УДК 004.9+556

К. О. Бондалєтов[1]
В. Б. Мокін[1]

АРХІТЕКТУРА КАСКАДНИХ ПАРСЕРІВ ГЕТЕРОГЕНИХ ДЖЕРЕЛ ДАНИХ ДЛЯ АВТОМАТИЗОВАНОЇ ПОБУДОВИ ДАТАСЕТІВ ДЛЯ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ МОНІТОРИНГУ

[1] Вінницький національний технічний університет

Анотація

У роботі представлено архітектуру каскадних парсерів гетерогенних джерел даних для автоматизованої побудови датасетів інтелектуальних систем моніторингу. Запропоновано поєднати розрізнені вхідні дані за типами джерел інформації (веб-ресурси у пошуковій видачі, статичні файли звітності та інтеграції по API з провайдерами тематичних даних) в єдиному конвеєрі збирання, нормалізації та валідації даних. Для кожного типу даних передбачається власна стратегія створення конвеєру підготовки даних до уніфікованого формату, придатного для подальшого аналітичного оброблення.

Запропонований підхід апробовано на прикладі автоматизованої побудови датасету для інтелектуальної веб-системи з інформацією про екологічні проблеми р. Південний Буг "WISEST SBB".

Ключові слова: видобування даних, автоматизація, інфраструктура, data pipeline, ГІС.

Abstract

The paper presents the architecture of cascade parsers of heterogeneous data sources for automated construction of datasets of intelligent monitoring systems. It is proposed to combine disparate input data by types of information sources (web resources in search results, static reporting files and integration via API with thematic data providers) in a single pipeline for data collection, normalization and validation. For each type of data, its own strategy for creating a pipeline for preparing data to a unified format suitable for further analytical processing is provided.

The proposed approach was tested on the example of automated construction of a dataset for an intelligent web system with information on environmental problems of the Southern Bug River "WISEST SBB".

Keywords: data mining, infrastructure, data pipeline, automation, GIS.

Вступ

Для підвищення якості роботи інтелектуальних інформаційних систем моніторингу пропонується підвищити повноту датасетів джерелами гетерогенних даних. Пропонується архітектура каскаду узагальнених типів джерел, як то ЗМІ, публікації в соціальних мережах, державні звіти, офіційні державні та громадські API.

Наведені джерела даних містять інформацію у різних форматах (html, pdf, csv, та ін.), різних режимах доступу (згадка у ЗМІ або API під паролем), неоднорідну структуру та неоднорідний ступінь достовірності. Існуючі інструменти видобування мають обмежену підтримку методів видобування даних і не мають підтримки єдиного конвеєру від неструктурованих даних до нормалізованого датасету.

Наприклад, у р. Південний Буг - 1090 масивів вод, з яких організована державна мережа постів спостережень включає 50 масивів вод. Для максимально комплексного оцінювання стану на окремих масивах рекомендовано охопити більше джерел даних, що актуалізує задачу автоматизованого збирання та нормалізації гетерогенних екологічних даних.

Метою даного дослідження є розроблення каскадної архітектури парсерів гетерогенних джерел даних, що видобуває та нормалізує дані у уніфікований формат, оптимальний для подальшого використання в інтелектуальних системах моніторингу.

Ідея розв'язання задачі

Головна ідея представленого підходу полягає у класифікації гетерогенних джерел інформації на три

базові типи з реалізацією спеціалізованих каскадів парсерів для кожного з них.

До першого типу пропонується віднести доступні пошуковим системам веб-ресурси (сайти ЗМІ, блоги, форуми, веб сайти органів влади, контент у соціальних мережах). Пропонується каскад парсерів в залежності від технічної реалізації веб-ресурсу, як то швидкий, але обмежений парсинг HTML, універсальний, але більш повільний парсинг, та браузерно-подібне рішення для покриття динамічних сайтів. Каскад будується за принципом резервування. У разі, якщо більш «дешевий» інструмент не може обробити сторінку, то вона передається наступному у черзі. Для кожного домену оптимально визначений підхід зберігається і перевикористовується при повторних зверненнях, що оптимізує надлишкові обчислення.

До другого типу пропонується віднести статичні (незмінні) файли, зазвичай це – звіти офіційних та некомерційних організацій та лабораторій. Пропонується універсальний API для обробки такого контенту.

До останнього типу відносяться публічні API провайдерів даних. Пропонується розробити персоналізовані парсери під формальні контракти для кожного провайдера даних з нормалізацією даних до єдиного внутрішнього формату системи.

Для кожного типу даних пропонується розробляти адаптовані під їх специфіку конвеєри агрегації даних, що включають валідатори та інструменти фільтрації потенційно шкідливого вмісту та механізми дедуплікації даних. Запропонована архітектура не залежить від доменної зони та може бути застосована у будь-якій інтелектуальній системи, що потребує автоматизованого збирання інформації з розрізнених джерел даних.

Результати дослідження

Відповідно до наведеної вище ідеї було реалізовано архітектуру (рис.1) каскадних парсерів гетерогених джерел даних для інтелектуальної веб-системи з інформацією про екологічні проблеми, природоохоронні заходи тощо про масиви вод басейну р. Південний Буг «WISEST Southern Bug Basin» (скорочено – «WISEST-SBB» або «WISESTR-SBB» – англ.: Water Information SystEm with Spatial and Temporal references for the Southern Bug Basin – «Водна інформаційна система з просторовою і часовою прив'язкою для басейну Південного Бугу»). Архітектура підсистеми базується на модульному моноліті, де окремі модулі є слабопов'язаними Docker-контейнерами зі спільною інфраструктурою, та через сховища redis та minio.

Для першого типу підготовлено Programmable Search Engine [6] з фокусом на українському сегменту інтернету. На вхід подаються 2 набори ключових слів – ті, по яких потрібно здійснювати пошук, та ті, які вказують на необхідність виключити певні результати з пошукової видачі (так звані «мінус-слова»). Ключові слова обох типів поповнюються через механізм зворотного зв'язку після оброблення видобутого контенту. Посилання зберігаються у вигляді структурованих json-файлів у локальному сховищі minio. На наступному кроці відбувається певне оброблення та валідація (як то нормалізація посилань – усунення utm та дедуплікація, таким чином, до наступного етапу потраплять тільки унікальні посилання). Пошук працює асинхронно, імплементовано пул API-ключів з автоматичною ротацією та обмеженням щоденних лімітів по кожному з ключів. Для оптимізації видобування контенту з сайтів запропоновано каскад парсерів: BeautifulSoup [5] для швидкого парсингу статичних сайтів, спеціалізований newspaper4k для сайтів ЗМІ, trafilatura як універсальний інструмент та playwright у headless режимі для динамічних веб сайтів (як то SPA сторінки та ін). Для кожного домену зберігаються визначені налаштування оптимального способу парсингу.

Для другого типу, а саме – оброблення державних звітів та інших статичних файлів підготовлено API, що дозволяє завантажувати файли у форматах csv, doc(x), xls(x), pdf, застосовувати необхідні правила обробки та зберігати у сховищі файлів.

Для конвеєру даних третього типу запроваджена опціональна підтримка інтеграції провайдерів публічних даних через формалізовані контракти даних та нормалізацію до єдиного формату.

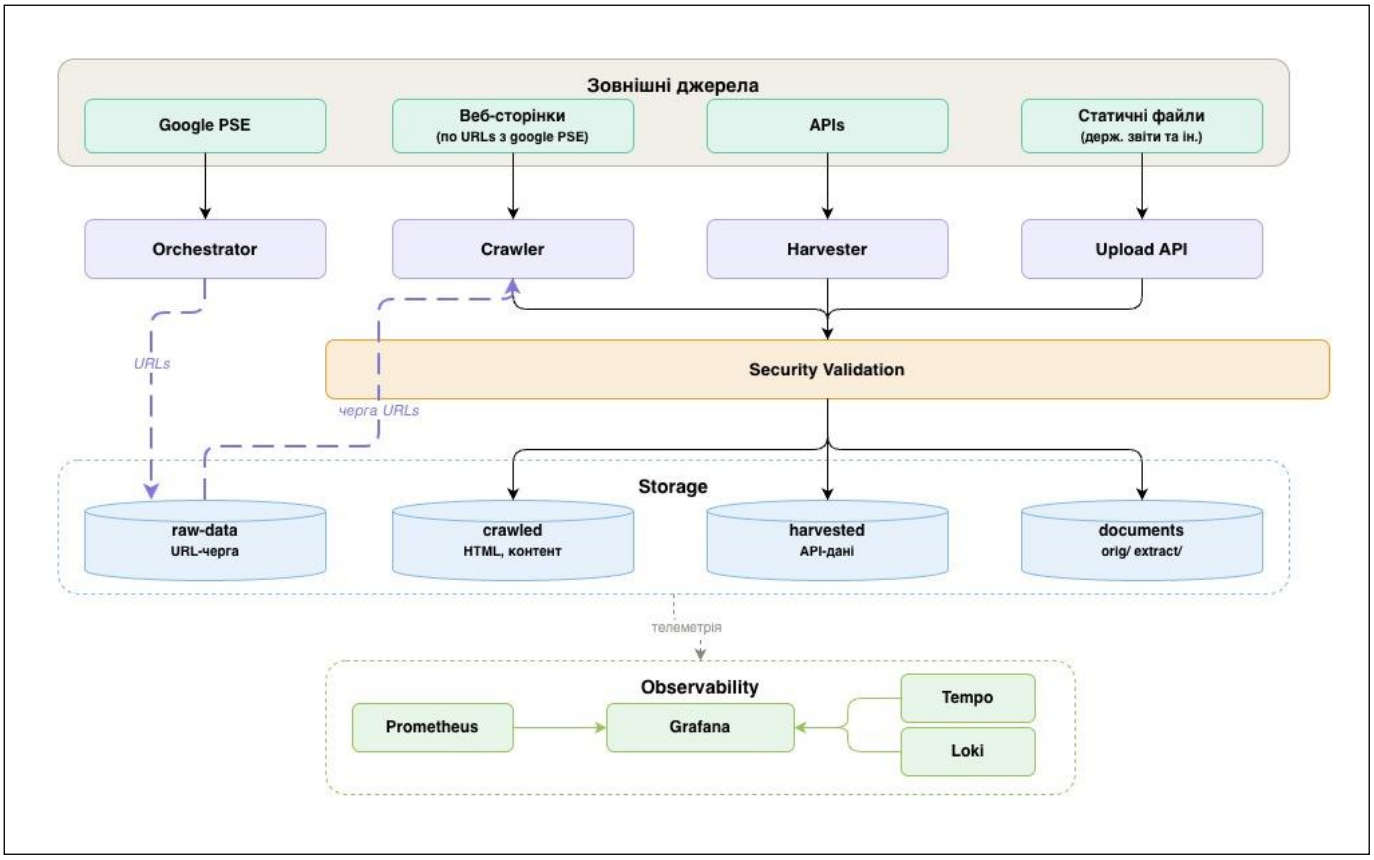

Рис. 1. Архітектура підсистеми каскадних парсерів гетерогенних джерел даних системи WISEST-SBB

На базі раніше побудованого датасету [3] було видобуто 151 цільове ключове слово для першого проходу системи. У результаті було отримано 2970 веб-посилань (з них 706 – дублікати поміж різними ключовими словами, 1134 унікальні веб домени; найбільше інформації – по річках та водосховищах, найменше – по містах). На подальших кроках нерелевантні результати поповнюють колекцію «мінус-слів», а на базі релевантних формуються низькочастотні уточнюючі ключові слова, що дозволяє покрити більше джерел інформації під час наступних ітерацій збирання.

Висновки

Спроєктовано та розроблено архітектуру каскадних парсерів гетерогенних джерел 3-х типів, із побудовою індивідуально налаштованих конвеєрів, нормалізацією та уніфікацією даних. Апробовано на прикладі підсистеми моніторингу та збирання даних системи WISEST-SBB. З практичної точки зору такий підхід дозволяє агрегувати всю доступну (довільні дописи у ЗМІ, державні звіти, звіти некомерційних організацій, видачу API провайдерів екологічної інформації тощо) експертам інформацію під єдиною «парасолькою», що очікувано підвищити якість розширеної аналітики основної системи.

Моніторинг поверхневих вод річки Південний Буг ускладняється через розрізнену природу наявних даних та динаміки їх оновлення, тож запропонована архітектура забезпечує промислову якість та масштабованість побудованого конвеєру даних, що дозволяє значно розширити обсяг доступних даних для побудови датасету.

Запропонована архітектура може бути застосована й для іншого роду інтелектуальних систем, які передбачають автоматичне або автоматизоване збирання різнорідної інформації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. В. Б. Мокін, К. О. Бондалєтов, Є. М. Крижановський, і В. О. Караваєв, «Метод аугментації текстів про стан масивів вод на основі інтелектуальної прив’язки до багатозв’язних геоінформаційних систем іменованих сутностей», Вісник ВПІ, вип. 3, с. 55–65, Черв. 2023.
2. К. О. Бондалєтов, В. Б. Мокін, І. М. Штельмах, і О. В. Слободянюк, «Автоматичне видобування знань з екологічних звітів з прив’язкою до часу та до просторових координат масивів вод», Вісник ВПІ, вип. 3, с. 101–110, Черв. 2025.

3. Бондалєтов К. О., Мокін В. Б., Григорчук М. В., Джура С. В., Кищук М. О., Нереуцький О. В., Неволя С. Д., Фурман А. М., Гіжевський В. В. Побудова датасету для тренування інтелектуальних моделей веб-системи з інформацією про екологічні проблеми та заходи у масивах вод басейну р. Південний Буг WISEST-SBB // Матеріали LII Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 21 – 23 червня 2023 р. – Електрон. текст. дані. – 2023. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/allfksa2023/paper/view/18958/ 15723
4. К. О. Бондалєтов і В. Б. Мокін, «Інтелектуальна технологія видобування та верифікації числових та текстових даних у багатозв'язних багатостадійних геоінформаційних системах», Вісник ВПІ, вип. 5, с. 51–60, Жовт. 2025.
5. Бондалєтов К.О. Collecting Google Search Data with BeautifulSoup. Kaggle: The World's AI Proving Ground. URL: https://www.kaggle.com/code/bondaletov/collecting-google-search-data-with-beautifulsoup (date of access: 20.03.2026).
6. Programmable Search Engine Help. Google Help. URL: https://support.google.com/programmable-search/ (date of access: 20.03.2026).
7. Портал відкритих даних. Головна сторінка - Data.gov.ua. URL: http://www.data.gov.ua (дата звернення: 20.03.2026).

Бондалєтов Костянтин Олегович — аспірант кафедри системного аналізу та інформаційних технологій; e-mail: bondaletov.k@gmail.com;

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій; e-mail: vbmokin@vntu.edu.ua.

Вінницький національний технічний університет, Вінниця

Bondalietov Kostiantyn O. — Post-graduate student of the Chair of System Analysis and Information Technologies, e-mail: bondaletov.k@gmail.com;

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@vntu.edu.ua.