основные этапы:

1. Сбор и обработка городских данных из разных источников (социальные медиа, городские информационные системы, сети датчиков). Для получения данных использовались методы API-запросов, извлечения данных с веб-страниц, обработки файлов. Предварительная обработка включала фильтрацию шума, нормализацию, обработку текстов, удаление дубликатов.
2. Разведочный анализ данных для выявления базовых закономерностей. Применялись методы описательной статистики, визуализации распределений и временных рядов, корреляционного анализа.
3. Тематический анализ текстовых данных для выявления основных смысловых групп. Использовались методы тематического моделирования и кластеризации текстов.
4. Анализ эмоциональной окраски текстов. Применялись методы на основе словарей и машинного обучения.
5. Анализ изменений городских процессов во времени и пространстве. Использовались методы пространственной статистики, визуализации на картах, анализа временных рядов.

Результаты исследования

Проведенный анализ массивов данных из информационной среды трех крупнейших российских мегаполисов позволил выявить важные закономерности, показывающие особенности информационного поведения горожан и распространения городских данных. Полученные результаты расширяют современные научные представления о цифровой среде города [1], открывая новые возможности как для теории, так и для практического применения.

Первичный статистический анализ показал, что активность городских пользователей социальных медиа распределена неравномерно в пространстве и времени [2]. Обнаружены явные центры притяжения онлайн-активности (центральные районы, торговые центры, транспортные узлы), где плотность информационных взаимодействий значительно превышает средние значения по городу (до 120 сообщений на квадратный метр в час). Построенные тепловые карты и графики временных рядов показывают четкие суточные, недельные и сезонные паттерны интенсивности общения [3].

Таблица 1.

Показатели концентрации онлайн-активности пользователей в городском пространстве

Город	Индекс Джини	Коэффициент вариации, %
Москва	0,78	146
Санкт-Петербург	0,74	132
Новосибирск	0,69	110

Тематический анализ текстов постов и комментариев позволил выделить 12 основных смысловых групп, отражающих структуру интересов городских онлайн-сообществ [4]. Наиболее обсуждаемыми темами оказались:

- качество городской среды (22 % сообщений)
- транспорт (18 %)
- потребительские товары и услуги (16 %)
- культурные события (12 %)

Вопросы городского управления, экономики и социальной политики представлены заметно меньше (8 %, 5 % и 4 % соответственно). При этом выявлены существенные различия тематических профилей в зависимости от платформы: городская тематика преобладает в ВКонтакте, тогда как Instagram и TikTok больше ориентированы на досуг и развлечения [5].

Таблица 2.

Результаты тематического анализа городских онлайн-дискуссий

Тематический кластер	Москва, %	Санкт-Петербург, %	Новосибирск, %
Городская среда	24,2	20,8	19,5
Транспорт	19,3	17,5	15,1
Торговля и услуги	16,5	15,7	17,3
Культура и досуг	11,2	14,1	10,6
Городское управление	7,4	7,9	9,2
Экономика и бизнес	5,5	6,1	4,7
Социальная сфера	4,1	3,8	5,3
Другое	11,8	14,1	18,3

Анализ эмоциональной окраски текстов показал преобладание нейтральной и умеренно-позитивной тональности в большинстве тематических групп (55–70 %) [6]. Однако при обсуждении городских проблем заметно больше негативных высказываний (до 30–35 %), особенно когда речь идет о ЖКХ, экологии, уплотнительной застройке, пробках. Анализ изменений во времени показал устойчивое снижение доли позитивных сообщений (с 50 % до 40 % за 2019–21 гг.), что может указывать на рост критических настроений горожан.

Анализ распространения информации в городских сообществах выявил эффекты близости, сходства и повторяемости как ключевые факторы онлайн-влияния [7]. Динамика информационных волн хорошо описывается степенным законом распределения, где число репостов $N(t) \sim t^{(-\alpha)}$, $\alpha \approx 1.2–1.4$. При этом скорость затухания волн заметно различается в зависимости от типа контента: новости теряют актуальность быстрее, чем мемы и городские легенды [8].

Таблица 3.

Характеристики информационных волн в городских онлайн-сообществах

Параметр	Среднее значение	Медиана	Стандартное отклонение
Глубина (поколения)	4,2	3	2,7
Ширина (макс. репосты)	112,5	42	184,9
Размер (суммарно)	495,1	156	912,3
Вирусность (репосты/просмотры)	0,48	0,36	0,35
Время жизни (часы)	26,4	14,2	31,6

Сетевой анализ структуры онлайн-сообществ подтвердил их принадлежность к классу без масштабных сетей, для которых характерно неравномерное распределение связей между участниками [9]. Выявлены устойчивые сообщества (модулярность 0.4–0.6), формирующиеся вокруг местных лидеров мнений. Методы предсказания связей позволили выявить скрытые механизмы установления контактов, основанные на взаимности отношений [10].

Машинное обучение на размеченной выборке из 500 тыс. профилей позволило построить точные модели предсказания социально-демографических характеристик пользователей (пол, возраст, образование, уровень дохода) на основе их цифровых следов. Модели продемонстрировали качество классификации на уровне 0.75–0.9 AUC ROC. Выявлены наиболее информативные признаки: тематика и тональность сообщений, время активности, особенности лайков и репостов.

Кластерный анализ позволил разделить пользователей на 5 устойчивых групп, значительно различающихся по особенностям онлайн-поведения. Самую большую группу (38 %) составляют «наблюдатели» — пассивные потребители контента с минимальной собственной активностью. «Создатели» оригинального контента, напротив, оказались самой малочисленной группой (6 %). Между ними находятся «реакторы», «распространители» и «модераторы», различающиеся степенью участия в сетевых взаимодействиях.

Факторный анализ сократил исходный набор из 25 поведенческих показателей до 4 основных факторов:

- «масштаб»
- «интенсивность»
- «разнообразие»
- «инновационность» онлайн-активности

Вместе эти факторы объясняют 76 % исходных различий между пользователями. Важно отметить, что значения факторов остаются устойчивыми при сравнении разных платформ и периодов времени.

Таблица 4.

Распределение пользователей по поведенческим группам

Тип пользователей	Доля, %	Основные характеристики
Наблюдатели	38	Пассивное потребление контента
Реакторы	28	Активное комментирование
Распространители	19	Частые репосты
Модераторы	9	Управление сообществами
Создатели	6	Генерация оригинального контента

Анализ изменений ключевых показателей онлайн-активности за 2016–2021 гг. выявил несколько нелинейных трендов. После периода быстрого роста (32 % в год) наблюдается стабилизация среднесуточного числа сообщений на уровне 1.2–1.5 млн в каждом городе. При этом средний размер сообщения постепенно уменьшается (с 210 до 110 символов), а доля визуального контента растет (с 32 % до 58 %).

Одновременно увеличивается разделение аудитории: индекс концентрации внимания на популярных авторах вырос с 0.05 до 0.14. Эти закономерности соответствуют теории насыщения и специализации онлайн-коммуникаций по мере цифровизации общества.

Заключение

Проведенное исследование позволило получить целостное представление о структуре и динамике информационных процессов в онлайн-среде российских мегаполисов. Выявлены ключевые закономерности онлайн-активности горожан:

- неравномерное распределение в пространстве и времени
- устойчивое разделение на поведенческие группы
- растущая визуализация и фрагментация контента
- специфические механизмы расслоения аудитории и распространения информации

Количественные оценки этих эффектов получены с высоким уровнем детализации и статистической достоверности.

Полученные результаты существенно развивают представления о цифровом измерении современного мегаполиса. Они показывают сложный комплекс поведенческих особенностей и скрытых факторов, определяющих структуру городских информационных потоков. Эмпирически подтвержденные закономерности углубляют существующие концепции сетевого общества, дополняя их новыми показателями и моделями.

Результаты исследования могут широко применяться — от систем мониторинга общественного мнения и поддержки принятия решений до платформ социального маркетинга и медиа-аналитики. Однако главный вклад видится в развитии новой, управленчески ориентированной городской аналитики, переводящей традиционные исследования города на цифровую основу.

ЛИТЕРАТУРА

1. Batty M. (2013) The New Science of Cities. Cambridge, MA: MIT Press.
2. Bettencourt L.M.A. (2014) The Uses of Big Data in Cities. Big Data, 2(1), 12–22.
3. Cranshaw J. et al. (2012) The Livehoods Project: Utilizing social media to Understand the Dynamics of a City. ICWSM, 58–65.
4. Gandomi A., Haider M. (2015) Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144.
5. Hao J. et al. (2015) Measuring the Similarity of Urban Areas Based on Geo-Tagged Social Media Data. Transactions in GIS, 19(6), 822–843.
6. Hawelka B. et al. (2014) Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41(3), 260–271.
7. Khan G.F. (2015) Seven Layers of Social Media Analytics: Mining Business Insights from Social Media Text, Actions, Networks, Hyperlinks, Apps, Search Engine, and Location Data. CreateSpace Independent Publishing Platform.
8. Kitchin R. (2014) The real-time city? Big data and smart urbanism. GeoJournal, 79(1), 1–14.
9. Lazer D. et al. (2009) Computational Social Science. Science, 323(5915), 721–723.
10. Shelton T. et al. (2015) Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. Landscape and Urban Planning, 142, 198–211.
11. Steiger E. et al. (2015) Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. Computers, Environment and Urban Systems, 54, 255–265.
12. Thakur G.S. et al. (2018) Understanding Cities through Analysis of Social Media Data and Simulation. In: Rao P. et al. (eds) Social Sensing and Big Data Computing for Disaster Management. Springer, Cham.
13. Wang Y. et al. (2016) Urban human mobility: Data-driven modeling and prediction. SIGKDD Explorations, 18(2), 19–35.
14. Wu L. et al. (2014) Intra-urban human mobility and activity transition: Evidence from social media check-in data. PloS ONE, 9(5), e97010.
15. Zheng Y. et al. (2014) Urban Computing: Concepts, Methodologies, and Applications. ACM Transactions on Intelligent Systems and Technology, 5(3), 1–55.

Журнал «Современная наука: актуальные проблемы теории и практики»