DOI 10.37882/2223-2966.2026.03.31

ПРОЕКТИРОВАНИЕ МЕТОДА КОЛЛЕКЦИОНИРОВАНИЯ ИНФОРМАЦИИ О НЕДВИЖИМОСТИ ЗА РУБЕЖОМ В УСЛОВИЯХ ОТСУТСТВИЯ ШАБЛОНИЗИРОВАННЫХ ИСТОЧНИКОВ ДАННЫХ

DESIGN OF A METHOD FOR COLLECTING INFORMATION ABOUT REAL ABROAD IN CONDITIONS OF THE ABSENCE OF STANDARDIZED DATA SOURCE

A. Pisarev
A. Oganesyan
A. Lomakin

Summary. The article examines the design of an adaptive parsing method for information collection and the development of an automated system for collecting data on foreign real estate oriented toward the Turkish market. The system, developed based on the modeled method, makes it possible to automate the process of obtaining data on price, location, and property characteristics from various sources without the need to consider predefined templates of the structural layout features of the sources. The system includes a module for automated adaptive parsing of web pages and a graphical interface that provides users with the ability to manage the collection process and monitor statuses. The conducted tests confirmed the effectiveness of the system in comparison with classical parsesrs in accelerating data collection processes, minimizing errors, and providing structured data of analytics.

Keywords: automatic information collection, real estate, automation, parsing.

Писарев Андрей Константинович
Волгоградский государственный технический университет
andrey-pisarev-0@mail.ru
Оганесян Артем Артакович
Волгоградский государственный технический университет
gdoq123@yandex.ru
Ломакин Арсений Сергеевич
Волгоградский государственный технический университет;
Волгоградский государственный медицинский университет
arseny.lomakin@gmail.com

Аннотация. В статье рассматривается проектирование метода адаптивного парсинга для сбора информации и разработка автоматизированной системы для коллекционирования данных о зарубежной недвижимости, ориентированной на рынок Турции. Система, разработанная на основе смоделированного метода, позволяет автоматизировать процесс получения данных о цене, местоположении и характеристики объектов, из различных источников без необходимости учитывать заранее заданные шаблоны структурных особенностей вёрстки источников. Система включает в себя модуль автоматизированного адаптивного парсинга веб-страниц и графический интерфейс, предоставляющий пользователям возможность управления процессом сбора и мониторинга статусов. Проведенные тестирования подтвердили эффективность системы в сравнении с классическими парсерами в ускорении процессов сбора данных, минимизации ошибок и обеспечении структурированных данных для аналитики.

Ключевые слова: автоматический сбор информации, недвижимость, автоматизация, парсинг.

Актуальность

Рынок зарубежной недвижимости продолжает демонстрировать рост, обусловленный его инвестиционной привлекательностью и высокой востребованностью среди россиян. Недвижимость остается одним из наиболее надежных способов инвестирования, предоставляя возможности для сохранения и приумножения капитала [1]. Помимо финансовых выгод, такие страны, как Турция и ОАЭ, предлагают дополнительные возможности для нерезидентов, включая программы получения гражданства через покупку недвижимости.

По данным опросов международного брокера недвижимости Tranio русскоязычные покупатели приобретали недвижимость в 2022 [6] прежде всего с целью получить вид на жительство или гражданство (52 % опрошенных), поэтому можно сделать предположение о том, что спрос на покупку недвижимости за рубежом у россиян будет стабильно высоким ближайшее время, в связи с популярностью стран с теплым климатом [7], являющихся стабильными партнёрами России.

Однако для эффективного анализа этого рынка и дальнейшего принятия решений о выгодной покупке

Таблица 1.

Статистика популярных стран
для покупки недвижимости у россиян

Страна	Доля заявок на покупку недвижимости
Турция	23,90 %
ОАЭ	11 %
Грузия	8,90 %
Испания	8,00 %
Греция	7,50 %
Таиланд	5,70 %
Кипр	5,50 %
Черногория	4,30 %
Франция	3,70 %
Италия	3,20 %

недвижимости требуется сбор и обработка большого объема данных, что является затруднительным если использовать только традиционные методы ручного сбора данных или простые парсеры.

Современные технологии автоматизированного сбора данных способны решать проблему высокой трудоемкости и риска ошибок, характерных для ручного мониторинга множества источников [2]. Автоматический сбор информации с веб-страниц, как отмечено в современных исследованиях, предоставляет доступ к актуальной информации напрямую с сайтов, исключая необходимость заключения дополнительных соглашений или использования посредников [3]. Такой подход позволяет получать структурированные данные быстро и без потери точности, но обычно у простых самописных парсеров присутствует проблема, связанная с шаблонизацией алгоритма парсинга под один конкретный источник данных, что исключает применение технологии на разных источниках, даже если они схожи между собой.

Основная сложность — отсутствие единого стандарта представления данных на сайтах, что приводит к необходимости адаптивных методов. Статичные парсеры быстро устаревают (URL: https://www.numberanalytics.com/blog/automated-data-collection-real-estate-insights) при изменении структуры страниц, что требует постоянного ручного обновления кода.

Целью данной работы является создание метода, который позволит оперативно собирать данные о зарубежной недвижимости, минимизировать временные затраты, исключая ошибки ручного мониторинга и невозможность использования классических парсеров с различными источниками данных.

Внедрение метода адаптивного алгоритма парсинга с динамической подстройкой под изменения структуры веб-страниц позволит сделать автоматизированные системы сбора и накопления данных более универсальным инструментом. Предположительно, это повысит эффективность обработки информации, особенно в условиях динамически развивающегося рынка, снижая временные затраты на её поиск и возможные ошибки, связанные с человеческим фактором.

В данной работе предлагается рассмотреть систему для автоматизированного сбора информации о недвижимости, основанной на адаптивном парсере. Работа включает следующие ключевые направления:

1. Анализ актуальности зарубежного направления на рынке недвижимости для российских покупателей;
2. Проектирование метода шаблонно-независимого автоматизированного сбора данных, обеспечивающей эффективную обработку информации из различных источников данных;
3. Разработка системы, включающей автоматизированный сбор данных и графический интерфейс для пользователей.

Анализ предметной области

Рынок недвижимости привлекателен тем, что помимо финансовой прибыли от ее продажи, недвижимость также и сама может приносить пассивный доход. Также можно получать дополнительный доход со сдачи недвижимости в аренду, что является распространенной практикой на сегодняшний день. Платежеспособность домохозяйств и объем предложения объектов недвижимости в краткосрочной перспективе [8] являются ключевыми факторами, определяющими рыночный спрос на недвижимость за рубежом.

Для брокеров недвижимости актуальна задача сбора информации в реальном времени с площадок объявлений, необходимой для постоянного мониторинга и анализа рынка. Обычно, они используют для этого ручной сбор данных, занося интересующие объявления в таблицы, однако некоторые прибегают к современным технологиям и используют самописные парсеры.

В целом, бизнес заинтересован в построении систем адаптивного парсинга, так как это позволяет оптимизировать рабочие процессы (Умный парсинг: [сайт]. URL: https://eclsoft.ru/parsing) и сократить время простоев, которое тратится на рутинную работу, в связи с чем в этой области активно проводятся научные исследования.

На данный момент в научном сообществе активно проводятся исследования, в которые включаются проектирования методов адаптивного парсинга. Так, в ра-

боте (Зеленский И.С., Парыгин Д.С., Савина О.В. Оценка привлекательности недвижимости при комплексном развитии участка территории // Информационное общество: образование, наука, культура и технологии будущего. 2022. № 6. С. 197–206) учёных из Волгоградского государственного технического университета описан алгоритм, позволяющий адаптировать структуру данных и оценочные критерии без изменения кода, что повышает универсальность и адаптивность системы анализа недвижимости. Реализованная подсистема для аналитической обработки открытых данных используется для оценки потребительской привлекательности объектов недвижимости.

В статье (Зеленский И.С., Парыгин Д.С., Савина О.В., Финогеев А.А., Шуклин А.А., Антюфеев А.Ю. Интеллектуальная поддержка решений по использованию объектов недвижимости для управления урбанизированными территориями // International Journal of Open Information Technologies. 2020. №11. С. 13–29) описана архитектура системы, включающая микросервисы «Text Parser» и «Ad Filter», которые обеспечивают адаптивный сбор и анализ данных о недвижимости, что повышает качество принятия решений в управлении урбанизированными территориями.

Зарубежные авторы также активно работают в этом направлении. Так, в работе (Barzin F., Yernaux G., Vanhoof W. ScrImmo: A Real-time Web Scraper Monitoring the Belgian Real Estate Market // Proceedings of the 22nd IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2023). 2023. P. 335–338) описывается разработка и внедрение веб-скрейпера SCRIMMO, который в режиме реального времени собирает данные о рынке недвижимости Бельгии. Отдельно в нём рассматривается техническая реализации с использованием JavaScript и RESTful API, а также возможностям анализа собранных данных для поддержки принятия решений.

В другом исследовании (Tsonev A., Dyankova V., Yusufov Y. Using web scraping for real estate price analysis // Mathematical and Software Engineering. 2022. Vol. 8, № 1–2, P. 19–23) представлен обзор областей применения веб-скрейпинга для адаптации к динамическим изменениям на рынке недвижимости, а также пример реализации для анализа цен. Помимо прочего, в нём рассматриваются технологии интеграции и обработки больших объёмов данных с различных веб-источников.

В академической среде тема также довольно популярна. Так, работа [17] Оскара Уильямса из MIT посвящена использованию веб-скрейпинга и алгоритмов поиска для выявления возможностей девелопмента и изменения зонирования на основе публичных данных. Разработанная платформа Aiden демонстрирует эффективность автоматизированного сбора и обработки данных для рынка недвижимости.

У самописных парсеров есть ограничения. Чаще всего, это скрипты, занимающиеся разовыми или периодическими выгрузками данных с определенного сайта или источника. У такого подхода есть ряд проблем, связанных, прежде всего, с необходимостью постоянно поддерживать парсер в актуальном состоянии, подстраивая под вводимые изменения в структуру данных и наполнение сайта, а также невозможность использовать подобный парсер в целях сбора аналогичной информации с другого источника данных.

Одним из способов решения проблемы является создание комбинированной методики проверки достоверности данных, сочетающей анализ повторяемости информации на разных источниках, машинное обучение для выявления аномалий и адаптации к изменениям источников объявлений, позволяющий минимизировать частоту доработок парсера при обновлении верстки целевых платформ.

Проектирование системы на основе предложенного метода

Адаптивный метод парсинга веб-страниц, представленный на рисунке 1 основан на комбинировании методов машинного обучения, в частности моделей компьютерного зрения и методов статического парсинга сайтов.

Для реализации метода необходимо создать систему, которая будет позволять брокеру недвижимости собирать заданную целевую значимую информацию (например цена, метраж, геолокация) из разных источников в условиях отсутствия шаблонизированной вёрстки и собирать их в единую коллекцию данных средствами СУБД.

Для этих целей, на текущий момент наиболее рационально использовать библиотеку Nokogiri [9] для составленных требований, потому что она является наиболее надежной библиотекой, так как ее разработка продолжается и по сей день, а следовательно, проблем с совместимостью не возникнет в ближайшей перспективе. Также она предоставляет все необходимые инструменты для сбора данных. Возможные проблемы с загрузкой оперативной памяти, при этом, маловероятны, так как будут обрабатываться данные небольших размеров.

Для адаптации к динамическим изменения вёрстки сайтов и выборки информации с разных видов шаблонов каждая полученная статическим парсером веб-страница обрабатывается методами компьютерного зрения, для чего используется библиотека PyTesseract.

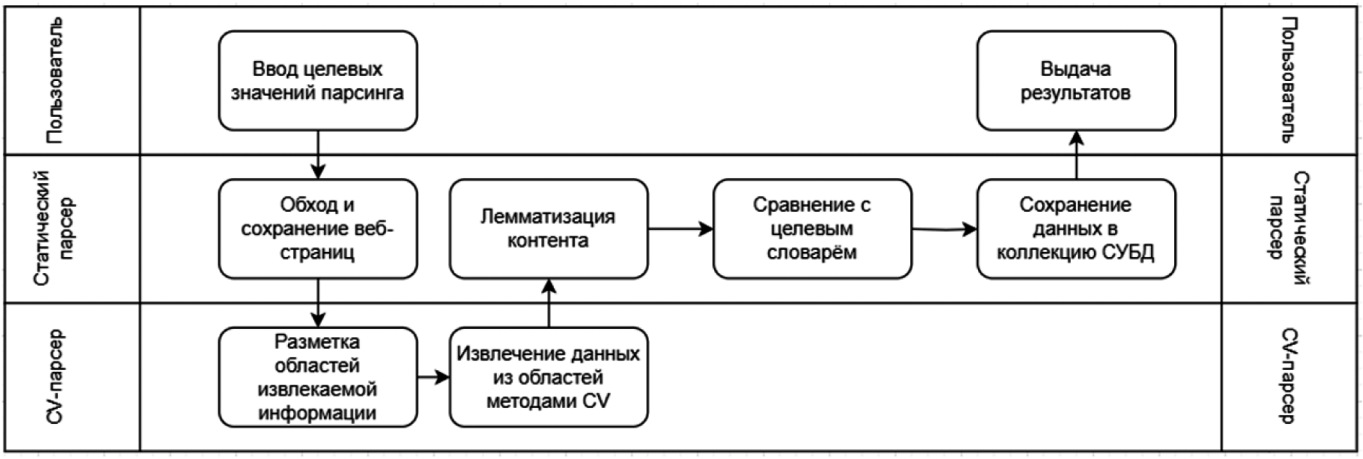

Рис. 1. Предложенный адаптивный метод парсинга

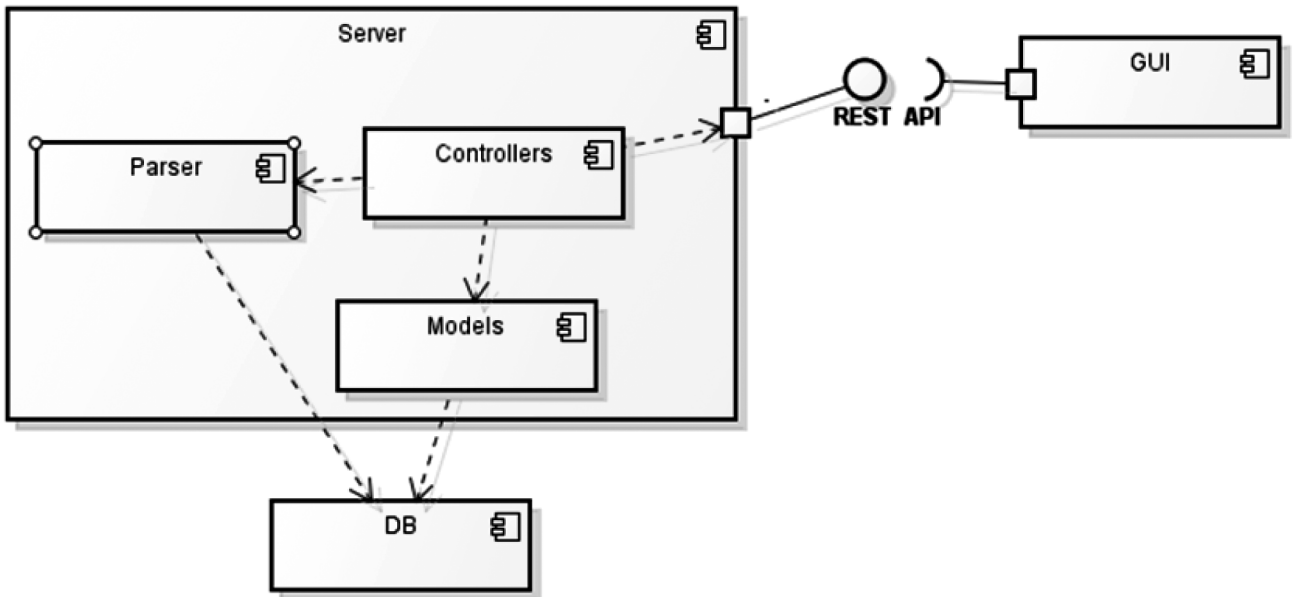

Рис. 2. Диаграмма компонентов системы сбора информации в нотации UML 2.0

У библиотеки есть заданный набор параметров, которые ищутся в полученном файле. Html-документ размечается в соответствии с ожидаемыми значениями на области данных, после чего значения из них лемматизируются.

Полученные наборы данных записываются в структуры данных, именуемые словарями (Map), численные значения которых предварительно переводятся в соответствующий тип данных, после чего данные словари извлекаются в документоориентированный формат JSON.

Полученная структурированная информация затем используется для накопления в СУБД и передаётся пользователю в шаблонизированном виде.

На рисунке 2 представлена диаграмма компонентов разрабатываемой системы, позволяющая применить описанный выше метод в реальной практике.

Разработка и внедрение системы на основе предложенного метода

Разработанная система предназначена для автоматизации составления базы знаний по объектам зарубежной недвижимости, которые содержатся на различных площадках объявлений о недвижимости.

Система состоит из двух модулей: модуля сбора информации и графического интерфейса пользователя, и имеет два способа запуска: через консольный интерфейс и через графический.

Графический интерфейс пользователя предназначен не только для запуска модуля сбора информации, он также имеет дополнительный функционал. С помощью GUI можно в понятном для человека формате следить за процессом парсинга, просматривать результаты работы.

Также добавлена возможность просматривать количество времени, которое уходит на сбор информации, а также получать подробные отчеты в файлах с подробным логгированием о том, как проходил процесс.

Модуль состоит из 9 классов, включая логгер и, классы специфицированные для парсера турецкой недвижимости с веб-страниц источников информации о ней. Их можно разделить на несколько категорий:

— классы, относящиеся к агентству;
— классы, относящиеся к объектам недвижимости;
— общие для первых двух групп классы.

Графический интерфейс пользователя реализуется на базе тестового стенда, в котором уже реализованы механизмы авторизации и регистрации. Для работы с системой автоматизированного сбора информации необходимо авторизоваться. В меню, при наведении на вкладку «Справочники» должен отображаться выпадающий список, последним пунктом которого должен быть, раздел «Парсеры».

Для дальнейшей работы с графическим интерфейсом системы необходимо перейти на вкладку «Парсеры». После перехода на странице отображается таблица с парсерами, содержащая информацию о модулях сбора (название агентства, дата и время начала процесса сбора, дата и время завершения, статус), ссылка на таблицу с логами и кнопка, чтобы можно было запустить процесс сбора информации.

Чтобы запустить модуль сбора информации необходимо нажать на кнопку «Начать парсинг». Кнопка приведёт в действие срабатывание POST-запроса на сервер, который запустит парсер и перезагрузит страницу с таблицей парсеров [11]. В результате статус выбранного

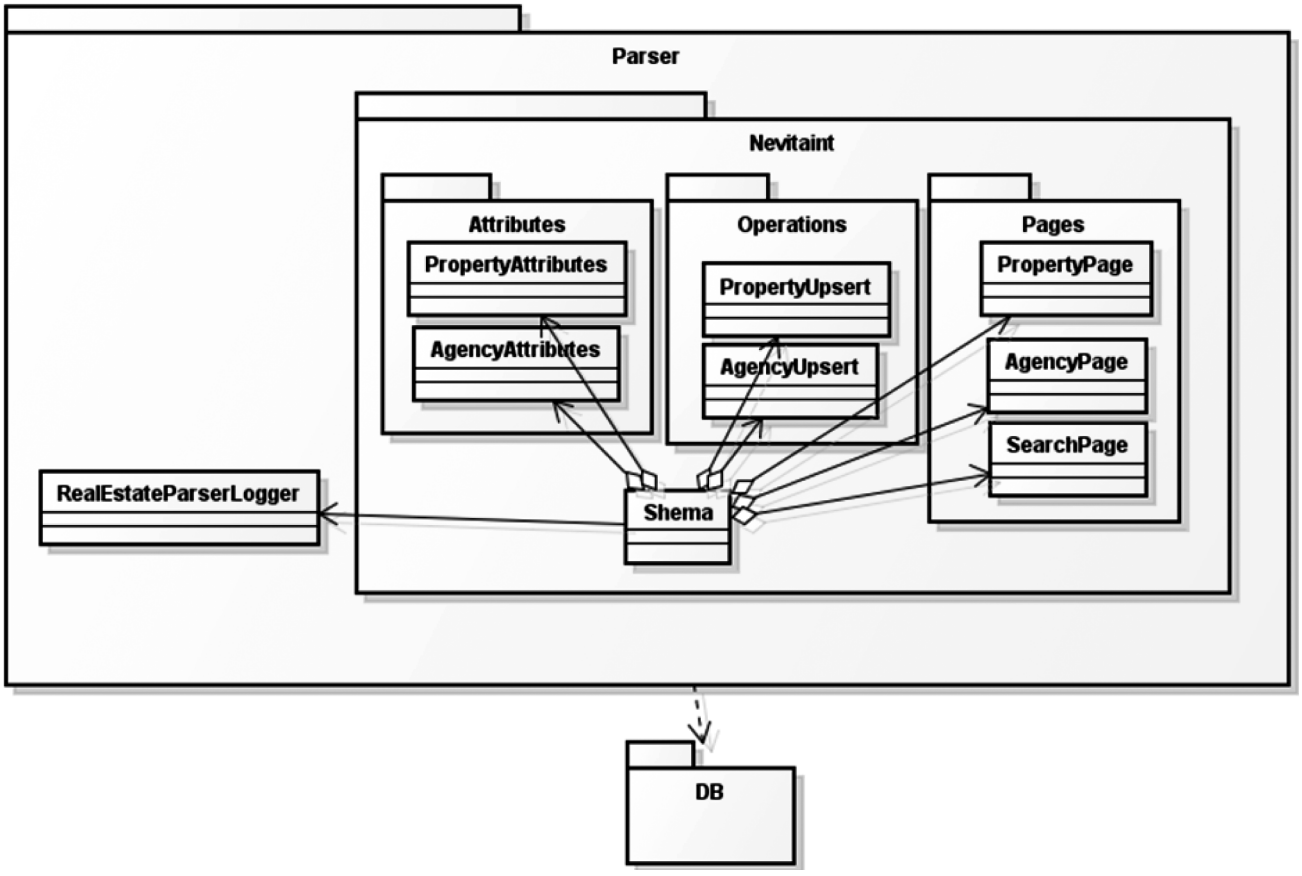

Рис. 3. Диаграмма классов модуля автоматизированного сбора информации о недвижимости в нотации UML 2.0

Главная / Парсеры

Парсеры

№	Агенство	Дата начала	Дата завершения	Статус		
4	nevitaint	30.05.24 19:06	30.05.24 19:07	неуспешно	Логи	Начать парсинг

Рис. 4. Макет таблицы парсеров

Таблица 2.

Полученные экспериментальные данные

Сборщик данных	Участник 1	Участник 2	Участник 3	Участник 4	Участник 5	Система шаблонизированного парсинга	Система с методом ААП
Время, затраченное на сбор информации, мин	45	90	59	43	65	Невозможно применить для разных источников данных	0,9

парсера изменится на статус «в процессе», а дата и время запуска обновятся. Одновременно с этим, в фоновом режиме будет выполняться сбор информации с сайта-источника.

В таблице парсеров отображается информация только из последнего сообщения в лог-файле. Чтобы посмотреть всё содержимое лог-файлов необходимо нажать на ссылку «Логи», после чего откроется страница с таблицей логов, которая содержит информацию о лог-файлах (дата и время начала и завершения парсинга, статус процесса, количество добавленных, удаленных и обновленных объектов недвижимости) и ссылка, позволяющая открыть лог-файл, содержащий все полученные исторические данные о запущенном процессе парсинга в браузере.

Заключение

Для проверки эффективности разработанной системы был проведен эксперимент. Суть эксперимента заключалась в том, чтобы сравнить ручной сбор информации и автоматизированный. В эксперименте приняли участие 5 уверенных пользователей ПК, которым было предложено собрать данные об агентстве недвижимости и объекте недвижимости из объявлений о продаже с веб-страниц площадок объявлений. Для сокращения времени проведения эксперимента. В качестве контрольной выборки, были использованы результаты группы, которой было предложено собрать информацию вручную на ограниченном объёме данных с 10 объявлений.

Во второй части эксперимента была запущена система автоматизированного сбора информации для сбора данных с такого же количества объявлений и того же агентства для повышения объективности результатов эксперимента.

Среднее арифметическое результатов эксперимента показало, что время, потраченное на ручной сбор, составило 60,4 минуты, что более, чем в 108 раз превышает автоматизированный сбор информации. При этом, в ходе ручного сбора были допущены ошибки и опечатки, которые невозможно допустить при парсинге. Так, внедрение автоматизированного подхода позволит многократно сократить время сбора информации о недвижимости.

Итак, предложенный метод автоматического анализа и адаптации к изменениям структуры сайтов с недвижимостью кроме плюсов, состоящих в минимизации частоты доработок парсера при обновлении верстки целевых платформ, также более чем на 99% быстрее справляется с задачей сбора информации и не совершает ошибок и опечаток в рамках разработанной системы, по данным проведенного эксперимента.

Таким образом, можно сделать вывод о том, что разработанная система, использующая в своей основе метод адаптивного алгоритма парсинга для динамических веб-страниц значительно эффективнее решает проблему, чем ручной сбор информации или применение автоматизированных системы без использования данной методологии, вследствие невозможности применения последних при наличии нескольких динамически изменяющихся источников информации.

ЛИТЕРАТУРА

1. Стерник С.Г. Рынок недвижимости и тенденции его развития: учебник. Москва: КноРус, 2023. 130 с.
2. Seferović A., Pejović N. The Integration of AI in Modern Real Estate Market Analysis // Advances in Computational Intelligence and Data Science. Springer, 2024. P. 123–145.
3. Костяшин Н.А. Применение автоматизированных средств сбора информации по сайтам // Информационные технологии и системы: управление, экономика, транспорт, право. 2020. № 3(39). С. 11–17.
4. Яхшисарова Р.М. Программная реализация парсинга данных раздела поиска недвижимости сайта avito.ru // Математическое моделирование процессов и систем: материалы VII Международной молодежной научно-практической конференции, Уфа, 07–09 декабря 2017 года / отв. ред. С. А. Мустафина. Часть II. — Уфа: Стерлитамакский филиал ФГБОУ ВО «Башкирский государственный университет», 2017. — С. 437–441.
5. Габдрахманова Л.З. Проблемы инвестирования в зарубежную недвижимость // Современные проблемы развития техники, экономики и общества: материалы II Международной научно-практической очно-заочной конференции, Казань, 04 апреля 2017 года / под ред. А.В. Гумерова. Казань: «Ро́кета Союз», 2017. С. 244–246.

6. Исследование: Русскоязычные покупатели зарубежной недвижимости — 2023 [Электронный ресурс] // Tranio: официальный сайт. URL: tranio.ru/articles/issledovanie-russkoyazychnye-pokupateli-zarubezhnoi-nedvizhimosti-2023/ (дата обращения: 15.12.2025).
7. Статистика: В 2022 году россияне скупали квартиры в Турции, ОАЭ и Грузии тысячами [Электронный ресурс] // Tranio: официальный сайт. URL: tranio.ru/articles/statistika-v-2022-godu-rossiyane-skupali-kvartiry-v-turcii-oae-i-gruzii-tysyachami/ (дата обращения: 15.12.2025).
8. Василенко Ж.А. Методика оценки объектов жилой недвижимости с учетом инвестиционных предпочтений // Инженерный вестник Дона, 2010, №4. URL: ivdon.ru/magazine/archive/n4y2010/270 (дата обращения 15.12.2025).
9. Hunter Powers Instant Nokogiri: Learning Data Scraping and Parsing in Ruby Using the Nokogiri Gem // Packt Publishing. 2013. P. 52.
10. Земцов А.Н, Болгов Н.В, Божко С.Н. Многокритериальный выбор оптимальной системы управления базы данных с помощью метода анализа иерархий // Инженерный вестник Дона, 2014, №2. URL: ivdon.ru/ru/magazine/archive/n2y2014/2360 (дата обращения: 15.12.2025).
11. Ломакин А.С. Разработка встраиваемой библиотеки для парсинга PDF-отчётов и автоматического интегрирования данных в карту пациента в области офтальмологии // NovaUm.Ru. 2023. № 44. С. 4–7.
12. Harnessing Automated Data Collection for Real Estate Insights [Электронный ресурс]. URL: https://www.numberanalytics.com/blog/automated-data-collection-real-estate-insights (дата обращения: 15.12.2025).
13. Умный парсинг [Электронный ресурс]: сайт. URL: https://eclsoft.ru/parsing (дата обращения: 15.12.2025).
14. Зеленский И.С., Парыгин Д.С., Савина О.В. Оценка привлекательности недвижимости при комплексном развитии участка территории // Информационное общество: образование, наука, культура и технологии будущего. 2022. № 6. С. 197–206.
15. Зеленский И.С., Парыгин Д.С., Савина О.В., Финогеев А.А., Шуклин А.А., Антюфеев А.Ю. Интеллектуальная поддержка решений по использованию объектов недвижимости для управления урбанизированными территориями // International Journal of Open Information Technologies. 2020. №11. С. 13–29. URL: https://cyberleninka.ru/article/n/intellektualnaya-podderzhka-resheniy-po-ispolzovaniyu-obektov-nedvizhimosti-dlya-upravleniya-urbanizirovannymi-territoriyami (дата обращения: 15.12.2025).
16. Barzin F., Yernaux G., Vanhoof W. ScrImmo: A Real-time Web Scraper Monitoring the Belgian Real Estate Market // Proceedings of the 22nd IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2023). 2023. P. 335-338.
17. Williams O. Identifying Real Estate Development Opportunities: Web-Scrapping, Regex Patterns & String-Searching Algorithms: Bachelor's thesis. Massachusetts Institute of Technology, 2021. URL: https://dspace.mit.edu/bitstream/handle/1721.1/139272/williams-oscarw-msred-cre-2021-thesis.pdf (дата обращения: 15.12.2025).