УДК 004.031.43

ПРОКСІ СЕРВІС ДЛЯ АГРЕГАЦІЇ ДАНИХ ПРО ТОВАРИ В РЕЖИМІ РЕАЛЬНОГО ЧАСУ

Вальтер А.А., Воропаєв В.О.
e-mail: anna.valter@nure.ua, vladyslav.voropaiev1@nure.ua
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

This paper is devoted to implementing a proxy service for aggregating and updating product information in real time. The objectives of the research include creating a server-side proxy service built on .NET platform and the integration of real-time updates into the data aggregation process. The paper also looks at conventional data extraction methods, analyzing their advantages and disadvantages in a real-time context. The work demonstrates the possibility of integrating server proxy, real-time, and security mechanisms, laying the groundwork for further improvements.

У контексті сучасних веб-додатків агрегація даних про товари з різних веб-джерел у режимі реального часу залишається критично важливим завданням через технічні обмеження, такі як CORS та неузгоджене форматування даних. Традиційні підходи, такі як клієнтські HTTP-запити або інструменти автоматизації браузерів, часто мають проблеми з масштабуванням, безпекою та синхронізацією в реальному часі.

Метою дослідження є розробка проксі сервісу, для агрегації, нормалізації та синхронізації даних про товари в режимі реального часу. Сервіс інтегрує гібридні методи обробки даних, а саме поєднання регулярних виразів та контекстного аналізу, та механізми безпеки для подолання проблем CORS.

Існуючі серверні інструменти, якнаприклад Selenium [1], пропонують часткові рішення, але не можуть інтегрувати обробку в реальному часі з надійним клієнт-серверним розмежуванням, що залишає прогалини в ефективності та адаптивності.Такі популярні бібліотеки, як Beautiful Soup (Python) і Jsoup (Java), вже давно використовуються для аналізу HTML-документів і вилучення релевантної інформації. Крім того, спеціалізовані фреймворки, такі як Scrapy, надають комплексні рішення для операцій парсингу веб сторінок. Однак ці методи в першу чергу орієнтовані на автономні завдання і не завжди є найкращим ріщенням для інтеграції у веб-додатки, що працюють у режимі реального часу. Сторонні проксі-сервіси, такі як Squid, або пропрієтарні API-шлюзи пропонують готові рішення для обробки перехресних запитів. Хоча ці рішення ефективні в загальних сценаріях, їм не вистачає гнучкості для налаштування логіки обробки даних відповідно до вимог додатків для роботи в режимі реального часу. Пропонується рішення на основі серверного проксісервісу розробленого на платформі .NET. Сервіс отримує та попередньо обробляє

вихідні веб-дані, використовуючи гібридну стратегію синтаксичного аналізу, яка поєднує зіставлення шаблонів на основі правил, наприкладвизначення валюти за допомогою регулярних виразів, з контекстно-орієнтованими методами вилучення. Нормалізація даних забезпечує узгодженість між різними форматами, перетворюючи результати в стандартизовану схему JSON.

Оновлення в реальному часі здійснюється за допомогою SignalR, який забезпечує миттєву двосторонню передачу даних між сервером та підключеними системами. На відміну від звичайних механізмів передачі даних, SignalR встановлює з'єднання в режимі реального часу, щоб надсилати оновлення, як тільки нові або змінені дані про продукт обробляються проксі-сервісом.

Використання серверного проксі дозволяє централізувати логіку отримання та обробки даних, тим самим зменшуючи ризики, пов'язані з прямою взаємодією з третіми сторонами. Програмна реалізація включаєперевірку вхідних даних та обробку помилок, що зменшує потенційні вразливості, притаманні загальним методам вилучення даних, такі як ін'єкційні атаки, ризики SSRF, витоки інформації через деталізацію виключень, тощо. Крім того, керуючи проблемами CORS на рівні сервера, підхід дозволяє уникнути обробку потенційно шкідливого контенту.

Запропонований підхід має потенціал для розширення можливостей проксі для підтримки додаткових атрибутів товару та інтеграції алгоритмів машинного навчання для прогнозування невідповідностей даних та застосування у системах різного напрямку, зокрема, в системах медичного призначення [2].

Запропоновано застосовувати комплексне рішення для агрегації даних про товари в реальному часі шляхом об'єднання вилучення атрибутів товарів, динамічної синхронізації та безпечної взаємодії з підключеними системами.

Список використаних джерел:

1. Мишко М. М. Аналіз можливостей використання SELENIUM для веб-скрапінгу задля отримання даних з онлайн ресурсів та автоматизації взаємодії програмними засобами / М. М. Мишко, А. І. Костромицький // Радіоелектроніка та молодь у XXI столітті : матеріали 27-го Міжнар. молодіж. форуму, 10–12 травня 2023 р. – Харків : ХНУРЕ, 2023. – Т. 4. – С. 147-148. URL: https://openarchive.nure.ua/handle/document/23928 (дата звернення: 09.03.2024).

2. Intelligent information system of heterogeneous medical data analysis / A. Yerokhin та ін. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), м. Lviv, 5–8 верес. 2017 р. 2017. URL: https://doi.org/10.1109/stc-csit.2017.8098798 (дата звернення: 09.03.2025).