УДК 004.89
DOI:10.25729/ESI.2024.33.1.014

Применение методов географического информационного поиска для анализа новостных данных

Авдюшина Анна Евгеньевна, Королёва Юлия Александровна, Маркина Татьяна Анатольевна, Бессмертный Игорь Александрович
Университет ИТМО РФ, Санкт-Петербург, *avdiushina@itmo.ru*

Аннотация. Статья посвящена выделению неформальных районов на основе данных из новостных источников и социальных сетей по критерию географической близости. Предложена методика извлечения геоданных из текстов для пространственной кластеризации. Из текстов извлекаются географические названия, которые преобразуются в геолокации с помощью геокодирования. Выделенные геоточки кластеризуются по плотности и для каждого кластера определяется распределение тем. Данный подход позволяет абстрагироваться от административного деления и выявить кластеры, близкие к восприятию горожан. Результаты кластеризации перспективно применять в разнообразных задачах управления городской инфраструктурой: мониторинг общественной жизни, анализ качества городской среды, общественная безопасность. Отличие предложенной методики от аналогов состоит в синтезе геоданных для группировки объектов. Разработанные на основе методики программные средства позволяют принимать решения в области урбанистики: развитие микрорайонов города и транспортной инфраструктуры, размещение социально значимых объектов и обеспечение безопасности.

Ключевые слова: интеллектуальный анализ данных, умный город, информационная модель, системы поддержки принятия решений, кластеризация, геопространственные данные.

Цитирование: Авдюшина А.Е. Применение методов географического информационного поиска для анализа новостных данных / А.Е. Авдюшина, Ю.А. Королёва, Т.А. Маркина, И.А. Бессмертный // Информационные и математические технологии в науке и управлении. – 2024. – № 1(33). – С. 154-165. – DOI:10.25729/ESI.2024.33.1.014.

Введение. Объем данных, публикуемых каждый день, неуклонно растет, для преобразования его в полезную информацию требуются эффективные методы анализа неструктурированной информации. Анализ городских новостных данных из различных информационных источников является актуальной задачей для понимания происходящих процессов и перспектив развития умного города.

Предварительное исследование показало, что существуют работы по извлечению геолокации из данных социальных сетей, только в том случае, если пользователь отметил геолокацию явно. Так же существует отдельное направление по анализу и извлечению информации из текстов, но нет работ, описывающих алгоритм (конвейер) извлечения геоданных из непривязанных текстовых данных явно, и дальнейшее разбиение на кластеры, соответствующие выделенным темам.

Для достижения поставленной цели перспективно использование методов из двух, на первый взгляд не связанных, научных направлений. Обработка естественного языка (NLP – Natural Language Processing) является дисциплиной на стыке лингвистики, компьютерных наук и искусственного интеллекта, направленной на анализ и понимание человеческого языка с помощью алгоритмов. Наука о географической информации (GIScience) изучает методы сбора, анализа и визуализации геопространственных данных. Хотя NLP и GIScience являются различными направлениями, их комплексное использование позволяет связать анализ текстовых данных с географическим контекстом. При этом NLP помогает извлекать информацию из текста, а GIScience – интерпретировать и визуализировать географическое расположение этой информации.

В данной работе предлагается метод определения тематического образа местности. Обработка данных даёт возможность получить список агрегированных тем, показывающий полное распределение тем по областям, определяет контекст происходящего в отдельных

частях города, а именно, какие темы и события наиболее характерны для различных частей города, которые не привязаны к административному делению.

1. Обзор существующих подходов к выделению тем и кластеризации городских данных. Городские данные являются разновидностью геопространственных данных, включают в себя информацию о местоположении объектов и явлений в городском пространстве, представленную обычно через географические координаты. Они предоставляют уникальную возможность для анализа и понимания городских процессов, позволяя идентифицировать и визуализировать пространственные взаимосвязи и динамику городской среды. Этот вид данных объединяет информацию о событиях и их местоположении, что критически важно для градостроительного планирования, управления городскими ресурсами и разработки стратегий улучшения качества жизни в городах.

Существующее административное деление регионов или районов не всегда совпадает с реальным делением. Нахождение такого представления регионов можно составить на основе характеристик ключевых признаков, отличающих один регион от другого, и формирующих их уникальность. Описание подобного представления городской среды можно сделать с помощью ментальных карт [1], которые включают в себя социокультурный анализ и проведение проектных сессий.

Использование ГИС позволяет отображать активность и особенности территории. Однако, данный подход ограничен тем, что объем данных, публикуемых каждый день, неуклонно растет. Преобразование такого огромного объема информации в знания, особенно касающиеся местоположения и контекста территорий, требует эффективных методов анализа неструктурированной информации, выделения границ неформальных районов.

В рамках интеллектуального анализа текста тематическое моделирование, в частности, скрытое распределение Дирихле (LDA – Latent Dirichlet Allocation), представляет собой мощный метод обнаружения и анализа абстрактных тем в коллекции текстовых документов [2, 3]. Тематическое моделирование используется в разных областях исследований – от анализа социальных сетей до геоинформационного поиска.

Количество информации в социальных сетях растет экспоненциально, а также информации о географических местах становится всё больше, именно такие данные могут стать источником новых возможностей исследования. Одним из важных направлений является развитие интеллектуальных систем поддержки принятия решений, где анализ геотегированных данных позволяет определить популярные места и интересы пользователей [3]. Дополнительное применение тематического моделирования с использованием LDA, pLSA (probabilistic Latent Semantic Analysis) и ml-PLSI (вероятностная тематическая модель) позволяет обнаруживать городские события, включая стихийные бедствия, а также классифицировать многозначные текстовые документы с использованием вероятностного подхода [4]. Отслеживание изменений тем во времени позволяет отслеживать тенденции в обществе, строить онтологии с помощью систем семантического моделирования и связывать непространственные концептуальные иерархии с онтологией места, основанной на семантике классификации для создания интегрированной меры семантической близости, которую можно использовать для ранжирования релевантности извлеченных объектов, или моделировать поведение и реакцию городских жителей [5, 6]. Иерархическая мера расстояния сочетается с евклидовым расстоянием между центроидами мест для создания гибридной меры пространственного расстояния. Это может объединить данные из социальных сетей, часто содержащих хэштеги, ключевые слова, классифицирующие темы, поэтому тематическое моделирование с кластеризацией хэштегов [7] становится многообещающим методом анализа коротких текстов.

Следует отметить, что для решения задач географического информационного поиска активно используется подход коллабаративного обучения, основанного на временной, географической и социальной информации для поиска географической информации с использованием социальных сетей, Интернета и баз геоданных [8, 9].

Исследования, включающие кластеризацию геопространственных данных с применением плотностных и адаптивных алгоритмов, используются для анализа новостных данных с учетом географического контекста, в социальных исследованиях, интеллектуальном анализе текста и в геоинформационных системах. Плотностные методы кластеризации текстовых данных с учетом местоположения улучшают классифицирование и анализ новостных данных за счёт определения объектов, семантически близких друг к другу [10-12]. Важной областью исследований является анализ аварий и прогнозирование. Исследование [13] представляет адаптивный метод анализа аварий и кластеризации с использованием географических данных, что имеет большое значение для обеспечения безопасности и предотвращения несчастных случаев.

В статье [14] используется подход, основанный на географическом тематическом моделировании данных из социальных сетей с пространственным контекстом. Методика включает анализ текстовой информации из социальных сетей, обогащенной географическими тегами, для выявления и визуализации тематических и пространственных закономерностей. Это позволяет исследователям лучше понимать социокультурные явления в различных географических регионах, а также способствует более эффективному мониторингу и анализу общественных настроений и трендов в реальном времени. Однако, основным недостатком подхода является использование исключительно геотегированных постов, что приводит к потере значительного количества семантики текстов и создает сложности с обогащением данных при использовании нескольких источников. Это ограничивает анализ, делая его менее гибким для более широкого исследования, например, для определения точек интереса за пределами прямо геотегированных упоминаний.

2. Методология. В основе данной разработки лежит идея о восприятии городской среды через яркие литературные тексты или цитаты. Для составления такого образа необходимо построить конвейер, концептуальная схема которого представлена на рисунке 1.

Данная диаграмма описывает методологию, а для практического применения необходимо определить этапы конвейера. В качестве источника данных использованы новостные статьи и социальные сети. После сбора данных из выбранных источников и их предобработки следует этап извлечения именованных сущностей. На описанных этапах применялись методы NLP. На последних этапах конвейера реализуется геокодирование именованных сущностей, пространственная кластеризация и анализ распределения тем внутри каждого геокластера с использованием методов GIScience. Конечная цель – сформировать актуальное распределение тем, которое освещается официальными источниками и отображает реальную кластеризацию районов города.

2.1. Сбор данных. Исходные данные для данного исследования были получены с использованием двух основных методов: веб-краулинга новостного сайта «Фонтанка» и подключения к социальной сети "ВКонтакте" с помощью API (Application programming interface). Эти данные включают в себя заголовки новостей, тексты статей и информацию о местоположении событий. Информация собиралась с целью создания набора данных, который позволит исследовать события, происходящие в различных районах и территориях. Для хранения, обновления и обработки информации создана база данных под управлением СУБД PostgreSQL.

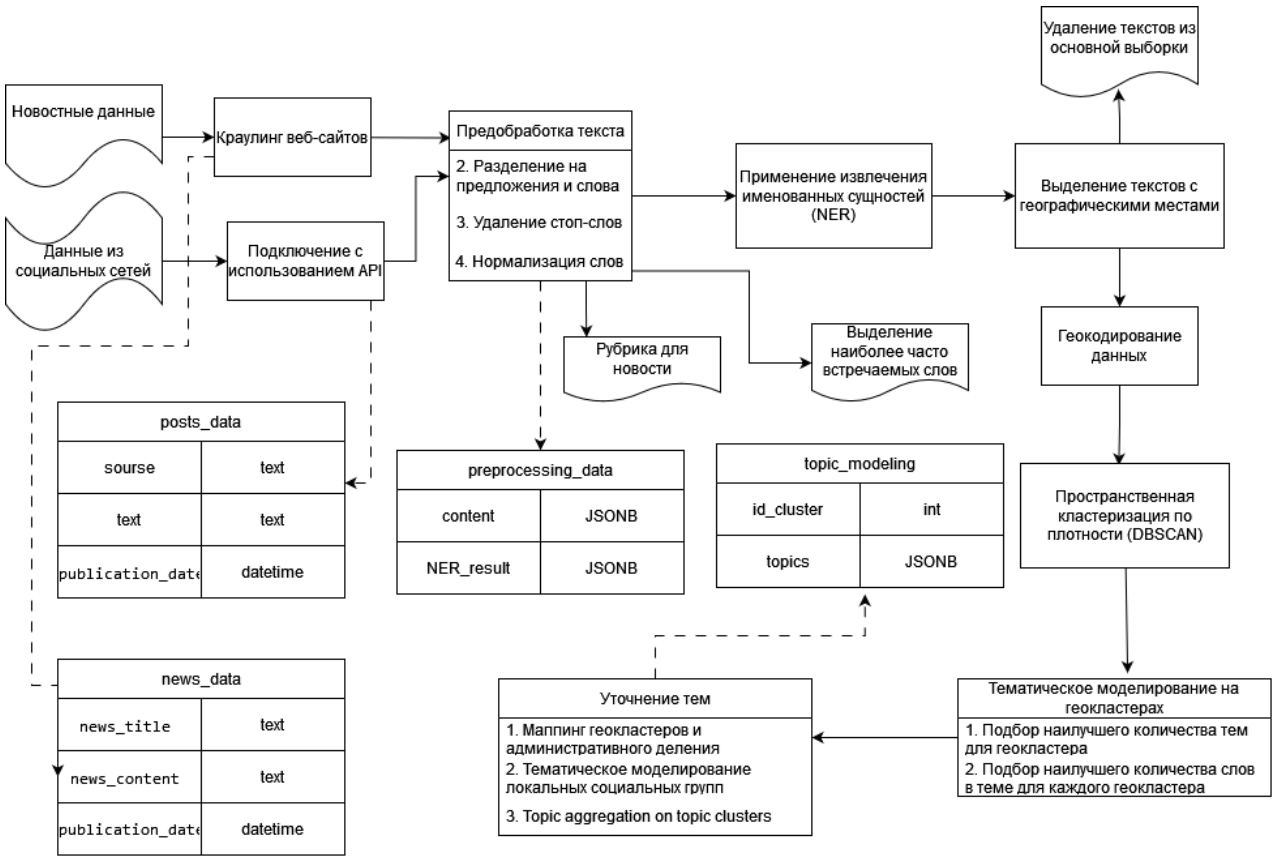

Рис. 1. Концептуальная диаграмма для обнаружения распределения тем по геокластерам

2.1.1. Веб-краулинг новостного сайта. Для сбора новостных данных использовался веб-краулинг, что позволило извлечь информацию с новостного сайта. Веб-краулинг – это метод сбора информации с веб-ресурсов, при котором агент автоматически переходит по страницам сайта, извлекая необходимую информацию, такую, как заголовки новостей, текст статей, даты публикации и географические координаты, если они доступны. Собранная информация была сохранена в базе данных PostgreSQL для дальнейшего анализа.

2.1.2. Подключение к социальной сети "ВКонтакте" по API. Для получения данных из социальной сети "ВКонтакте", был использован API данной платформы. API "ВКонтакте" предоставляет доступ к различным группам и сообществам и позволяет извлекать данные постов и комментарии пользователей. Также разработан инструмент для дополнения в созданную базу данных этой информации в потоковом режиме.

2.1.3. Объединение данных и удаление дублирующихся новостей. После сбора данных из двух источников, информация была объединена в одной базе данных для дальнейшей обработки. Объединение позволило снизить трудозатраты на анализ и вероятность ошибок.

При объединении данных также проведена процедура удаления дублирующихся новостей, которые могли возникнуть из-за перекрытия информации между новостным сайтом и социальной сетью. Были применены алгоритмы сравнения и удаления дубликатов, что позволило уменьшить избыточность данных и улучшило качество исходных данных для последующего анализа.

2.2. Предобработка текста. На этом этапе происходило удаление лишних символов, преобразование текста в нижний регистр, а также удаление стоп-слов, которые не несут смысловой нагрузки. Такая предобработка позволяет сделать текст более структурированным и готовым для дальнейшего анализа.

Для предобработки текстовых данных применены разные библиотеки языка Python, которые добавлены в этапы конвейера. В рамках исследования использованы следующие методы и инструменты:

1. **Сегментация текста и морфологический анализ:** с помощью сегментатора и морфологического словаря Natasha текст был разбит на отдельные слова, а также проведен морфологический анализ. Это позволило определить части речи и грамматические характеристики слов.
2. **Морфологическая разметка:** для детального морфологического анализа текста применялся инструмент SpaCy, из которого применялись функции лемматизации и определение частей речи.
3. **Синтаксический анализ:** Для изучения структуры предложений и связей между словами использовался Stanford NLP Parser, который определяет синтаксические зависимости для каждого предложения.

Описанный этап предобработки необходим для формирования понимания исходных данных, проведения разведывательного анализа данных.

2.3. Извлечение именованных сущностей (NER – Named Entity Recognition). Для выявления именованных сущностей, включая названия организаций, географические местоположения, использовался инструмент Natasha, при этом выделялись такие типы сущностей, как наименования организаций, географические наименования, а также координаты. На данном этапе определялись тексты для дальнейшего использования. Тексты, в которых не получилось выявить именованную сущность, вероятнее всего, не представляют ценности. Использование NER может быть улучшено путём дообучения модели с использованием топонимов.

2.4. Геокодирование данных. Извлеченные именованные сущности, связанные с местоположением, преобразовывались в географические координаты с помощью геокодирования, это является важным этапом для анализа геопространственной информации. Использована библиотека yandex_geocoder, предоставляющая возможность обращения к геокодированию с использованием API. Так как не все географические наименования и названия организаций могут быть преобразованы в координаты, поэтому на этом этапе формируется основная выборка (словарь) для дальнейшей работы, но также остаётся подвыборка с неопределёнными местоположениями. Таким образом, геокодирование позволяет преобразовать текстовые локации в географические координаты, что делает возможным дальнейший анализ и визуализацию данных. Для сохранения данных с географическими координатами используется расширение PostgreSQL – PostGIS.

2.5. Пространственная кластеризация. Следующим этапом является геопространственная кластеризация геокодированных данных с использованием алгоритма DBSCAN [15]. Данный метод позволяет группировать географические точки в кластеры на основе плотности расположения объектов. Это помогает выделить географические области, в которых события близки друг к другу. Результатом пространственной кластеризации является набор геокластеров, обозначаемых как C.

2.6. Анализ распределения тем внутри каждого кластера с использованием LDA. Для каждого из геокластеров, обозначаемых как C_i в наборе C, происходит извлечение текстовых данных, связанных с событиями внутри кластера (новости и посты из социальных сетей). Затем применяется модель LDA для анализа распределения тем внутри каждого кластера. LDA позволяет определить, какие темы наиболее характерны для текстовых данных в каждом кластере. Например, это позволяет определить, какие события и темы наиболее

актуальны в определенных районах. Модель базируется на вероятностном распределении слов по темам и может быть представлена следующим образом [2]:

$$P\left(\theta\middle|\alpha\right)=\mathrm{Dir}\left(\theta\middle|\alpha\right)=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}, \quad (1)$$

где

θ – распределение Дирихле на темы внутри кластера;

α – вектор гиперпараметров Дирихле для распределения тем внутри кластера;

K – количество тем в модели LDA.

Таким образом, результатом этого этапа является набор распределений тем θ, каждое из которых соответствует одному из геокластеров в C.

3. Результаты. В результате сбора данных сформирован датасет, включающий в себя 20 тысяч записей (15 тысяч постов и комментариев из «Вконтакте» и 5 тысяч новостей сайта «Фонтанка»). После предварительной обработки, включающей удаление неинформативных элементов, таких, как ссылки, специальные символы и стоп-слова, а также лемматизацию, объем данных был сокращен примерно на 15%, составив 17 тысяч текстов. Размер каждого текста составляет от 10 до 1600 символов.

На рисунке 2 представлена карта города Санкт-Петербурга, на которой отмечены точки, соответствующие географическим координатам новостей и постов. Каждая точка на карте представляет собой местоположение, связанное с определенным текстом. Так как тесты определялись по схожести и для того, чтобы избежать дублирования, к каждой точке привязаны от 1 до 5 текстов. Тексты сравнивались с использованием косинусной меры в рамках одного дня публикации для определения схожих и объединения их в один. Рисунок 2 демонстрирует пространственное распределение данных, что позволяет визуально оценить, какие районы города наиболее активны с точки зрения информационных событий.

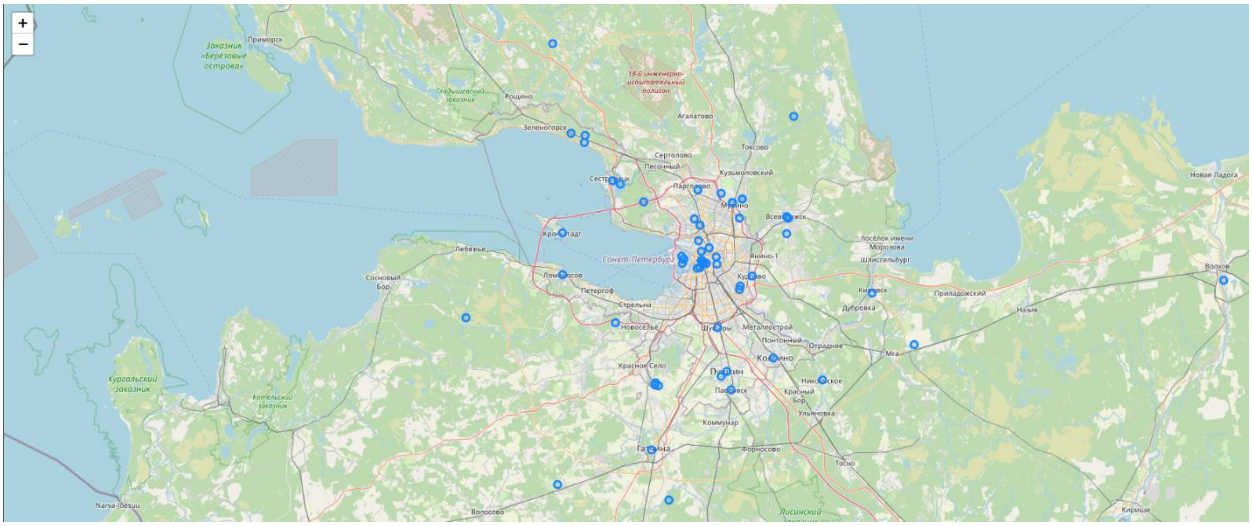

Рис. 2. Карта Санкт-Петербурга с нанесёнными точками координат новостей и постов

Для кластеризации и тематического моделирования использовались методы DBSCAN и LDA соответственно, где количество кластеров и тем определялось на основе анализа метрик качества (коэффициент силуэта для кластеризации и коэффициент когерентности для тематического моделирования).

Коэффициент силуэта [15] является инструментом для оценки качества кластеризации и определения степени схожести объектов внутри кластеров. Значение силуэта позволяет судить о том, насколько хорошо объекты внутри кластера схожи между собой и насколько четко разделяются кластеры.

Формула для вычисления коэффициента силуэта имеет следующий вид:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (2)$$

где

$S(i)$ – силуэтный коэффициент для объекта i;

$a(i)$ – среднее расстояние между объектом i и всеми другими объектами в том же кластере;

$b(i)$ – наименьшее среднее расстояние между объектом i и объектами в кластере, к которому i не принадлежит.

Для проведения моделирования и максимизации коэффициента силуэта проводились эксперименты с различным количеством кластеров (5-15 кластеров) и различным минимальным количеством точек к одном кластере (5-20). Максимальное значение, которое получилось достичь, равно 0,63 при следующих параметрах: количество кластеров, равное 15 с параметрами ε = 0.5 и MinPts = 5, что свидетельствует о достаточной степени схожести объектов внутри кластеров и хорошем разделении кластеров. Общее значение силуэта оценивает, насколько хорошо объекты были сгруппированы в кластеры, а в контексте данного исследования это означает, что тексты внутри каждого кластера имеют достаточное семантическое сходство, в то время как тексты из разных кластеров значительно отличаются друг от друга. Такой уровень кластеризации является весьма приемлемым для анализа городских данных, подчеркивая эффективность выбранных методов для выделения информативных и географически релевантных паттернов в больших объемах текстовой информации. Это значение силуэта возможно улучшить путем обогащения данных. Интеграция дополнительных источников информации позволит сделать кластеризацию более точной: увеличить объем кластеров и более полно исследовать тематический контент и географические данные.

Когерентность для тем в модели LDA в данном исследовании вычислялась с использованием меры C_V когерентности [3]. Эта мера оценивает степень семантической связности слов в теме на основе их совместной встречаемости в текстах. Когерентность C_V для темы T может быть определена как:

$$C_V(T) = \frac{1}{M}\sum_{m=1}^{M}\frac{\sum_{i=2}^{N}\sum_{j=1}^{i-1}\text{score}(w_i.\, w_j)}{N(N-1)/2}, \quad (3)$$

где

M – количество топ-слов в теме;

N – количество слов, для которых вычисляется когерентность;

w_i и w_j – слова в теме;

score $(w_i,\ w_j)$ – мера семантической связности между парами слов, часто основанная на их совместной встречаемости или других статистических данных из корпуса.

Алгоритм вычисления когерентности включает выбор топ- N слов из каждой темы, вычисление score для каждой пары слов в этом наборе, и усреднение этих значений для

получения общей меры когерентности для темы. Затем среднее значение когерентности по всем темам используется для оценки качества модели LDA в целом.

Для проведения расчетов использовались различные параметры: количество тем варьировалось от 10 до 20, а количество слов – от 5 до 10. После моделирования с использованием меры C_V для оценки когерентности максимальное значение получилось 0,55, где было выявлено 10 наиболее выраженных тем, а каждая тема включает 6 слов.

В таблице 1 представлена информация о кластерах, их географических координатах (долгота и широта), и связанных с ними темах. Каждая строка таблицы соответствует отдельному кластеру, который имеет свой уникальный идентификатор (ID CLUSTER). Наименованием кластера (NAME) является именованная сущность, которая была выделена из текста, а координаты (LON и LAT) – являются результатом геокодирования. Темы, связанные с каждым кластером, отражают ключевые слова или концепции, выявленные в данных, связанных с этим кластером. В таблице собраны необходимые данные для более подробного анализа результатов кластеризации, что, в свою очередь, также помогает понять, какие темы преобладают в разных географических областях.

Таблица 1. Распределение тем по геокластерам

	NAME	LON	LAT	ID CLUSTER	TOPIC
0	Южный Всеволожск	30.648415	59.990431	19	пожар, ремонт, цены
1	Ленинградская область	29.608975	59.337017	13	суд, завод, празднования
2	Колтушское шоссе	30.647885	60.023717	19	штраф, авария, отдел
3	Ленобласть	29.608975	59.337017	13	эксперт, фестиваль
4	Мурино	30.438578	60.051284	21	происшествия, движение, квартира, пострадавший, мигрант, новый
...	...	...	...	...	
12	Тосненский район	31.017569	59.372039	9	авария, следователи
13	Петербург	30.315644	59.938955	5	почта, концерт, мчс, парковка, эрмитаж
14	Василеостровский район	30.248045	59.941430	5	рабочий, бизнес, тариф, дтп
15	Шушары	30.379523	59.807224	22	ремонт, долг, новый, метро

На рисунке 3 представлена карта, на которой отмечены кластеры, выделенные с использованием алгоритма. Кластеры представлены разными цветами и обозначают группировку точек схожих координат. Этот рисунок помогает визуализировать результаты кластеризации и выделить географические области, в которых события схожи или связаны между собой. Каждый кластер на рисунке 3 соответствует кластеру из таблицы 1.

В результате проведенного исследования разработан полноценный конвейер обработки данных, начиная со сбора информации из новостных ресурсов и завершая анализом распределения тем внутри географических кластеров. Применение разработанной методологии позволило получить тематический контекст по геокластерам на основе текстовых данных, полученных из новостных и социальных медийных источников. Визуализация тематического контекста выявила области на карте города, которые отображают восприятие жителями неформальных районов.

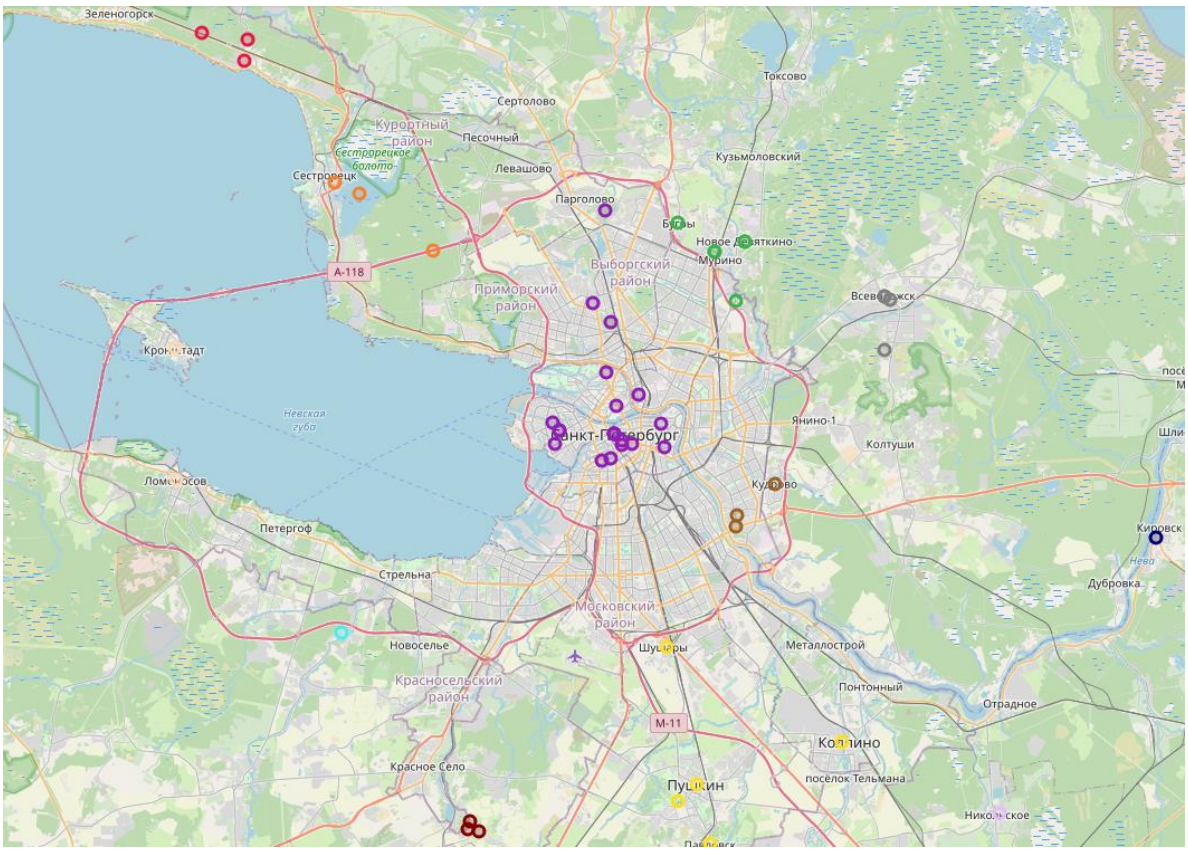

Рис. 3. Разделение географических точек на кластеры

Заключение. Таким образом, разработана методология анализа тематического контента на геокластерах, основанная на данных из социальных сетей и новостных источников. Представленный подход с учетом обобщенности и модульности позволяет применить его к различным источникам неструктурируемых текстовых данных. Использование методов геокодирования и кластеризации целесообразно применять для территорий, где нет возможности получить административное деление или необходимо более точечное и локальное исследование.

В работе продемонстрировано, что этапы агрегирования и сегментации могут успешно выделить темы в географически разбросанных данных. Полученные агрегированные темы могут быть использованы для поиска по ключевым словам и позволяют интерпретировать информацию, содержащуюся в данных. Это открывает возможности для создания рекомендательных систем, обнаружения интересов пользователей и анализа настроений сообщений.

В дальнейших исследованиях планируется расширить подход, добавив анализ настроений, что позволит определить характер новостей и текстов и, как следствие, создать более точный портрет районов, рекомендации для пользователей относительно мест и событий. Кроме того, метод будет протестирован на анализе рекламных сообщений, с целью создания тематического рекламного образа города. Текстовые посты в социальных сетях и новостные данные представляют собой богатый источник информации, который, после агрегирования и кластеризации, помогает выявить интересы пользователей и предоставить им соответствующие рекомендации и рекламу.

Полученные результаты полезны для градостроителей, руководства города и органов местного самоуправления, так как они могут служить основой системы поддержки принятия решений в условиях неопределенности для развития городской инфраструктуры, планирования мероприятий и оптимизации городских служб. Анализ пространственной

организации информации позволяет лучше понимать динамику событий и взаимосвязи между различными районами города.

Данная методология имеет широкий спектр практических применений, начиная от анализа пользовательских интересов в социальных сетях и заканчивая созданием интеллектуальных систем рекомендаций и анализа общественного мнения в различных сферах.

Список источников

1. Вампилова Л.Б. Регионы и города России: Атлас ментальных карт / Л.Б. Вампилова, В.Н. Калуцков, И.И. Митин, В.М. Матасов // Русское географическое общество. Электронное (сетевое) научное издание, 2018. – 130 с. – ISBN 978-5-600-02139-6.
2. Rüdiger M, Antons D, Joshi A.M, Salge T-O. Topic modeling revisited: New evidence on algorithm performance and quality metrics. PLoS ONE, 2022, no. 17(4), DOI:10.1371/journal.pone.0266325.
3. Jelodar H., Wang Y., Yuan C. [et al.]. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 2019, no. 78, pp. 15169-15211, DOI:10.1007/s11042-018-6894-4.
4. Bodrunova S.S. Topic modeling in Russia: current approaches and issues in methodology. The Palgrave handbook of digital Russia studies. Palgrave Macmillan, Cham, 2020, pp. 409-426, DOI:10.1007/978-3-030-42855-6_23.
5. Карпович С.Н. Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI / С.Н. Карпович // Труды СПИИРАН. – СПб.: СПб ФИЦ РАН, 2016. – Вып. 47(4). – С. 92–104. – URL: http://proceedings.spiiras.nw.ru/index.php/sp/article/view/3359/1942 (дата обращения: 25.10.2023).
6. Jones C.B., Alani H., Tudhope D. Geographical information retrieval with ontologies of place. In: Montello, D.R. (eds) Spatial Information Theory. COSIT 2001. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2001, vol. 2205. pp 322-335, DOI:10.1007/3-540-45424-1_22.
7. Тен Л.В. Тематическое моделирование в задаче автоматической рубрикации новостных текстов / Л.В. Тен // Terra Linguistica, 2023. – Т.14. – №2. – URL: https://cyberleninka.ru/article/n/tematicheskoe-modelirovanie-v-zadache-avtomaticheskoy-rubrikatsii-novostnyh-tekstov (дата обращения: 25.10.2023).
8. Wu X., Fang L., Wang P. [et al.]. Performance of using LDA for Chinese news text classification. 2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE), Halifax, NS, Canada, 2015, pp. 1260-1264, DOI:10.1109/CCECE.2015.7129459.
9. Mata-Rivera F., Torres-Ruiz M., Guzmán G. [et al.] A collaborative learning approach for geographic information retrieval based on social networks. Computers in Human Behavior, 2015, vol. 51(B), pp. 829–842, DOI:10.1016/j.chb.2014.11.069.
10. Nguyen M.D., Shin W.Y. an improved density-based approach to spatio-textual clustering on social media. IEEE Access 7, 2019, pp. 27217-27230, DOI:10.1109/ACCESS.2019.2896934.
11. Jiang B., Ma D., Yin J. [et al.] Spatial distribution of city tweets and their densities. Urban remote sensing: monitoring, synthesis, and modeling in the urban environment, second edition, 2021, pp. 115-129, DOI:10.1002/9781119625865.ch6.
12. Hu Y. Geo-text data and data-driven geospatial semantics. Geography Compass, 2018, DOI:10.1111/gec3.12404.
13. Dadwal R., Funke T., Demidova E. An adaptive clustering approach for accident prediction. IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 2021, pp. 1405-1411, DOI:10.1109/ITSC48978.2021.9564564.
14. Funkner A.A., Elkhovskaya L.O., Lenivtceva I.D. [et al.] Geographical topic modelling on spatial social network data // Procedia computer science, 2021, Vol. 193, pp. 22-31, DOI:10.1016/j.procs.2021.10.003.
15. Sander J., Ester M., Kriegel H-P, Xiaowei Xu. Density-Based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data mining and knowledge discovery, Berlin. Springer-Verlag, 1998, vol. 2, iss. 2, pp. 169-194, DOI:10.1023/A:1009745219419

Авдюшина Анна Евгеньевна. *Аспирант, факультет программной инженерии и компьютерной техники университет ИТМО, AuthorID: 57221719751, ORCID: 0000-0002-8235-902X, avdiushina@itmo.ru, 197101, Санкт-Петербург, Кронверкский проспект 49.*

Королева Юлия Александровна. *Доцент, к.т.н., факультет программной инженерии и компьютерной техники университета ИТМО, ORCID: 0000-0003-1462-1599, jakoroleva@itmo.ru, 197101, Россия, Санкт-Петербург, Кронверкский пр 49.*

Маркина Татьяна Анатольевна. *Доцент, к.т.н., факультет программной инженерии и компьютерной техники университета ИТМО, ORCID: 0009-0009-9146-433X, markina_t@itmo.ru, 197101, Россия, Санкт-Петербург, Кронверкский пр 49.*

Бессмертный Игорь Александрович. *Профессор, д.т.н. факультет программной инженерии и компьютерной техники университета ИТМО, AuthorID: 36661767800, ORCID: 0000-0001-6711-6399, bessmertny@itmo.ru, 197101, Россия, Санкт-Петербург, Кронверкский пр 49.*

UDC 004.89
DOI:10.25729/ESI.2024.33.1.014

Application of geographic information retrieval methods to analyze new's data

Anna E. Avdiushina, Yulia A. Koroleva, Tatyana A. Markina, Igor A. Bessmertny
ITMO University, Russia, St. Petersburg, *avdiushina@itmo.ru*

Abstract. The article focuses on identifying informal urban areas based on data from news sources and social networks, utilizing geographical proximity as a criterion. A methodology for extracting geodata from texts for spatial clustering is proposed. Geographic names extracted from texts are transformed into geolocations through geocoding. The identified geopoints are then clustered by density, and a distribution of themes is determined for each cluster. This approach allows for an abstraction from administrative divisions to reveal clusters that are closer to the citizens' perception. The clustering results are promising for application in various urban infrastructure management tasks: monitoring public life, analyzing the quality of the urban environment, and public safety. The distinction of the proposed methodology lies in the synthesis of geodata for grouping objects. The software tools developed based on this methodology enable decision-making in the field of urban planning, including the development of city districts and transport infrastructure, the placement of socially significant objects, and ensuring safety.

Keywords: data mining, smart city, information model, decision support systems, clustering, geospatial data

References

1. Vampilova L.B., Kalutskov V.N., Mitin I.I., Matasov V.M. Regions and cities of Russia: An atlas of mental maps. Russian geographical society. Elektronnoye (setevoye) nauchnoye izdaniye [Electronic (online) scientific publication], 2018, 130 p., ISBN 978-5-600-02139-6.
2. Rüdiger M, Antons D, Joshi A.M, Salge T-O. Topic modeling revisited: New evidence on algorithm performance and quality metrics. PLoS ONE, 2022, no. 17(4), DOI:10.1371/journal.pone.0266325.
3. Jelodar H., Wang Y., Yuan C. [et al.]. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 2019, no. 78, pp. 15169-15211, DOI:10.1007/s11042-018-6894-4.
4. Bodrunova S.S. Topic modeling in Russia: current approaches and issues in methodology. The Palgrave handbook of digital Russia studies. Palgrave Macmillan, Cham, 2020, pp. 409-426, DOI:10.1007/978-3-030-42855-6_23.
5. Karpovich S.N. Mnogoznachnaya klassifikaciya tekstovy`x dokumentov s ispol`zovaniem veroyatnostnogo tematicheskogo modelirovaniya ml-PLSI [Multi-label classification of text documents using probabilistic topic model ml-PLSI]. SPIIRAS Proceedings, 2016, iss. 4(47), pp. 92–104, available at: http://proceedings.spiiras.nw.ru/index.php/sp/article/view/3359/1942 (accessed: 10/20/2023).
6. Jones C.B., Alani H., Tudhope D. Geographical information retrieval with ontologies of place. In: Montello, D.R. (eds) Spatial information theory. COSIT 2001. Lecture notes in computer science, Springer, Berlin, Heidelberg, 2001, vol. 2205. pp 322-335, DOI:10.1007/3-540-45424-1_22.
7. Ten L.V. Tematicheskoe modelirovanie v zadache avtomaticheskoj rubrikacii novostny`x tekstov [Topic modeling in automatic categorization of news] Terra Linguistica, 2023, vol. 14, no. 2, DOI: 10.18721/JHSS.14207 (accessed: 10/25/2023).
8. Wu X., Fang L., Wang P. [et al.]. Performance of using LDA for Chinese news text classification. 2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE), Halifax, NS, Canada, 2015, pp. 1260-1264, DOI:10.1109/CCECE.2015.7129459 (accessed: 10/20/2023).
9. Mata-Rivera F., Torres-Ruiz M., Guzmán G. [et al.] A collaborative learning approach for geographic information retrieval based on social networks. Computers in Human Behavior, 2015, vol. 51(B), pp. 829-842, DOI:10.1016/j.chb.2014.11.069 (accessed: 10/20/2023).
10. Nguyen M.D., Shin W.Y. an improved density-based approach to spatio-textual clustering on social media. IEEE Access 7, 2019, pp. 27217-27230, DOI:10.1109/ACCESS.2019.2896934.

11. Jiang B., Ma D., Yin J. [et al.] Spatial distribution of city tweets and their densities. Urban remote sensing: monitoring, synthesis, and modeling in the urban environment, second edition, 2021, pp. 115-129, DOI:10.1002/9781119625865.ch6.
12. Hu Y. Geo-text data and data-driven geospatial semantics. Geography compass, 2018, DOI:10.1111/gec3.12404.
13. Dadwal R., Funke T., Demidova E. An adaptive clustering approach for accident prediction. IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 2021, pp. 1405-1411, DOI:10.1109/ITSC48978.2021.9564564.
14. Funkner A.A., Elkhovskaya L.O., Lenivtceva I.D. [et al.] Geographical topic modelling on spatial social network data // Procedia computer science, 2021, Vol. 193, pp. 22-31, DOI:10.1016/j.procs.2021.10.003.
15. Sander J., Ester M., Kriegel H-P, Xiaowei Xu. Density-Based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data mining and knowledge discovery, Berlin. Springer-Verlag, 1998, vol. 2, iss. 2, pp. 169-194, DOI:10.1023/A:1009745219419.

Avdyushina Anna Evgenievna. *Postgraduate student, Faculty of Software Engineering and Computer Technology, ITMO University, AuthorID: 57221719751, ORCID: 0000-0002-8235-902X, avdiushina@itmo.ru, 197101, St. Petersburg, Kronverksky Prospekt 49.*

Koroleva Yulia Alexandrovna. *Associate professor, Ph.D., faculty of software engineering and computer technology, ITMO University, ORCID: 0000-0003-1462-1599, jakoroleva@itmo.ru, 197101, Russia, St. Petersburg, Kronverksky Ave 49.*

Markina Tatyana Anatolyevna. *Associate professor, Ph.D., faculty of software engineering and computer technology, ITMO University, ORCID: 0009-0009-9146-433X, markina_t@itmo.ru, 197101, Russia, St. Petersburg, Kronverksky Ave 49.*

Bessmertny Igor Alexandrovich. *Professor, doctor of technical sciences faculty of software engineering and computer technology, ITMO University, AuthorID: 36661767800, ORCID: 0000-0001-6711-6399, bessmertny@itmo.ru, 197101, Russia, St. Petersburg, Kronverksky Ave 49.*

Статья поступила в редакцию 01.11.2023; одобрена после рецензирования 04.03.2024; принята к публикации 13.03.2024.

The article was submitted 11/01/2023; approved after reviewing 03/04/2024; accepted for publication 03/13/2024.