УДК 004.6:311.2:311.213

Подход к интегрированной обработке открытых данных об инфраструктуре города

Голубев А. В., Парыгин Д. С., Финогеев А. Г.

Постановка задачи: *поддержка принятия решений по управлению территорией города требует всестороннего анализа ситуации, который следует проводить с использованием объективной информации, характеризующей все составляющие элементы инфраструктуры, входящие в неё объекты и протекающие процессы. Однако необходимая информация в своей основе размещается в гетерогенных источниках, и, как следствие, ее использование требует разработки целого ряда решений для организации эффективного доступа.* ***Целью работы*** *является разработка комплексного подхода к обработке информации из гетерогенных источников, предоставляющих открытый доступ к данным об инфраструктуре города. Данный подход включает в себя взаимосвязанные процедуры сбора и предварительной обработки всего массива получаемых данных об объектах в городе из источников в сети Интернет. При этом источники классифицируются для применения соответствующих инструментов извлечения информации и организации многоуровневой системы ее хранения для целей последующего использования внешними сервисами.* ***Используемые методы:*** *решение многоуровневой задачи интеграции информации об инфраструктуре территории проводится с использованием методов интеллектуального анализа геопространственных данных, извлекаемых из сетевых ресурсов. Сбор данных реализуется с помощью алгоритмов извлечения данных с web-страниц (parser, grabber) и прямого доступа к API сервисов. Для предобработки изображений применяются подходы теории распознавания образов, в том числе, определения опорных точек и перцептивных хэш-алгоритмов. Извлечение дополнительных свойств объектов реализуется с использованием методов «Text Mining» и математической статистики.* ***Новизна*** *проведенного исследования заключается в разработанной модели интеграции разнородной информации и созданной на её основе последовательности обработки данных, реализующей принцип перекрывающего резервного хранения структурированной и исходной информации.* ***Результаты:*** *выделены ключевые источники информации о городской инфраструктуре, анализ которых позволил сформировать представление о структуризации данных об инфраструктуре территории и разработать модель их интеграции с учётом сквозной геопространственной привязки. В рамках реализации интеллектуального анализа геопространственных данных разработаны методы извлечения информации с сетевых ресурсов по объектам недвижимости, подход к сбору и обобщению данных с картографических сервисов, алгоритм извлечения дополнительных фактов об объектах инфраструктуры из записи на естественном языке на основе формальных грамматик, подход к валидации информации на основе сравнения изображений с помощью хэш-алгоритма, а также подход к определению схожести объявлений с использованием предложенной системы коэффициентов с помощью статистического анализа исходных данных по объектам недвижимости. Предложен подход к организации связанных файловых хранилищ и баз данных для хранения предобработанных, структурированных и исходных данных.* ***Практическая значимость*** *заключается в построении на основе разработанных подходов единой системы обработки информации, состоящей из модулей, реализующих операции парсинга онлайн-ресурсов и доступа к их API, анализа текстов на естественном языке, изображений и структурированных данных об объектах недвижимости и инфраструктуры города, а также создании SQL и NoSQL баз данных для хранения собранной информации наравне с ее размещением в виде файлов исходных данных.*

Библиографическая ссылка на статью:
Голубев А. В., Парыгин Д. С., Финогеев А. Г. Подход к интегрированной обработке открытых данных об инфраструктуре города // Системы управления, связи и безопасности. 2018. № 2. С. 84-107. URL: http://sccs.intelgr.com/archive/2018-02/06-Golubev.pdf

Reference for citation:
Golubev A. V., Parygin D. S., Finogeev A .G. The Approach to Integrated Processing of Open Data about the City Infrastructure. *Systems of Control, Communication and Security*, 2018, no. 2, pp. 84–107. Available at: http://sccs.intelgr.com/archive/2018-02/06-Golubev.pdf (in Russian).

Ключевые слова: *открытые данные; гетерогенные источники данных; сетевой ресурс; геопространственные данные; инфраструктура территории города; анализ данных; предобработка данных; объект недвижимости; обработка изображений; «сырые» данные; естественный язык; Открытое правительство; Data Mining; геоинформационная система; SQL; NoSQL.*

Введение

Актуальные данные о процессах в городской среде становятся одним из ключевых факторов обеспечения эффективного управления усложняющейся инфраструктурой расширяющихся урбанизированных территорий [1]. В условиях современных информационных технологий можно говорить о возможности получения данных, характеризующих состояния любых объектов. Благодаря встроенным системам «самоконтроля» технических объектов они сами способны накапливать и даже распространять информацию о своем состоянии [2]. Но более значимо, что благодаря развившимся системам сбора данных, источником информации становится не только техническая среда, но и вообще все процессы и состояния всех элементов сложных территориально-распределенных систем в городе [3, 4].

Работа с данными об инфраструктуре территорий в самом широком смысле выходит на новый уровень, становясь повседневным инструментом для обычных пользователей, а также базой для анализа ситуации и поддержки принятия решений для специалистов в долгосрочной перспективе [5]. Интеграция данных разных уровней и источников, обеспечение надежного доступа к ней и прозрачности представления является основой информированного устойчивого развития для целых городов [6].

В рамках данного исследования предлагается разработать модель интеграции разнородных данных, способную обеспечить унифицированное представление информации из гетерогенных источников, возможность расширения и настройки на новые источники, а также географическую привязку всех обрабатываемых данных. Кроме того, необходимо сформировать подходы к сетевому сбору открытых данных об инфраструктуре города, которые позволят реализовать технологии извлечения информации с онлайн-ресурсов.

1. Анализ открытых источников данных об инфраструктуре города

Принцип открытости данных основывается на организации их свободного доступа для использования. Открытые данные должны распространяться максимально широко через общие каналы связи, преимущественно сети Интернет, и не содержать ограничений на обработку, в том числе, для коммерческих и любых иных целей [3].

С точки зрения контента, открытые данные в первую очередь связаны с информацией, распространяемой государственными и муниципальными органами власти в рамках концепции «Открытое правительство». Однако любые сведения, будь то база данных торговой компании о потреблении различных видов товаров или данные использования мобильной сети, могут укладываться

в концепцию открытых данных, если они не нарушают тайну личной жизни или коммерческую тайну [7].

Отдельной задачей является соблюдение некоторых правил в отношении формата распространения данных. Информация должна быть интероперабельной и поставляться в машиночитаемом виде, пригодном для использования в автоматических системах обработки [8]. Хотя в широком смысле, любые данные, передаваемые по сети, пригодны для машинной обработки, даже если они загружаются и создаются пользователями в «ручном» режиме, и лишь определяют сложность алгоритмов работы с ними.

С учётом обозначенных подходов в определении открытых данных, можно выделить ряд ключевых источников информации о городской инфраструктуре:

1) «Открытое правительство» [9, 10]. Данные различного уровня детализации, относящиеся к определенным ведомствам, системам жизнеобеспечения, ситуации с инфраструктурой в целом и в ее отдельных компонентах;
2) социальные сети и микроблоги. Все современные онлайн-площадки для коммуникации являются поставщиками разнородной информации. Ключевым отличием разных соцсетей (Facebook, Вконтакте, Одноклассники, Twitter, др.) является их политика открытости, позволяющая получать определенный объем данных, относящихся к некоторой территории, через их программный интерфейс приложения (API) [11];
3) геоинформационные сервисы (ГИС). Цифровые ГИС-данные об объектах территории включают сведения об их местоположении и свойствах, пространственных и непространственных атрибутах. Такие данные можно получать как со специализированных сайтов, так и от пользователей различных мобильных приложений или онлайн-сервисов, использующих геометки [12];
4) статистика. Статистические сведения включают в себя данные Росстата и его региональных подразделений в России, а также иные источники официально распространяемой систематизированной информации, описывающей некоторые объекты наблюдения и их признаки. Данные обычно поставляются в файлах формата «.csv», «.xml», «.xlsx», «.json», «.geojson» [13];
5) коммерческие сетевые сервисы. Тематические и многопрофильные ресурсы сети Интернет, сайты магазинов, новостные агрегаторы, онлайн-карты, и т.д. хранят и распространяют большие объемы структурированной для собственных целей и разрозненной информации о конкретных объектах или состоянии инфраструктуры в целом;
6) естественный язык. Оцифрованные данные опросов населения, записи пользователей на естественном языке могут являться предметом машинного анализа («Text Mining») и содержать как эмоциональную окраску отношения, так и фактические сведения об описываемом объекте [14];

7) сенсоры. Данные удаленного измерения и сбора какой-либо информации, полученные с датчиков, могут фиксировать уровень шума, загрязнение воздуха, климатические изменения (осадки, ветер, влажность) [15, 16].

2. Модель интеграции гетерогенных данных

Доступные данные об инфраструктуре городов имеют разные типы хранения. В связи с этим, первоочередной задачей является унификация представления данных пользователям. Однако такое единообразие в первую очередь основано на объединении данных различных источников, заключающимся в интеграции исходной информации.

В основе интеграции лежит комплексный процесс, включающий извлечение данных из гетерогенных источников, их преобразование к виду, пригодному для хранения в определенной структуре, и загрузку в соответствующую базу или хранилище данных. Для реализации этого процесса в рамках проводимого исследования, а также с учётом специфики геопространственной информации, был предложен подход к интеграции разнородной информации (рис. 1).

Разработке модели предшествовало изучение имеющихся данных и сферы их возможного применения. Городская среда является пространственно обусловленной композицией сложных территориально-распределенных систем. Каждая такая система состоит из некоторого набора обладающих присущими им внутренними свойствами объектов. Поэтому центральным понятием, описывающим любую структурную единицу городского пространства, был выбран «Объект».

Использование такой единицы, как «Объект» допускает максимальную функциональность аналитической работы с данными, имеющими географическую привязку. При этом можно выделить четыре основные конфигурации («Точка», «Периметр», «Набор» и «Линия») базовой единицы территории, определяющих пространственную вложенность и границы представления на картографической основе.

Конфигурация объекта в виде точки может быть стационарного, подвижного или событийного типа и представляется единичной парой координат. Стационарный точечный объект описывает расположение недвижимого оборудования или элементов инфраструктуры, производственные точки, точки на земле и т.п. Подвижный точечный объект характеризует временную локацию заведомо предназначенных к перемещению объектов, техники, транспортных средств, людей, др. Событийный точечный объект описывает привязку фактических новостных, информационных или тревожных сообщений о происшествиях, авариях, преступлениях, нарушениях процессов функционирования или обеспечения требуемыми ресурсами, др.

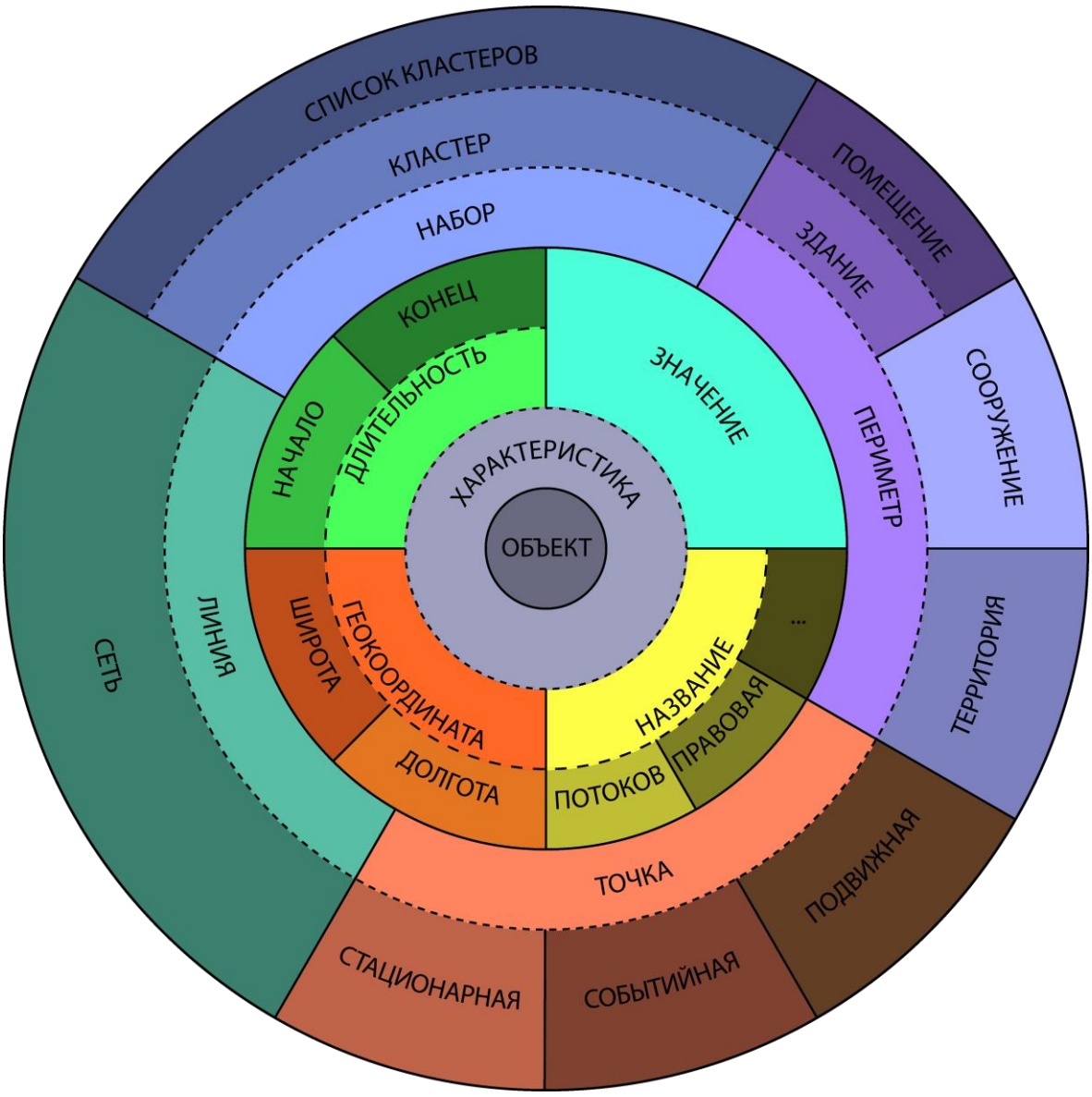

Рис. 1. Модель интеграции гетерогенных данных

Объекты в виде периметров описывают территориальные структуры, здания и сооружения. При этом территориальные или площадные структуры подразумевают, например, земельные ресурсы, зоны воздействия, населенные пункты, которые могут включать в себя и другие объекты разной конфигурации. «Сооружения» включают строения, не имеющие внутренних помещений или имеющие неэксплуатируемые внутренние пространства, главная ценность которых заключается в их целостном представлении, как то опоры ЛЭП, памятники, др. «Здания» являются объемными строениями с некоторым числом эксплуатируемых внутренних помещений, которые выделяются как самостоятельные вложенные или составляющие целое объекты.

Наборы точек описывают структурированные и хаотично размещенные на территории массивы объектов. Формализацией их представления являются кластеры, входящие в списки кластеров.

Линейные объекты описывают замкнутые, разомкнутые, пересекающиеся последовательности координат и другие структуры, одной из главных характе-

ристик которых является значение потока (ресурсов, людей, транспорта, др.), перемещающегося между связанными точками. Линии могут группироваться в сети. Примером таких конфигураций служат дороги, маршруты, линии электропередач, трубопроводы, коммунальные сети.

Каждый объект имеет собственные характеристики. «Характеристика» детерминируется значением, являющимся мерой состояния объекта, которое может изменяться с течение времени.

«Длительность» присуща всем характеристикам и определяется периодом реального времени. Для ряда характеристик такой период не будет иметь ограничений, обозначая постоянство конкретного свойства и указывая фактически на период существования самого объекта. При этом для характеристик процессов, значимых на протяжении некоторого промежутка времени, период их существования будет описываться начальным и конечным значением на шкале «дата-время».

Объекты любой конфигурации привязаны к местности через географические координаты. Справедливо также обратное утверждение, что каждая координата или группа координат являются некоторым объектом. Геокоордината всегда описывается парой «широта, долгота».

«Название» свойства является частью формализованного представления объекта в сети взаимовлияний. Так, например, «Правовая» характеристика определяет субъектно-объектный статус. Субъектность фиксируется правовым статусом индивидуальной или организационной принадлежности. В то же время для обозначения права собственности на объект справедливо применение характеристики, которая может обозначать постоянную или арендную имущественную принадлежность.

Предложенный подход к интеграции данных систематизирует хранение и формирует логику обработки разнородной информации о городской территории через ее пространственную привязку, определение ее места в структуре территории, указание временных и прочих характеристик. Созданная модель позволила перейти к разработке соответствующих инструментов для хранения получаемых из сети, извлекаемых и анализируемых данных различных открытых источников.

3. Подход к сбору и интеллектуальному анализу геопространственных данных об инфраструктуре города

Актуальность данных и точность сведений об объектах, которые они характеризуют, влияет на качество сделанных на их основе выводов и принятых впоследствии управленческих решений. Наиболее эффективным является анализ информации из первоисточника – ресурса, который генерирует или агрегирует исходные данные. При этом оптимальная скорость получения и обработки данных может быть достигнута, если ресурс находится в доступе онлайн.

Анализ многочисленных источников информации в Интернете является трудоемким процессом. Уже на стадии сбора необходимых данных приходится реализовывать специализированные решения. Для некоторых сайтов и задач существуют стандартизованные самими владельцами ресурсов инструменты,

распространяемые в форме API. Это особенно удобно для задач сбора больших данных и широко распространено для крупных площадок, таких как социальные сети, поисковые системы, новостные порталы и др.

Однако регулярные решения не покрывают всех возможных задач. Для многих сайтов они вообще не предлагаются. В таком случае необходимо применять технологии парсинга, включающие поиск и извлечение искомых данных со страниц сетевых ресурсов. Парсинг позволяет автоматизировать процесс сбора контента в режиме реального времени.

3.1. Сбор данных с онлайн-ресурсов

В рамках проведенного исследования на практике были изучены возможности библиотеки BeautifulSoup и программной платформы (фреймворка) Scrapy [17]. Оба эти решения способствуют реализации полного цикла сбора данных с сетевых ресурсов, однако имеют отличия с точки зрения технологии их применения.

Так, можно говорить, что за относительную простоту реализации решений на основе BeautifulSoup приходится расплачиваться ограниченным функционалом. По факту, данная библиотека позволяет лишь анализировать загруженный HTML-код и извлекать из него информацию. В то время как Scrapy, полноценная и мощная платформа с множеством дополнительных функций. К примеру, в Scrapy есть своя библиотека «scrapy-proxies», которая позволяет отправлять HTTP-запросы с использованием случайного прокси-сервера из списка [17].

Примечание. Применить одно из изученных решений было предложено студентам второго курса специальности «Информатика и вычислительная техника» ФГБОУ ВО «Волгоградский государственный технический университет» (ВолгГТУ). Больше 80% обучающихся выбрали для реализации подход на основе библиотеки BeautifulSoup, что по итогам оказалось оправданным и позволило всем в отведенное время при работе с разными сайтами создать функционирующие прототипы программ для сбора данных.

3.1.1. Парсинг сетевых ресурсов и социальных сетей

Оценка применимости различных подходов к сбору открытых данных об инфраструктуре территории проводилась на основе анализа популярных ресурсов сети Интернет, содержащих информацию в формате объявлений об операциях с объектами недвижимости, таких как «Яндекс.Недвижимость», «Квадрум», «Авито» и др. [18-27]. Отдельно были рассмотрены ведущие электронные торговые площадки России с информацией о выставленных в аукционных лотах объектах, такие как «РТС-тендер», «Сбербанк-АСТ» и др. [28-31], а также федеральный реестр данных по объектам недвижимости Росреестра [32].

Для каждого из выбранных источников была разработана структура данных, содержащихся в объявлениях и лотах. Рассматривались основные действия с недвижимостью (покупка или продажа, сдача аренду или найм) и типы объектов.

По полученным видам объявлений была проведена полная детализация данных. Такая процедура проводился в целях согласования порядка сбора информации с сайтов-источников.

Сравнение показало, что для всех сайтов получились разные структуры данных по объектам. Поэтому было решено для каждого из источников разработать отдельный модуль по сбору данных об объектах недвижимости на языке Python на основе упомянутых выше библиотеки BeautifulSoup или фреймворка Scrapy.

Результаты исследований и разработки модулей позволили сформировать единую процедуру работы с подобными онлайн-площадками. Получившийся порядок сбора данных был обобщен в виде метода парсинга сетевых ресурсов по объектам недвижимости, состоящего из следующих этапов [17]:

1) определяется «тяжелая» ссылка, перейдя по которой можно получить ссылки на объявления со всех регионов страны;
2) с использованием указателя ресурса составляются ссылки на объявления. Таким образом, парсер получает ссылки для дальнейших действий по сбору данных;
3) определяется количество страниц, на которых находятся ссылки на объявления. Это необходимо для определения количества итераций по получению ссылок на страницы объектов;
4) проводится постраничный сбор ссылок на объекты, а также выборка данных из заголовков объявлений;
5) после получения ссылок на все объекты, парсер начинает сбор оставшихся данных непосредственно со страниц объявлений;
6) все найденные данные на странице объявления заносятся в справочник, поля которого соответствуют основным характеристикам объекта;
7) данные сохраняются в файл формата JSON для последующей их обработки.

3.1.2. Парсинг картографических сервисов

Для получения комплексного описания объектов территории и организаций был разработан модуль сравнения данных картографических сервисов. В качестве источников входной информации использовались интерфейсы API картографических сервисов «Яндекс.Карты» и «Карты Google».

Собираемые данные обрабатывались согласно следующему принципу:

1) при наличии организации или объекта в обоих источниках, сведения о них обобщались и вносились в итоговую таблицу;
2) при наличии организации или объекта только в одном из источников, сохранялись имеющиеся по нему данные.

При этом сведения об организации, такие как номера телефонов, адреса, координаты, ссылки на официальные сайты приводились к общему виду. Так, для хранения адресной информации, был разработан единый нормализованный реестр. И каждая организация была отнесена к некоторой категории или категориям, соответствующим ее роду деятельности.

Получение данных об организациях с «Яндекс.Карт» возможно средствами компонента API «Яндекс.Организации». Сервис позволяет искать такие виды объектов, как дома, улицы, достопримечательности, кафе и другие объекты. Результаты поиска возвращаются в формате JSON или JSONP.

Средства API вводят ограничения на количество получаемых за один запрос данных (до 500 объектов) и общее количество запросов в сутки (всего до 500). Однако для оптимизации запросов возможно уточнение результатов за счёт ограничения области поиска в виде прямоугольной зоны, которая описывается угловыми географическими координатами, либо по радиусу вокруг точки поиска.

Поиск происходит как по имени организации, так и по ее категориям. Исходя из описанного механизма, было принято решение по использованию в качестве текстовых запросов имена категорий организаций, используемых в Яндекс. Например, для вывода перечня школ в городе Волгограде может быть использован следующий запрос:

```
https://search-
maps.yandex.ru/v1/?text=Общеобразовательная шко-
ла&bbox=44.10882555,48.41330622~44.68953248,48.8
8720444&rspn=1&type=biz&results=500&lang=ru_RU&a
pikey=КЛЮЧ_API_ЯНДЕКСА
```

Получаемый ответ в формате JSON состоит из нескольких разделов. В одном из них («features») находятся результаты поиска в виде вложенных объектов со сведениями об организациях, такими как название, адресная информация, перечень контактных данных, время работы, категории и т.д.

Получение данных с сервиса «Карты Google» в целом соответствует описанной выше процедуре для «Яндекс.Карт», включая, например, особенности ограничений на запросы. Итоги совместного сбора и сравнительной обработки данных обоих этих сервисов позволили отработать технологии получения ГИС-информации в виде, удобном для последующего комплексного исследования.

3.1.3. Парсинг ресурсов «Открытого правительства» и государственной статистики

Статистическая информация и ведомственные открытые данные содержат коррелирующие по содержанию сведения и, зачастую, имеют идентичную организацию порядка доступа к ним. Большая часть данных Росстата, а также данные, выкладываемые министерствами и госорганами регионов распространяются в форматах «.csv», «.xml», «.doc», «.xls» и др. Некоторым недостатком такого способа подачи является необходимость организации централизованной выгрузки с сетевых ресурсов большого количества разрозненных файлов и их предобработки, включая разбор и объединение данных в требуемые для дальнейшей работы структуры.

Однако часть данных «Открытого правительства» концентрируется в рамках специализированных автоматизированных систем, имеющих, в том числе, доступ через сайты в сети Интернет. Такие ресурсы могут предоставлять доступ к данным с использованием API. У этого способа также есть свои особенности, в частности, ограничения доступности серверов и, как следствие,

скорости получения данных, а также внутренняя политика доступа к информации.

Исследование такого класса ресурсов проводилось на примере портала «Реформа ЖКХ» [33]. Доступность данных тестировался с помощью официального API сайта «Реформа ЖКХ», а также с помощью сторонних программ-парсеров. При этом применение технологии парсинга в целом аналогично описанному выше подходу и показало свою эффективность, например, для задач сбора информации о домах конкретного региона.

Примечание. Для целей комплексного анализа жилой инфраструктуры в городах России более технологичным является комплексное получение данных с помощью API «Реформа ЖКХ». При этом имеется возможность сразу получать информацию по управляющим организациям и многоквартирным домам (МКД) в их управлении, получать данные о реализации региональных программ по переселению и мониторингу реализации региональных программ капитального ремонта.

В рамках взаимодействия ВолгГТУ и «Реформа ЖКХ» было получено официальное разрешение на чтение данных автоматизированной системы с помощью ее API. Обмен запросами между разработанным программным инструментом извлечения данных и системой «Реформа ЖКХ» был организован с помощью GET и SET методов согласно регламенту автоматизированной системы. Это обусловило специфику получения данных в соответствии со структурой хранения и доступа в системе «Реформа ЖКХ», позволив при этом получить масштабные выборки сразу целиком по нескольким регионам.

3.2. Обработка данных для выявления дополнительных свойств объектов инфраструктуры

3.2.1. Извлечение информации из текстов на естественном языке

Специфика данных о городской инфраструктуре требует учёта не только формализованных данных, представленных в структурированных выборках, но и информации от населения, получаемой зачастую в виде записей на естественном языке. Такая информация может оказаться наиболее подробной, проверенной и/или актуальной для конкретной ситуации.

Примеры таких данных об объектах инфраструктуры содержатся в не структурированном виде в пользовательских описаниях в социальных сетях и на тематических сайтах в объявлениях об операциях с недвижимостью. В связи с этим решалась задача по извлечению данных об объектах недвижимости из их описания в текстах на естественном языке. Соответственно необходимо было выбрать технологию семантического анализа содержащейся в них информации и разработать алгоритм её структуризации.

Анализ существующих решений показал соответствие технологии, реализованной в Томита-парсере, поставленным задачам извлечения структурированного описания (фактов). Вычленение фактов происходит при помощи контекстно-свободных грамматик и словарей ключевых слов. Парсер позволяет писать свои грамматики и добавлять словари для нужного языка [34].

Разработанный алгоритм, учитывающий функциональные возможности Томита-парсера, позволяет организовать автоматизированное извлечение фактов из записи на естественном языке в соответствии с имеющимися формальными грамматиками. Полученная из записи информация структурируется. В качестве результатов обработки записей получается описание объекта с извлеченными из текста фактами [14].

При создании алгоритма извлечения описания объектов недвижимости были решены задачи по обработке структуры и состава данных в записях. В качестве входных данных было организовано получение предварительно обработанных записей без знаков препинания в виде текстовых файлов.

Анализ объявлений об операциях с объектами недвижимости показал отсутствие четкой структуры в записях. Поэтому было предусмотрено при отсутствии некоторых фактов сохранять значения по умолчанию для соответствующих им полей описания объекта.

При этом были выделены несколько ключевых фактов, для которых в любом случае необходимо обнаружить значения в записях, чтобы полученное описание объекта было пригодно для дальнейшего анализа. Такими фактами являются: «Адрес», «Цена», «Площадь», «Район города» и «Количество комнат». В итоге, для поставленных условий определения в текстах записей фактов, задача написания формальных грамматик Томита-парсера была успешно решена.

3.2.2. Валидация информации на основе сравнения изображений

Вопрос валидации получаемых из сети данных неизменно возникает в процессе работы с любой информацией, а особенно свободно размещаемой пользователями. Речь идет о проверке одного из ключевых свойств информации – актуальности данных, получаемых из первоисточников. Так, например, часть объявлений об объектах недвижимости содержит заведомо поддельные сведения. Такие объявления размещаются в целях коммерческой рекламы, мошенничества или в результате халатного подхода к внесению данных. В связи с эти необходимо определять некоторые ключевые признаки, по которым можно было бы с определенной долей уверенности идентифицировать уникальность и достоверность каждого отдельного объявления.

Совокупность отдельных обозначаемых в объявлениях фактов об описываемом объекте недвижимости в тестовом формате может выступать в качестве валидационных признаков. Однако использование нескольких таких фактов в сочетании с проверкой оригинальности представленных в объявлении фотографий объекта недвижимости может оказаться более эффективным подходом. Наличие двух объявлений с одинаковыми фотографиями, но, например, различными контактами владельца или данными о состоянии описываемого объекта, является поводом для подозрений относительно подлинности одного из них. Соответственно, было решено в рамках данного исследования рассмотреть способы реализации процедуры сравнения изображений для применения в рамках единого процесса предобработки для валидации информации по объектам недвижимости.

На данный момент существует значительное количество алгоритмов сравнения изображений [35]. Они отличаются друг от друга методами получения конечного результата. И данное исследование проводилось для рассмотрения нескольких из них: методов на основе «опорных точек» и перцептивных методов.

Было выбрано решение, основанное на подсчете хэшей сравниваемых изображений. Такое решение является более простым в реализации и быстрым в работе, чем разработки на основе вычисления опорных точек изображений. При этом степень достоверности получаемых результатов является довольно высокой [36], хоть перцептивные методы и уступают по точности методам на основе «опорных точек». В итоге хэш-метод позволил с достаточной точностью определять, являются ли фотографии, прикрепленные к каждому конкретному объявлению, копиями фотографий из других объявлений.

Фотографии из источников не хранятся в созданной системе для обработки данных с сетевых ресурсов. Доступ к ним реализуется по прямым ссылкам, получаемым в результате парсинга. Поэтому используемый подход к сравнению изображений удовлетворяет ряду требований и, в первую очередь, позволяет соблюдать условия по скорости обработки большого потока визуальных данных, получаемых непосредственно с сетевых ресурсов.

Кроме того, было учтено, что пользователи, размещающие ложные объявления, могут искажать прикрепляемые к ним оригинальные изображения с целью обеспечения их псевдо уникальности. Анализ обрабатываемых источников объявлений позволил выявить искажения типа размещения водяных знаков, цветокоррекции, обрезки, деформации и др.

Примечание. Тестирование вариантов алгоритмов хэш-семейства проводилось на выборке из 2150 изображений, полученных с сайтов «МирКвартир» [21] и «ЦИАН» [22]. Каждое из изображений подвергалось ряду основных видов искажения (поворот на 3 градуса, обрезка на 5%, цветокоррекция и зеркальное отражение). Результат каждого отдельного действия сравнивался с исходным изображением. При этом на каждом этапе сравнения задавалось различное расстояние Хэмминга между хэшами, при котором изображения признавались одинаковыми.

Результаты исследования позволили сделать вывод, что наилучшим вариантом для использования является алгоритм pHash с порогом сравнения от 0 до 15. Выбранный алгоритм дает наибольшее количество успешно распознанных случаев искажения при относительно небольшом числе ложноположительных совпадений для заданной строгости сравнения.

3.2.3. Математические методы обработки и анализа

При решении задачи валидации данных об объектах инфраструктуры было решено не останавливаться на сравнительном анализе изображений. Кроме того, необходимо было по возможности точно выявлять источники не ложной, но дублирующей информации, общую корректность данных, а также их распределение при фильтрации по различным параметрам. С этой целью были исследованы методы статистического анализа.

Последовательность решения поставленной задачи включала следующие этапы:

1) считывание данных в виде, выгружаемом с сетевых ресурсов;
2) определение в описаниях объектов параметров, которые позволят выявлять уникальность записей, а также таких параметров, изменение которых влияет на изменение цены объектов. При этом стоимостную оценку объектов предполагается использовать как универсальную характеристика для их сопоставления;
3) исключение параметров, которые не будут учитываться в работе, т.к. оказывают незначительное влияние при формировании цены объекта или не позволяют однозначно определять уникальность записей;
4) создание алгоритмов определения дубликатов описаний. Для анализа уникальности описаний объектов недвижимости, среди всех параметров были выбраны следующие:
   - ссылки на изображения;
   - описание;
   - цена;
   - данные владельца (номер телефона и ФИО);
   - площадь;
   - адрес (район, улица, дом, этаж);
   - количество комнат;
   - год постройки.

   Определение схожести объявлений основано на предложенной системе коэффициентов, определяющих значимость параметров. Совпадение каждого из параметров двух объявлений увеличивает значение схожесть на определенную величину. Если схожесть будет выше определенного порогового значения, то объявления считаются одинаковыми и одно из них удаляется из выборки;
5) выгрузка данных в формате, удобном для чтения и обработки с помощью инструментария интерпретируемых программных языков (Python, R, др.). Такие языки удобны в частности тем, что зачастую имеют разработанные все необходимые математические методы;
6) кластеризация объявлений и определение тех, которые нельзя отнести к какому-либо кластеру с высокой доверительной вероятностью. Для кластеризации было принято решения использовать алгоритм нечеткой логики. Выборку предлагается разбивать на 3 кластера – кластеры с регулярной ценой, заниженной и завышенной;
7) проверка зависимости стоимости объектов от выбранных параметров и выявление наиболее значимых из них. На данном этапе исследования проверку предлагается проводить с помощью линейной регрессии;
8) анализ цен на объекты недвижимости в выборке и определение выбросов значений.

3.3. Организация хранения данных из гетерогенных источников

Инструменты сбора данных сетевых ресурсов, а также предобработчики исходной информации являются поставщиками разнородных данных, требующих организации их хранения. Разнообразие поступающих данных различных форматов и способов получения информации определили необходимость гибридного подхода на основе совместного использования NoSQL и SQL решений, а также хранилищ сырых данных.

На рис. 2 показан разработанный подход к последовательной обработке исходной информации. Передача данных в процессе их сбора и предобработки происходит, согласно приведенной схеме, от центра к окраинам представленной окружности путем перетекания через смежные области, т.е. пересечение способов их хранения. При этом данные из открытых источников (OD) классифицируются по принадлежности к основным типам ресурсов.

3.3.1. Онлайн-ресурсы и социальные сети

Основным способом хранения данных этих ресурсов был выбран метод сохранения «сырых» данных (RD), получаемых после программ парсинга и сохраняемых в файлах различных форматов на дисковой системе. Эти данные не имеют чёткой формализованной структуры, но они должны соответствовать формату, который будет иметь машиночитаемый вид.

Отличительной особенностью этого вида хранения является отсутствие возможности получить быстрый доступ к нужному блоку данных. Однако эти данные могут иметь документоориентированный вид. Основными форматами хранения «сырых» данных были выбраны «.json», «.csv» и «.xml».

В соответствии с разработанной моделью интеграции разнородной информации, все собранные и сохраненные данные об инфраструктуре территории должны, так или иначе, быть приведены к единой структуре и сохранены в базе данных. Для этого полученная информация в большинстве случаев проходит различные виды предобработки на предмет выявления дополнительных свойств с помощью разбора естественного языка (TP) или математических методов (MM), а также исключения дубликатов с помощью анализа изображений (HA) или MM.

Таким образом, «сырые» данные в основном являются промежуточной стадией между работой программ сборщиков и базами данных SQL или NoSQL вида. При этом в разработанных решениях для сбора данных из описанных далее источников, «сырые» данные используются для дублирующего резервного хранения одновременно с записью в базы данных.

3.3.2. Статистические данные и данные «Открытого правительства»

Работа в автоматизированном режиме с государственными информационными системами определила необходимость использования NoSQL решения. При реализации выбор был сделан в пользу MongoDB.

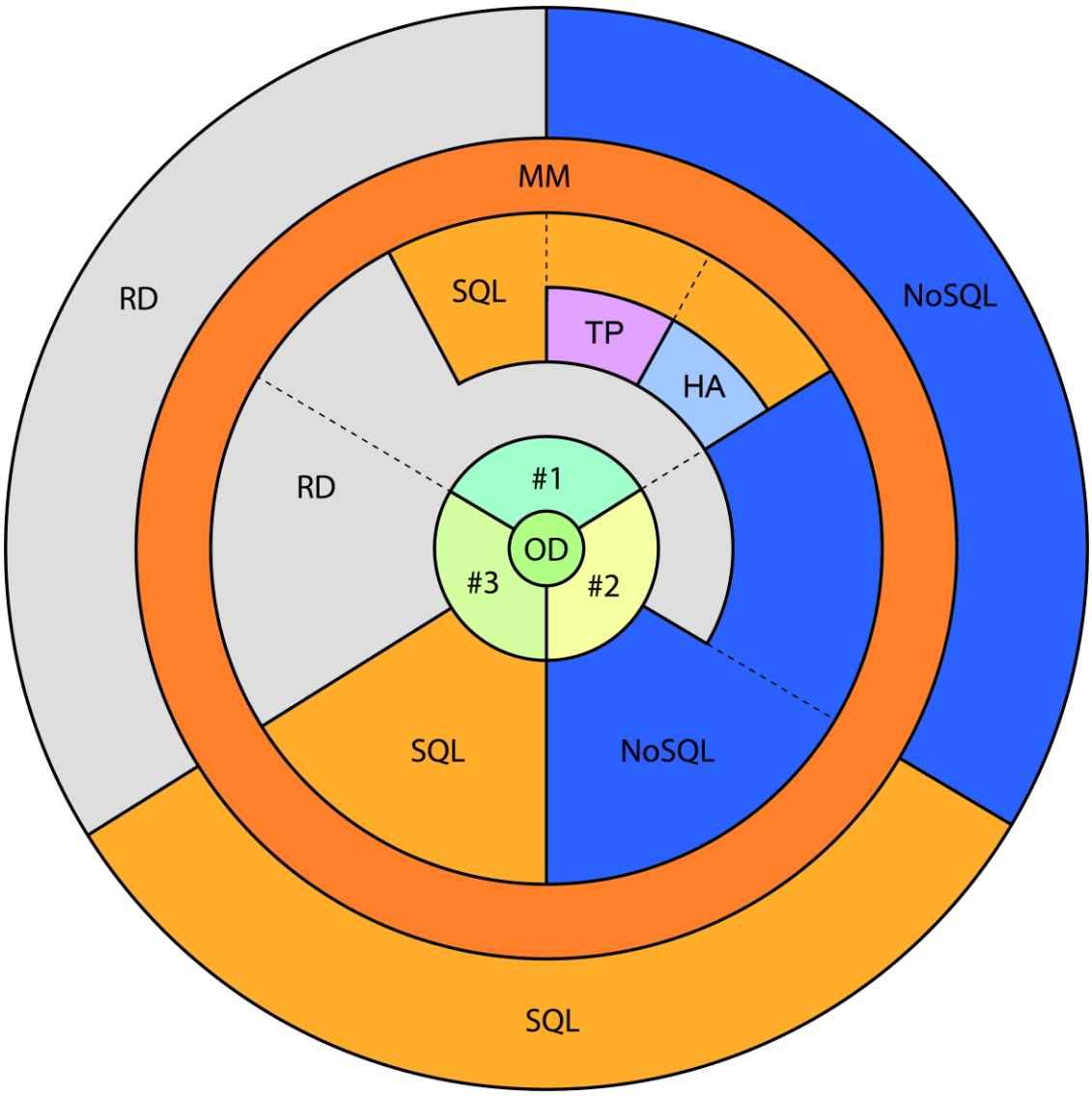

Рис. 2. Последовательность обработки данных (из центра к окраинам): OD – Open Data (от англ. «открытые данные»); #1 – онлайн-ресурсы и социальные сети; #2 – открытые правительственные данные; #3 – онлайн-карты и ГИС; RD – Raw Data (от анг. «сырые» данные"); TP – Tomita Parser; HA – хэш-алгоритм; MM – математические методы

Такое решение обусловлено тем, что структура получаемых документов, например, с ресурса «Реформа ЖКХ» не имеет чётко выраженной структуры. При этом использование базы данных на основе NoSQL не накладывает ограничений на типы хранимых данных и позволяет добавлять новые типы в процессе работы.

Кроме того, необходимо было придерживаться методологии быстрой разработки приложений (RAD). NoSQL базы данных не нуждаются в том же объёме подготовительных действий, которые обычно нужны для реляционных баз. Однако и отказаться вообще от использования базы данных в этом случае было нельзя, так как требовалось обеспечить контроль получаемых от источника коллекций данных.

3.3.3. Онлайн-карты и ГИС

Несмотря на то, что NoSQL-базы удобны благодаря быстродействию и хорошей масштабируемости, для ряда задач выбор был сделан в пользу структурированных SQL-хранилищ.

Жёсткое определение порядка взаимодействия транзакций с базой данных позволяет уменьшить вероятность неожиданного поведения системы и обеспечить целостность данных. А чёткая структура конечного результата, например, в задаче сравнения данных нескольких картографических сервисов позволила подготовить прозрачную структуру для хранения предобработанных данных.

Выводы

Управление данными в аналитических задачах предъявляет особые требования к их качеству и подаче. Разработчики сложных программных систем хотят видеть на входе чёткие требования к интерфейсу API, которые позволят им организовать прозрачное и быстрое взаимодействие для доступа к «чистым» и качественным данным. Однако в основе технологичных пользовательских решений лежит огромный пласт работ с данными от их источников до хранилищ, включая рутинные процедуры сбора информации из сети в условиях постоянно меняющихся политик доступа, трудности переформатирования и извлечения знаний, исключительные требования по надежности сохранения.

В проведенном исследовании рассмотрен комплексный подход к интегрированной обработке открытых данных. Предложены решения, учитывающие всю последовательность шагов, которую необходимо выполнить для сбора данных из гетерогенных источников в сети Интернет и их многоуровневой предварительной обработки.

Разработанная модель интеграции разнородных данных об объектах инфраструктуры территории позволила реализовать компоненты представленной последовательности обработки информации. Взаимосвязанная система хранения данных в базах и файловых хранилищах создает условия для формирования структурированных массивов данных, пополняемых и восстанавливаемых по требованию из независимо сохраняемых исходных наборов.

Сбор и анализ неструктурированных открытых данных о городской среде, а также организация их хранение входят в число приоритетных задач, которые должны решаться в цикле постоянного улучшения. Описанные подходы и технологии способны стать основой многих инструментов, востребованных уже сегодня и набирающих свою популярность, в том числе, для реализации методов автоматизированной оценки недвижимости, интеллектуального анализа и оценки инфраструктуры городской территории с учётом большого количества факторов, прогнозной аналитики для организации процессов управления и много другого.

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 17-37-50033 «мол_нр» и № 16-07-00388 «а».

Литература

1. Парыгин Д. С., Камаев В. А., Садовникова Н. П., Миронов А. Ю. Концепция информационно-аналитической системы управления развитием города // Инновационные информационные технологии: материалы Международной научно-практической конференции (Прага, Чехия, 22-26 апреля 2013 г.). – М.: МИЭМ НИУ ВШЭ, 2013. – Т. 4. – С. 205-213.

2. Барабанова Е. А., Мальцева Н. С. Коммутационные системы с параллельной обработкой. – Астрахань: АГТУ, 2012. – 164 с.

3. Парыгин Д. С. Модель интеркоммуникационной системы обеспечения потребностей жителей города // Известия Волгоградского государственного технического университета. Серия: Актуальные проблемы управления, вычислительной техники и информатики в технических системах. 2013. № 14 (117). С. 90-95.

4. Parygin D., Sadovnikova N., Kravets A., Gnedkova E. Cognitive and ontological modeling for decision support in the tasks of the urban transportation system development management // IISA 2015: proceedings of the Sixth International IEEE Conference on Information, Intelligence, Systems and Applications (Greece, Corfu, 6-8 July 2015). – IEEE, 2015. – P. 1-5.

5. Парыгин Д. С., Садовникова Н. П., Жидкова Н. П. Построение траекторий территориального развития на основе методов сценарного прогнозирования // Интернет-вестник ВолгГАСУ. Серия: Строительная информатика. 2012. № 8 (24). С. 1-9. – URL: http://vestnik.vgasu.ru/attachments/ParyginSadovnikova-2012_8(24).pdf (дата обращения: 10.03.2018).

6. Sadovnikova N., Parygin D., Gnedkova E., Sanzhapov B., Gidkova N. Evaluating the sustainability of Volgograd // The Sustainable City VIII: proceedings of the Eight International Conference on Urban Regeneration and Sustainability (Malaysia, Putrajaya, 3-5 December 2013). – WIT Press, 2013. – P. 279-290.

7. Парыгин Д. С., Садовникова Н. П., Шабалина О. А. Информационно-аналитическая поддержка задач управления городом. – Волгоград: ВолгГТУ, 2017. – 116 с.

8. Береснева Е. OPENDATA: зачем нужны открытые данные и что это такое // Электронное периодическое издание «Научная Россия» [Электронный ресурс]. 13.11.2014. – URL: https://scientificrussia.ru/articles/opendata-zachem-nam-nuzhny-otkrytye-dannye (дата обращения: 07.03.2018).

9. Открытые данные // Роскомнадзор [Электронный ресурс]. 14.03.2018. – URL: https://rkn.gov.ru/opendata/ (дата обращения: 14.03.2018).

10. Портал открытых данных Российской Федерации [Электронный ресурс]. 01.03.2018. – URL: http://data.gov.ru/ (дата обращения: 01.03.2018).

11. Russell M. A. Mining the Social Web. Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. – O'Reilly Media, 2013. – 448 p.

12. Беда Д. В., Гуртяков А. С., Парыгин Д. С., Потапова Т. А. Web-сервис для планирования туристических маршрутов на основе предпочтений

пользователей и геоданных о городских объектах // Известия Волгоградского государственного технического университета. Серия: Актуальные проблемы управления, вычислительной техники и информатики в технических системах. 2017. № 8 (203). С. 13-17.

13. Открытые данные // Федеральная служба государственной статистики [Электронный ресурс]. 17.02.2018. – URL: http://www.gks.ru/opendata/dataset (дата обращения: 17.02.2018).

14. Зеленский И. С., Донченко Д. С., Парыгин Д. С., Дегтяренко Д. Р., Петрова Т. М. Извлечение структурированного описания объектов недвижимости из пользовательских записей на естественном языке // Известия Волгоградского государственного технического университета. Серия: Актуальные проблемы управления, вычислительной техники и информатики в технических системах. 2017. № 14 (209). С. 41-46.

15. Источники метеорологических данных на территорию РФ по станциям [Электронный ресурс]. 15.05.2014. – URL: http://gis-lab.info/qa/meteo-station-sources.html (дата обращения: 05.03.2018).

16. Барабанова Е. А., Фролова А. В. Сенсорная телекоммуникационная система для мониторинга магистральных продуктопроводов. Наука, образование, инновации: пути развития: материалы Седьмой всероссийской научно-практической конференции (Петропавловск-Камчатский, 16-19 мая 2016 г.). – Петропавловск-Камчатский, 2016. – С. 98-100.

17. Cherkesov V., Malikov V., Golubev A., Parygin D., Smykovskaya T. Parsing of Data on Real Estate Objects from Network Resource // Advances in Computer Science Research: proceedings of the IV International research conference «Information technologies in Science, Management, Social sphere and Medicine» (Russia, Tomsk, 5-8 December 2017). 2017. Vol. 72. P. 385-388.

18. Доска объявлений от частных лиц и компаний на Avito [Электронный ресурс]. 12.03.2018. – URL: https://www.avito.ru/ (дата обращения: 12.03.2018).

19. Из рук в руки [Электронный ресурс]. 10.01.2018. – URL: http://irr.ru/ (дата обращения: 10.01.2018).

20. Портал объявлений по недвижимости [Электронный ресурс]. 04.01.2018. – URL: http://realty.dmir.ru/ (дата обращения: 04.01.2018).

21. Мир Квартир. Недвижимость [Электронный ресурс]. 14.02.2018. – URL: http://mirkvartir.ru/ (дата обращения: 14.02.2018).

22. ЦИАН. База недвижимости [Электронный ресурс]. 03.02.2018. – URL: https://www.cian.ru/ (дата обращения: 03.02.2018).

23. Яндекс.Недвижимость [Электронный ресурс]. 09.02.2018. – URL: https://realty.yandex.ru/ (дата обращения: 09.02.2018).

24. Квадрум [Электронный ресурс]. 02.03.2018. – URL: https://kvadroom.ru/ (дата обращения: 02.03.2018).

25. Продажа и аренда недвижимости. Недвижимость Mail.Ru [Электронный ресурс]. 16.01.2018. – URL: https://realty.mail.ru/ (дата обращения: 16.01.2018).

26. Недвижимость – продажа и аренда квартир, домов, комнат, офисов, цены на недвижимость. Домино [Электронный ресурс]. 22.01.2018. – URL: http://domino-rf.ru/nedvizimost/ (дата обращения: 22.01.2018).

27. Портал недвижимости Москвы и Санкт-Петербурга, Ленинградской и Московской областей [Электронный ресурс]. 20.02.2018. – URL: http://www.restate.ru/ (дата обращения: 20.02.2018).

28. Реестр имущественных торгов [Электронный ресурс]. 11.01.2018. – URL: https://www.etp-torgi.ru/market/ (дата обращения: 11.01.2018).

29. Торги по приватизации, аренде и продаже имущества [Электронный ресурс]. 19.01.2018. – URL: https://i.rts-tender.ru/main/auction/Trade/Search.aspx (дата обращения: 19.01.2018).

30. Инвестиционный портал города Москвы [Электронный ресурс]. 30.01.2018. – URL: https://investmoscow.ru/tenders/ (дата обращения: 30.01.2018).

31. Сбербанк-АСТ. Автоматизированная система торгов [Электронный ресурс]. 14.01.2018. – URL: http://www.sberbank-ast.ru/ (дата обращения: 14.01.2018).

32. Портал услуг Федеральной службы государственной регистрации, кадастра и картографии [Электронный ресурс]. 14.02.2018. – URL: https://portal.rosreestr.ru/ (дата обращения: 14.02.2018).

33. Реформа ЖКХ [Электронный ресурс]. 11.03.2018. – URL: https://www.reformagkh.ru/ (дата обращения: 11.03.2018).

34. Томита-парсер. Видеокурс [Электронный ресурс]. 07.01.2018. – URL: https://tech.yandex.ru/tomita/doc/video/index-docpage/ (дата обращения: 07.01.2018).

35. Простой и быстрый метод сравнения изображений для сходства // stack.io [Электронный ресурс]. 16.11.2010. – URL: http://qaru.site/questions/42410/simple-and-fast-method-to-compare-images-for-similarity (дата обращения: 08.03.2018).

36. Zauner C. Implementation and Benchmarking of Perceptual Image Hash Functions // pHash [Электронный ресурс]. 22.07.2010. – 108 p. – URL: https://www.phash.org/docs/pubs/thesis_zauner.pdf (дата обращения: 16.02.2018).

References

1. Parygin D. S., Kamaev V. A., Sadovnikova N. P., Mironov A. Yu. Koncepciya informacionno-analiticheskoj sistemy upravleniya razvitiem goroda [The concept of information-analytical system of management urban development]. *Materialy Mezhdunarodnoy nauchno-prakticheskoy konferentsii «Inno-vatsionnyye informatsionnyye tekhnologii»* [Materials of the International Scientific and Practical Conference «Innovative information technologies»]. Czech Republic, Prague, April 22-26, 2013. Moscow, Moscow Institute of Electronics and Mathematics A.N. Tikhonov, 2013, vol. 4, pp. 205-213 (in Russian).

2. Baranova E. A., Malceva N. S. *Kommutacionnye sistemy s parallelnoj obrabotkoj* [Switching systems with parallel processing]. Astrakhan, Astrakhan State Technical University, 2012. 164 p. (in Russian).

3. Parygin D. S. Model interkommunikacionnoj sistemy obespecheniya potrebnostej zhitelej goroda [Model intercommunication system of ensuring needs of city inhabitants]. *Izvestiya Volgograd State Technical University. Series Actual problems of management, computing hardware and informatics in engineering systems*, 2013, no. 14 (117), pp. 90-95 (in Russian).

4. Parygin D., Sadovnikova N., Kravets A., Gnedkova E. Cognitive and ontological modeling for decision support in the tasks of the urban transportation system development management. *Proceedings of the Sixth International IEEE Conference on Information, Intelligence, Systems and Applications*. Corfu, Greece, July 6-8, 2015. IEEE, 2015, pp. 1-5.

5. Parygin D. S., Sadovnikova N. P., Zhidkova N. P. Postroenie traektorij territorialnogo razvitiya na osnove metodov scenarnogo prognozirovaniya [Construction of territorial development trajectories based on methods of scenario forecasting]. *Internet-vestnik VolgGASU. Seriya Stroitel'naya informatika*, 2012, no. 8 (24). Available at: http://vestnik.vgasu.ru/attachments/ParyginSadovnikova-2012_8(24).pdf (accessed 10 March 2018) (in Russian).

6. Sadovnikova N., Parygin D., Gnedkova E., Sanzhapov B., Gidkova N. Evaluating the sustainability of Volgograd. *Proceedings of the Eight International Conference on Urban Regeneration and Sustainability «The Sustainable City VIII»*. Malaysia, Putrajaya, December 3-5, 2013. WIT Press, 2013, pp. 279-290.

7. Parygin D. S., Sadovnikova N. P., Shabalina O. A. *Informacionno-analiticheskaya podderzhka zadach upravleniya gorodom* [Informational and analytical support of city management tasks]. Volgograd, 2017. 116 p. (in Russian).

8. Beresneva E. *OPENDATA: why open data is needed and what is it*. Available at: https://scientificrussia.ru/articles/opendata-zachem-nam-nuzhny-otkrytye-dannye (accessed 7 March 2018) (in Russian).

9. Open data. *Roskomnadzor*. Available at: https://rkn.gov.ru/opendata/ (accessed 14 March 2018) (in Russian).

10. *Open Data Portal of the Russian Federation*. Available at: http://data.gov.ru/ (accessed 1 March 2018) (in Russian).

11. Russell M. A. *Mining the Social Web. Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, 2013. 448 p.

12. Beda D. V., Gurtyakov A. S., Parygin D. S., Potapova T. A. Web-servis dlya planirovaniya turisticheskih marshrutov na osnove predpochtenij polzovatelej i geodannyh o gorodskih obektah [Web-service for planning tourist routes based on user preferences and the geodata of urban sites]. *Izvestiya Volgograd State Technical University. Series Actual problems of management, computing hardware and informatics in engineering systems*, 2017, no. 8 (203), pp. 13-17 (in Russian).

13. Open Data. *Federal State Statistics Service*. Available at: http://www.gks.ru/opendata/dataset (accessed 17 February 2018) (in Russian).

14. Zelenskiy I. S., Donchenko D. S., Parygin D. S., Degtyarenko D. R., Petrova T. M. Izvlechenie strukturirovannogo opisaniya obektov nedvizhimosti iz polzovatelskih zapisej na estestvennom yazyke [Extracting a structured description of real estate objects from user posts in natural language]. *Izvestiya Volgograd State*

Technical University. Series Actual problems of management, computing hardware and informatics in engineering systems, 2017, no. 14 (209), pp. 41-46 (in Russian).

15. *Sources of meteorological data on the territory of the Russian Federation by stations*. Available at: http://gis-lab.info/qa/meteo-station-sources.html (accessed 5 March 2018) (in Russian).

16. Barabanova E. A., Frolova A. V. Sensornaya telekommunikacionnaya sistema dlya monitoringa magistralnyh produktoprovodov [Sensory telecommunication system for monitoring of main product pipelines]. *Materialy Sed'moy vserossiyskoy nauchno-prakticheskoy konferentsii «Nauka, obrazovaniye, innovatsii: puti razvitiya»* [Materials of the Seventh All-Russian Scientific and Practical Conference «Science, Education, Innovation: Ways of Development»]. Petropavlovsk-Kamchatsky, May 16-19, 2016. Petropavlovsk-Kamchatsky, Kamchatka State Technical University, 2016, pp. 98-100 (in Russian).

17. Cherkesov V., Malikov V., Golubev A., Parygin D., Smykovskaya T. Parsing of Data on Real Estate Objects from Network Resource. *Proceedings of the IV International research conference «Information technologies in Science, Management, Social sphere and Medicine»*. Russia, Tomsk, December 5-8, 2017. Atlantis Press, 2017, vol. 72, pp. 385-388.

18. *Bulletin board from private persons and companies on Avito*. Available at: https://www.avito.ru/ (accessed 12 March 2018) (in Russian).

19. *From hand to hand*. Available at: http://irr.ru/ (accessed 10 January 2018) (in Russian).

20. *Real Estate Portal*. Available at: http://realty.dmir.ru/ (accessed 4 January 2018) (in Russian).

21. *World of Apartments. Real estate.* Available at: http://mirkvartir.ru/ (accessed 14 February 2018) (in Russian).

22. *CIAN. Real estate database.* Available at: https://www.cian.ru/ (accessed 3 February 2018) (in Russian).

23. *Yandex.Realty.* Available at: https://realty.yandex.ru/ (accessed 9 February 2018) (in Russian).

24. *Kvadroom*. Available at: https://kvadroom.ru/ (accessed 2 March 2018) (in Russian).

25. *Sale and rental of real estate. Real estate Mail.Ru*. Available at: https://realty.mail.ru/ (accessed 16 January 2018) (in Russian).

26. *Real estate – sale and rent of apartments, houses, rooms, offices, property prices. Domino.* Available at: http://domino-rf.ru/nedvizimost/ (accessed 22 January 2018) (in Russian).

27. *Real estate portal of Moscow and St. Petersburg, Leningrad and Moscow regions.* Available at: http://www.restate.ru/ (accessed 20 February 2018) (in Russian).

28. *Register of property trades.* Available at: https://www.etp-torgi.ru/market/ (accessed 11 January 2018) (in Russian).

29. *Bidding for privatization, rent and sale of property.* Available at: https://i.rts-tender.ru/main/auction/Trade/Search.aspx (accessed 19 January 2018) (in Russian).

30. *Investment portal of the city of Moscow*. Available at: https://investmoscow.ru/tenders/ (accessed 30 January 2018) (in Russian).

31. *Sberbank-AST. Automated bidding system.* Available at: http://www.sberbank-ast.ru/ (accessed 14 January 2018) (in Russian).

32. *Portal of Services of the Federal Service of State Registration, Cadastre and Cartography*. Available at: https://portal.rosreestr.ru/ (accessed 14 February 2018) (in Russian).

33. *Reforma GKH*. Available at: https://www.reformagkh.ru/ (accessed 11 March 2018) (in Russia).

34. *Tomita-parser. Video course.* Available at: https://tech.yandex.ru/tomita/doc/video/index-docpage/ (accessed 7 January 2018) (in Russian).

35. A simple and fast method of comparing images for similarities. *stack.io.* Available at: http://qaru.site/questions/42410/simple-and-fast-method-to-compare-images-for-similarity (accessed 8 March 2018) (in Russian).

36. Zauner C. Implementation and Benchmarking of Perceptual Image Hash Functions. *pHash*, 2010, 108 p. Available at: https://www.phash.org/docs/pubs/thesis_zauner.pdf (accessed 16 February 2018).

Статья поступила 22 марта 2018 г.

Информация об авторах

Голубев Алексей Владимирович – соискатель ученой степени кандидата технических наук. Аспирант кафедры систем автоматизированного проектирования и поискового конструирования. Волгоградский государственный технический университет. Область научных интересов: машинное обучение; сбор, обработка и анализ данных; анализ больших данных; ГИС; геопространственные данные; eCity; Smart City; поддержка принятия решений; прогнозирование временных рядов. E–mail: ax.golubev@gmail.com

Парыгин Данила Сергеевич – кандидат технических наук. Доцент кафедры систем автоматизированного проектирования и поискового конструирования. Волгоградский государственный технический университет. Область научных интересов: город; информационно-коммуникационные технологии; геопространственные данные; моделирование; поддержка принятия решений; интеллектуальный анализ данных; Smart City; устойчивое городское развитие; управление социально-экономическим развитием города; местное самоуправление; управление муниципальными образованиями; урбанизация. E–mail: dparygin@gmail.com

Адрес: 400005, Россия, г. Волгоград, пр. им. Ленина, д. 28.

Финогеев Алексей Германович – доктор технических наук, профессор. Профессор кафедры систем автоматизированного проектирования. Пензенский государственный университет. Область научных интересов: сенсорные сети; поддержка принятия решений; САПР; беспроводные технологии; беспроводные сети; SCADA системы; информационная безопасность; 3D моделирование; вир-

туальная реальность; расширенная реальность; мониторинг. E–mail: alexeyfinogeev@gmail.com

Адрес: 440026, Россия, г. Волгоград, ул. Красная, д. 40.

The Approach to Integrated Processing of Open Data about the City Infrastructure

A. V. Golubev, D. S. Parygin, A .G. Finogeev

Purpose. *Support for decision-making on the management of the city requires a comprehensive analysis of the situation, which should be carried out using objective information that characterizes all the components of infrastructure, objects included in it and ongoing processes. However, the necessary information is placed in heterogeneous sources, and as a result, its use requires the development of a number of solutions for effective access organization. The purpose of the present paper is to develop a comprehensive approach to the processing of information from heterogeneous sources that provide open access to data on the city's infrastructure. This approach includes interrelated procedures for collecting and preliminary processing the entire array of received data on objects in the city from sources on the Internet. At the same time, sources are classified for the use of appropriate tools for information extraction and organization of a multi-level storage system for the purposes of subsequent use by external services.* ***Methods.*** *The solution of the multi-level task of information integration about the territory infrastructure is carried out using methods of intellectual analysis of geospatial data extracted from network resources. Data collection is realized using algorithms for retrieving data from web pages (parser, grabber) and direct access to the services API. Approaches of the theory of pattern recognition, including descriptions of reference points and perceptual hash algorithms are used for image preprocessing. Extraction of additional properties of objects is realized using the methods «Text Mining» and mathematical statistics.* ***Novelty.*** *Elements of novelty in the conducted research consists in the developed model of integration of the heterogeneous information and the sequence of data processing created on its basis, realizing the principle of overlapping backup storage of the structured and initial information.* ***Results.*** *Key sources of information on an urban infrastructure were identified. Their analysis allowed to form an idea of the structuring of data on the territory infrastructure and to developed an model of their integration taking into account a geospatial bind. Methods of information extraction from network resources for real estate objects, an approach to collecting and generalizing data from cartographic services, an algorithm for extraction of additional facts about infrastructure objects from a natural language record based on formal grammars, an approach to validation information based on image comparison using a hash algorithm, as well as the approach to determination the similarity of ads using the proposed system of coefficients with use statistical analysis of the initial data on real estate were developed in the framework of the implementation of geospatial data mining. The approach to organizing associated file storages and databases for store pre-processed, structured and raw data is proposed.* ***Practical relevance.*** *A single information processing system consisting of modules that implement operations of parsing online resources and accessing their APIs, analyzing natural language texts, images and structured data on real estate objects and city infrastructure has been realized and a complex of SQL and NoSQL databases to store the collected information on an equal basis with its placement in the form of source data files has been created on basis of the developed approaches.*

Key words: *open data; heterogeneous data sources; network resource; geospatial data; city infrastructure; data analysis; data preprocessing; real estate object; image processing; raw data; natural language; Open government; Data Mining; geoinformation system; SQL; NoSQL.*

Information about Authors

Alexey Vladimirovich Golubev – Doctoral Student. The postgraduate student of the Department of CAD. Volgograd State Technical University. Field of research: machine learning; collection, processing and analysis of data; Big Data analysis; GIS; geospatial data; eCity; Smart City; decision support; time series forecasting. E–mail: ax.golubev@gmail.com

Danila Sergeevich Parygin – Ph.D. of Engineering Sciences. Associate Professor at the Department of CAD. Volgograd State Technical University. Field of research: city; information and communication technologies; geospatial data; modeling; decision support; data mining; Smart City; sustainable urban development; management of social and economic development of the city; local government; management of municipalities; urbanization. E–mail: dparygin@gmail.com

Address: Russia, 400005, Volgograd, Lenina Ave., 28.

Alexey Germanovich Finogeev – Dr. habil. of Engineering Sciences, Full Professor. Professor at the Department of CAD. Penza State University. Field of research: sensor networks; decision support; CAD; wireless technologies; wireless network; SCADA systems; Information Security; 3D modeling; virtual reality; augmented reality; monitoring. E–mail: alexeyfinogeev@gmail.com

Address: Russia, 440026, Penza, Krasnaya Str., 40.