

Антипов Алексей Владимирович

студент, Амурский государственный университет, г. Благовещенск

Фомин Денис Васильевич

*канд. техн. наук, доцент кафедры информационной безопасности, Амурский
государственный университет, г. Благовещенск*

СИСТЕМА АВТОМАТИЗИРОВАННОГО СБОРА ДАННЫХ ИЗ ИНТЕРНЕТ-МАГАЗИНОВ. I

Аннотация. Статья посвящена проблеме автоматизированного сбора и обработки данных о товарах интернет-магазинов из открытых веб-источников. Рассматриваются существующие инструменты (ParseHub, WebHarvy, Диггернаут), анализируются их достоинства и недостатки: зависимость от облачной инфраструктуры, риски утечки коммерческой информации, низкая производительность при больших объёмах, сложность освоения и высокая стоимость. Выявлено, что ни одно из готовых решений не обеспечивает оптимального сочетания визуальной настройки, локального контроля, асинхронной обработки и экономической доступности. Обоснована целесообразность разработки собственного программного продукта, отвечающего указанным требованиям для эффективного мониторинга рынка.

Annotation. The article is devoted to the problem of automated collection and processing of data about products from online stores from open web sources. Existing tools (ParseHub, WebHarvy, Diggernaut) are considered, and their advantages and disadvantages are analysed: dependence on cloud infrastructure, risks of commercial information leakage, low performance with large volumes, difficulty of use, and high cost. It has been found that none of the ready-made solutions provides an optimal combination of visual configuration, local control, asynchronous processing, and economic accessibility. The feasibility of developing a proprietary software product that meets the specified requirements for effective market monitoring has been substantiated.

Ключевые слова: информационная система, автоматизированный сбор данных, веб-скрапинг, парсинг, интернет-магазин, нормализация данных.

Keywords: information system, automated data collection, web scraping, parsing, online store, data normalization.

Введение. Стремительное развитие информационных технологий и повсеместное распространение сети интернет привели к формированию огромных массивов данных, находящихся в открытом доступе. Однако они нередко присутствуют в человеко-, а не машинно-читаемом виде. То есть в открытом доступе в Интернете присутствуют серьёзные объёмы ценной информации, которые имеют запутанную нерегулярную структуру, форму и специфические способы доступа пригодные для человека. Всё это с одной стороны создаёт необходимость использования средств автоматической обработки (объёмы и скорость обновления данных), а с другой стороны существенно затрудняет из машинный сбор и обработку (структура, способы доступа). При этом возможность эффективного сбора и систематизации такой информации становится важной задачей для организаций различных типов и сфер деятельности.

Одним из таких открытых источников информации являются интернет-магазины. Сбор и обработка информации об их товарах приобретает повышенное значение, поскольку электронная коммерция продолжает активно развиваться и охватывать всё большую долю розничной торговли. Мониторинг цен конкурентов, анализ ассортимента товаров, отслеживание характеристик продукции требуют обработки больших объёмов информации, распределённой по множеству веб-ресурсов [1-3]. Ручной сбор и обработка таких данных является крайне трудоёмким процессом, зачастую неэффективным из-за больших объёмов и постоянного обновления информации в сети.

Технологии автоматизированного сбора и обработки данных позволяют решить проблему масштабируемости и оперативности получения информации

о товарах, однако их разработка сопряжена с рядом технических сложностей. Разнообразие форматов представления данных, различия в структуре веб-страниц, динамическая генерация контента, применение средств защиты от автоматизированного доступа создают существенные препятствия для создания универсальных решений. Существующие системы часто ориентированы на конкретные задачи или типы ресурсов, что ограничивает их применимость при работе с разнородными интернет-магазинами.

Таким образом, разработка системы автоматизированного сбора и обработки данных о товарах интернет-магазинов является актуальной задачей, решение которой позволит сократить трудозатраты, повысить качество мониторинга цен и ассортимента, а также обеспечить оперативную работу с информацией из интернет-магазинов игроков рынка.

Методы и инструменты. Рынок инструментов автоматизированного сбора и обработки данных с веб-ресурсов демонстрирует активное развитие в последние годы, что обусловлено растущим спросом на структурированную информацию для принятия бизнес-решений.

Существующие готовые решения можно классифицировать по нескольким ключевым характеристикам: типу клиентского приложения (облачные платформы, десктопные приложения), месту выполнения процесса сбора данных (локально на компьютере пользователя или на серверах провайдера), сложности освоения инструмента для новых пользователей, гибкости настройки и стоимости использования.

ParseHub является одним из наиболее известных десктопных приложений для автоматизированного сбора данных [4]. Программа предоставляет графический интерфейс, в котором пользователь может указывать элементы для извлечения непосредственно на отображаемой веб-странице (рис. 1).

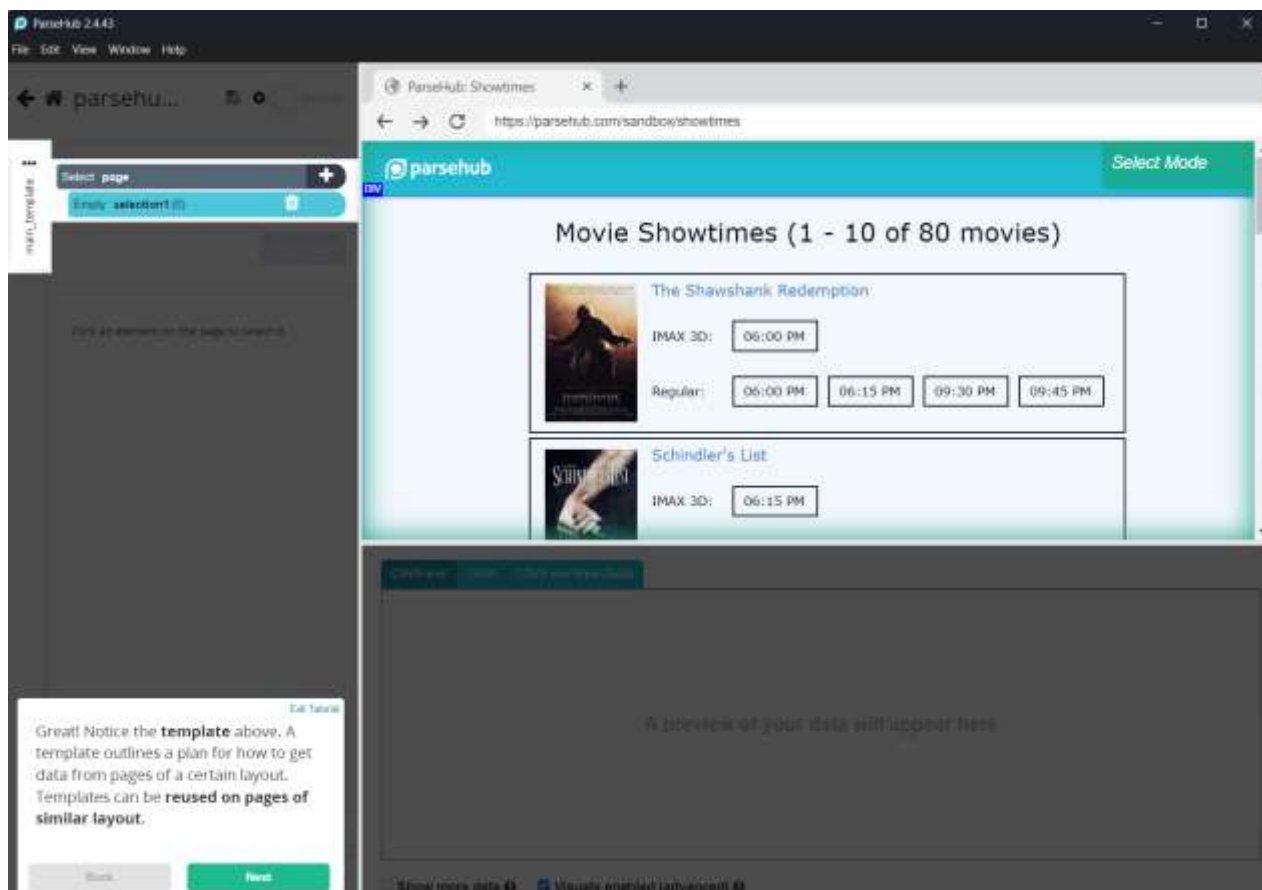


Рис. 1 – Интерфейс программы ParseHub

ParseHub автоматически определяет паттерны данных и формирует правила извлечения, что значительно упрощает процесс настройки по сравнению с программированием [3]. Инструмент поддерживает работу с динамическим контентом и пагинацией (постраничным разделением каталогов товаров с возможностью автоматического перехода между страницами для полного сбора данных).

Однако архитектура ParseHub предполагает выполнение процесса сбора данных в облачной инфраструктуре разработчика, что создает ряд существенных ограничений. Передача собираемых данных через серверы третьей стороны создает риски доступа к структурированной коммерческой информации о ценах, ассортименте и характеристиках товаров. Производительность процесса зависит от загруженности облачной инфраструктуры и качества интернет-соединения, что может приводить к непредсказуемым задержкам.

WebHarvy является альтернативным десктопным приложением для Windows, ориентированным на простоту использования (рис. 2) [5].

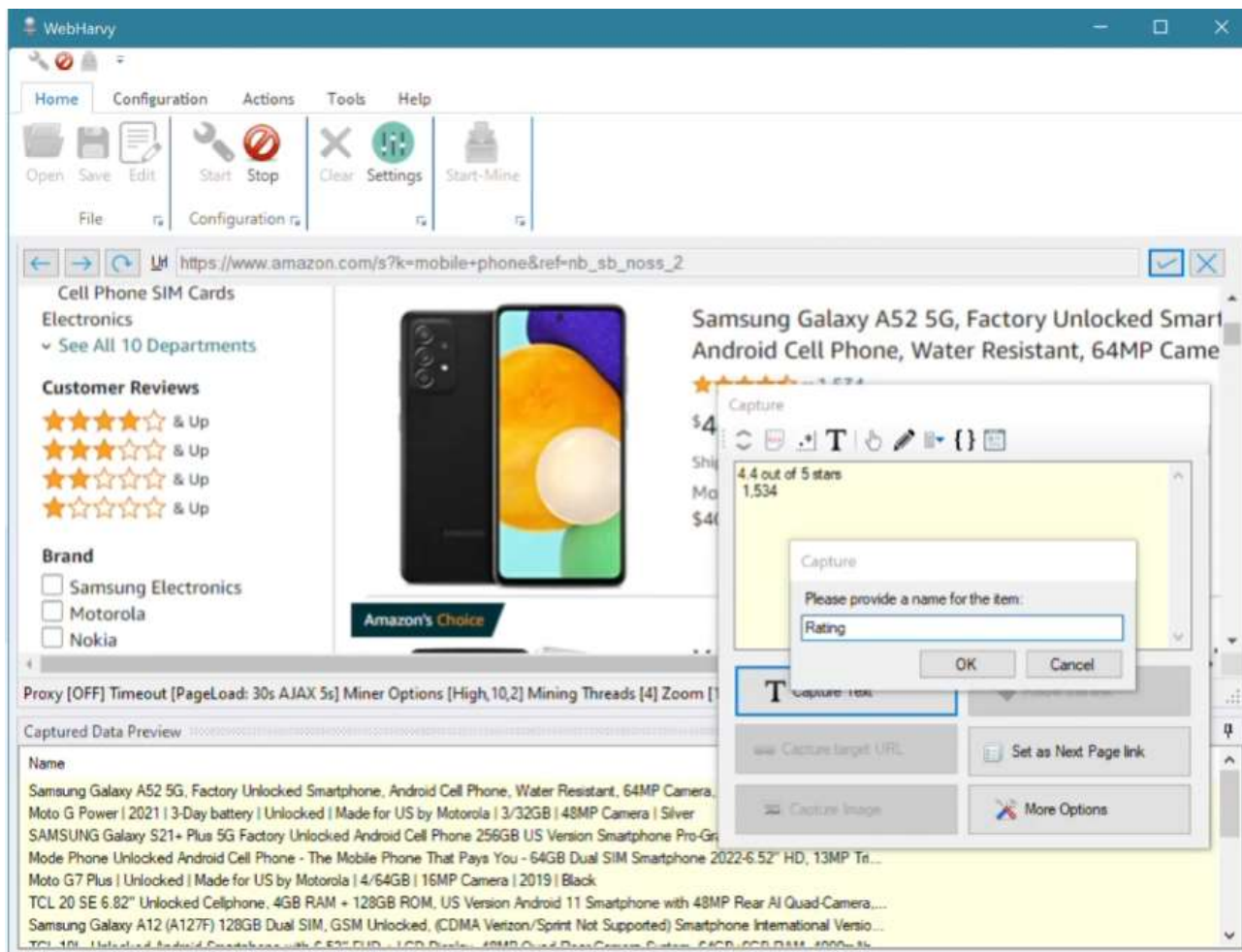


Рис. 2 – Интерфейс программы WebHarvy

Программа использует встроенный браузер для отображения веб-страниц и предоставляет интерфейс визуального выбора элементов – пользователь указывает мышью на данные, которые необходимо извлечь. WebHarvy автоматически распознает повторяющиеся паттерны данных и применяет правила извлечения к множеству страниц. Приложение поддерживает экспорт в XLSX, CSV, JSON, XML. WebHarvy работает полностью локально без облачной инфраструктуры, что обеспечивает полный контроль над структурированными данными.

Однако WebHarvy имеет значительное техническое ограничение. Приложение не поддерживает асинхронную обработку запросов, что при

работе с каталогами, содержащими сотни или тысячи товаров, приводит к длительному времени выполнения задачи.

На российском рынке присутствуют специализированные облачные сервисы сбора данных. Диггернаут представляет собой облачную платформу для создания парсеров (рис. 3) [6].

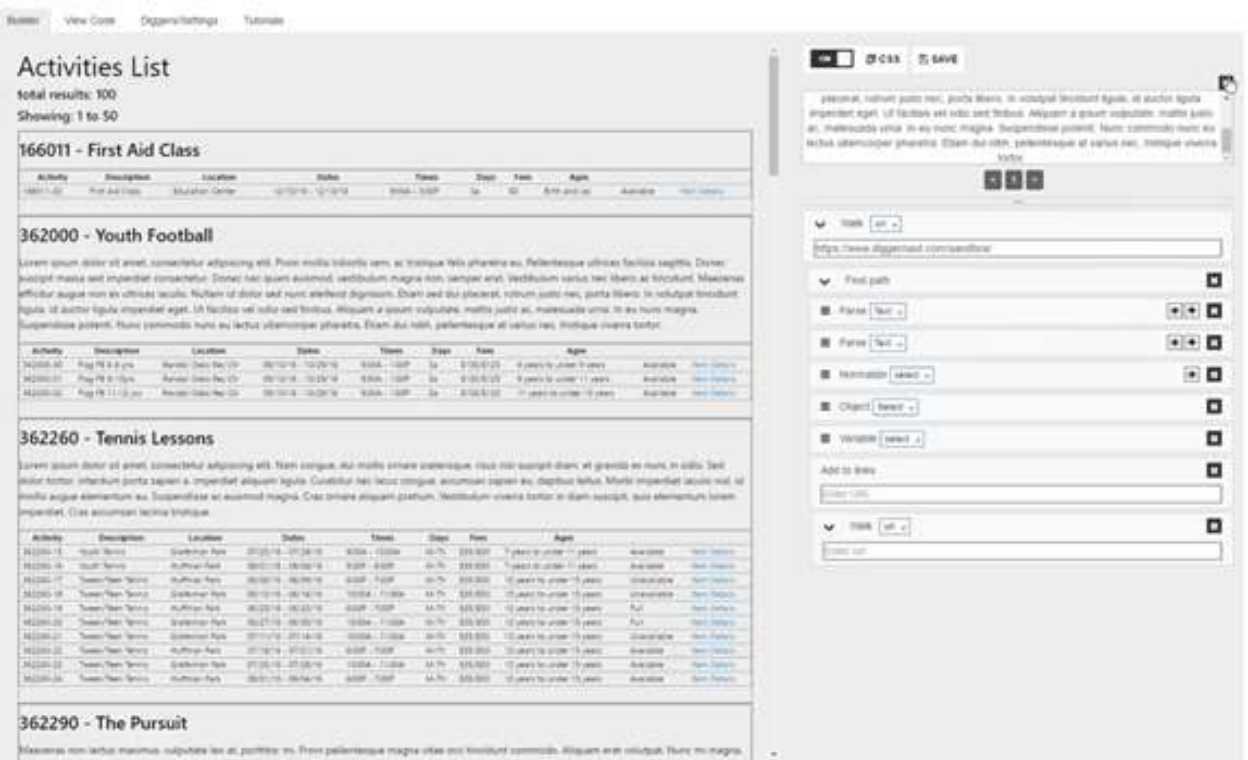


Рис. 3 – Интерфейс программы Диггернаут

Платформа предоставляет специализированный язык конфигурации для описания логики сбора данных, что занимает промежуточное положение между программированием и визуальной настройкой. Диггернаут поддерживает обработку динамического контента, работу с формами и авторизацией, экспорт данных в различные форматы.

Однако использование языка конфигурации требует изучения специфического синтаксиса и концепций платформы, что создает барьер для пользователей, поскольку освоение подобного инструментария требует значительных временных затрат и глубокого погружения в её внутреннюю логику. Все данные обрабатываются в облачной инфраструктуре, что создает риски доступа провайдера к агрегированной структурированной информации. Ценовая политика предполагает оплату на основе количества выполненных

запросов и объема обработанных данных, что может приводить к значительным расходам при регулярном мониторинге больших каталогов.

Анализ существующих решений выявляет системные проблемы, характерные для всех рассмотренных категорий инструментов. Локальные десктопные приложения имеют архитектурные ограничения из-за отсутствия асинхронной обработки. В свою очередь, решения, зависящие от облачной инфраструктуры, будь то облачные платформы или десктопные приложения, работающие через удалённые серверы, требуют передачи данных третьим сторонам, создают зависимость от провайдера и характеризуются высокой и непредсказуемой стоимостью использования.

Обзор существующих решений позволяет сделать вывод о целесообразности разработки программного решения, которое сочетало бы визуальный интерфейс настройки процесса сбора данных, доступный для специалистов без глубоких технических знаний, с полным локальным контролем над данными, приемлемой производительностью обработки и отсутствием зависимости от платных облачных сервисов, позволяя при этом работать с большинством веб-ресурсов без необходимости адаптации под каждый из них отдельно.

Заключение. Проведённый анализ современных инструментов автоматизированного сбора и обработки данных интернет-магазинов подтверждает высокую актуальность данной задачи в условиях стремительного роста электронной коммерции и увеличения объёмов неструктурированной информации. Рассмотренные решения демонстрируют разнообразие подходов к реализации парсинга, однако каждое из них обладает существенными недостатками, ограничивающими их практическую ценность для широкого круга пользователей.

Основные выявленные проблемы сводятся к трём ключевым аспектам: архитектурная зависимость от облачной инфраструктуры с сопутствующими рисками утечки коммерческой информации, недостаточная производительность при обработке больших каталогов из-за синхронной

архитектуры, а также высокий порог входа, связанный либо с необходимостью изучения специализированных языков конфигурации, либо с непредсказуемой стоимостью использования.

Таким образом, существующие решения не в полной мере удовлетворяют требованиям организаций, заинтересованных в эффективном, безопасном и экономически оправданном мониторинге рынка. В связи с этим представляется целесообразной разработка собственного программного продукта, который сочетал бы интуитивно понятный визуальный интерфейс настройки, локальное выполнение всех процессов для обеспечения полного контроля над данными, асинхронную обработку запросов для приемлемой производительности и гибкую систему правил извлечения, адаптируемую к разнородной структуре веб-страниц. Создание такого решения позволит не только снизить трудозатраты и операционные расходы, но и обеспечить независимость от внешних провайдеров и конфиденциальность собираемой информации.

Список литературы

1. Уколов А. Н. Веб-скрапинг: возможности, методы и инструменты [Электронный ресурс] // Молодой учёный. – 2024. – № 30 (529). – С. 29–31. – URL: <https://moluch.ru/archive/529/116859/> (дата обращения: 03.07.2026).
2. Web scraping вашего сайта: непрошенные гости и как их встречают [Электронный ресурс] // Хабр : блог компании Qrator Labs. – 2020. – URL: <https://habr.com/ru/company/qrator/blog/513014/> (дата обращения: 01.08.2026).
3. 15 топовых веб-скрапинг решений 2021 года [Электронный ресурс] // Хабр. – 2021. – URL: <https://habr.com/ru/articles/543760/> (дата обращения: 01.08.2026).
4. ParseHub: официальный сайт [Электронный ресурс]. – URL: <https://www.parsehub.com/> (дата обращения: 07.08.2026).
5. WebHarvy: официальный сайт [Электронный ресурс]. – URL: <https://www.webharvy.com/> (дата обращения: 07.08.2026).

6. Диггернаут: официальный сайт [Электронный ресурс]. – URL: <https://www.diggernaut.ru/> (дата обращения: 07.08.2026).

References

1. Ukolov A. N. Web scraping: possibilities, methods and tools [Electronic resource] // Young Scientist. – 2024. – № 30 (529). – Pp. 29-31. – URL: <https://moluch.ru/archive/529/116859/> (date of access: 03.07.2026).

2. Web scraping of your site: uninvited guests and how they are greeted [Electronic resource] // Habr : blog of Qrator Labs. – 2020. – URL: <https://habr.com/ru/company/qrator/blog/513014/> (accessed: 01.08.2026).

3. 15 top web scraping solutions of 2021 [Electronic resource] // Habr. – 2021. – URL: <https://habr.com/ru/articles/543760/> (date of request: 01.08.2026).

4. ParseHub: official website [Electronic resource]. – URL: <https://www.parsehub.com/> (accessed: 07.08.2026).

5. WebHarvy: official website [Electronic resource]. – URL: <https://www.webharvy.com/> (accessed: 07.08.2026).

6. Diggernaut: official website [Electronic resource]. – URL: <https://www.diggernaut.ru/> (accessed: 07.08.2026).