УДК 004.624

ГІБРИДНИЙ ПІДХІД ДЛЯ ЕКСТРАКЦІЇ ДАНИХ У ДИНАМІЧНОМУ ВЕБ-СЕРЕДОВИЩІ

Шатило І. Ю., Безкоровайний В. В., Чала Л. Е.

Харківський національний університет радіоелектроніки, Харків

Сучасні інформаційні системи реалізують процеси збору й обробки даних, левова частка яких генерується у динамічному веб-середовищі. Висока динамічність веб-сторінок, відсутність уніфікованих стандартів семантичної розмітки та постійні зміни у структурі DOM (Document Object Model) перетворюють екстракцію даних (Data Extraction) на складну науково-прикладну проблему. Традиційні підходи до її вирішення, засновані на регулярних виразах або статичних XPath-селекторах, демонструють низьку робастність та вимагають постійної «ручної» підтримки [1]. Це зумовлює перехід до інтелектуальних методів [2] і необхідності розробки нових засад для побудови стійких та адаптивних систем веб-екстракції.

Веб-середовище W пропонується розглядати як великомасштабну, відкриту, стохастичну та динамічну систему, особливостями якої є:

– *гетерогенність (heterogeneity)*: веб-система складається з елементів (веб-сторінок P) різної природи. Гетерогенність проявляється на кількох рівнях: структурному (різноманітність DOM-дерев, використання фреймворків), презентаційному (різні CSS, візуальне оформлення) та семантичному (відсутність єдиної онтології для опису даних);

– *динамічність (dynamism)*: система W знаходиться у стані постійної зміни $W(t)$. Це стосується як контенту $C(t)$, так і структури $S(t)$. Динамічність генерується як серверними технологіями (CMS), так і клієнтськими (JavaScript, AJAX), що модифікують DOM «на льоту»;

– *невизначеність (uncertainty)*: для зовнішнього спостерігача (системи екстракції) веб-середовище є «сірою» або «чорною» скринькою. Неможливо

точно передбачити структуру сторінки $p \in P$, яка буде отримана за запитом q.

Формалізуємо задачу екстракції. Існує множина цільових даних D, які необхідно вилучити. Веб-сторінка p є структурою, $p = \langle T_{DOM}, T_{CSS}, C_{data} \rangle$, де T_{DOM} – DOM-дерево, T_{CSS} – стилі, C_{data} – контент. Задача екстракції полягає у пошуку функції (оператора) $F_{extract}: P \to D'$ (де D' – вилучені дані, причому необхідно максимізувати метрику відповідності $M(D', D) \to \max$).

Проблема полягає в тому, що функція $F_{extract}$ є нестаціонарною. Внаслідок динамічності $W(t)$, оператор $F_{extract}(t-1)$, побудований для сторінки $p(t-1)$, може стати невалідним для $p(t)$. Класичні системи екстракції намагаються «зафіксувати» $F_{extract}$ через жорсткі правила (наприклад, XPath), що системно призводить до їхньої деградації.

На рисунку 1 зображено концептуальна схема візуалізації проблеми.

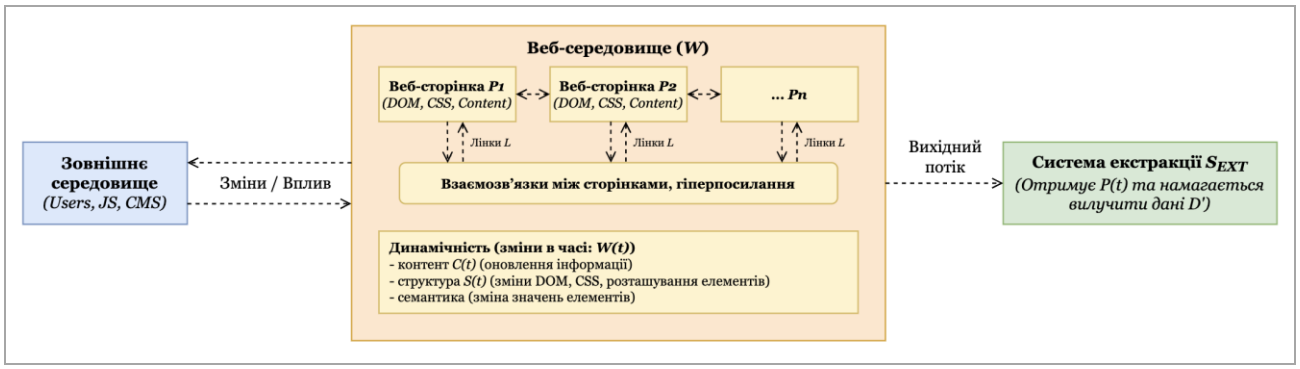

Рисунок 1 – Концептуальна схема веб-середовища W як динамічної системи

Мета системи S_{EXT} (рис. 2) полягає у підтримці стабільного стану (високої точності екстракції $M(D', D)$) попри збурюючі впливи $Z(t)$ з боку $W(t)$. Система S_{EXT} має такі компоненти: *входи (Inputs)* – цільова URL-адреса або запит (U_{in}), HTML/DOM-структура сторінки $p(t)$; *виходи (Outputs)* – структуровані дані D_{out}; *керуючий блок (Control Unit)* – модель екстракції M_{model}; *зворотний зв'язок (Feedback)*: механізм оцінки якості екстракції $E = 1 - M(D_{out}, D)$.

Традиційні системи працюють за принципом розімкненого контуру.

Системи на основі машинного навчання намагаються реалізувати замкнений контур, адаптуючи M_{model} через перенавчання. Однак більшість сучасних ML-моделей (зокрема, глибокі нейронні мережі) є «чорними скриньками». Це створює фундаментальну проблему з точки зору теорії управління: відсутність спостережуваності (observability) внутрішнього стану системи.

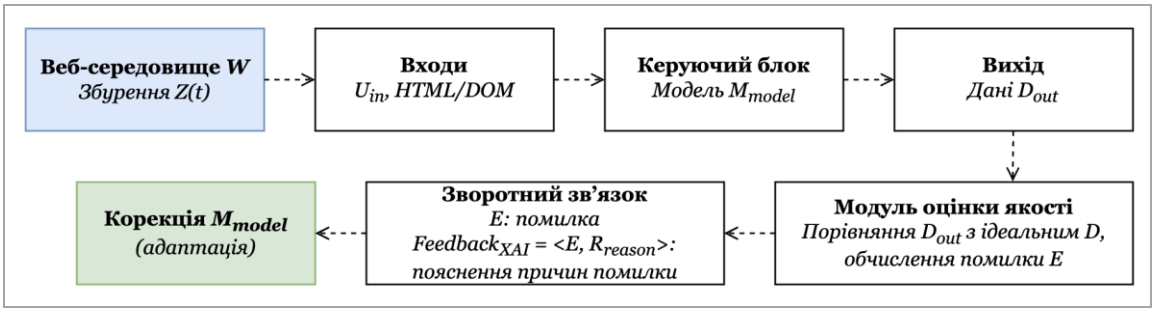

Рисунок 2 – Структурна схема системи екстракції S_{EXT}

Для усунення цього недоліку пропонується використати механізм пояснюваності (Explainability, XAI) [3], що реалізується додатковим каналом зворотного зв'язку. Це дозволяє отримувати інформацію не лише про факт помилки, але й про її причину $Feedback_{XAI} = \langle E, R_{reason} \rangle$ (де R_{reason} – набір причин або логічних пояснень, чому D_{out} було згенеровано саме таким чином). У такий спосіб буде реалізовано перехід від простої адаптації (перенавчання «всліпу») до керованої еволюції моделі M_{model}. Пропонується здійснювати вирішення цього завдання шляхом синтезу гібридної системи, у якій будуть інтегровані моделі різної природи для обробки різних аспектів W.

Розроблювана система S_{HYBRID} інтегрує:

– *підсистему візуального аналізу* S_{CV}: використовує методи комп'ютерного зору, зокрема візуально-семантичне представлення [4], для ідентифікації візуально значущих блоків контенту, ігноруючи невидиму DOM-структуру. Ця підсистема стійка до змін T_{DOM}, якщо візуальне представлення T_{CSS} залишається сталим;

– *підсистему структурно-семантичного аналізу* $S_{NLP/SYM}$: використовує нейромережеві методи (для семантики) та нейро-символьні підходи (для логіки та правил) [5]. Ця підсистема аналізує DOM та текстовий контент, шукаючи семантичні патерни, а не жорсткі структурні шляхи;

– *підсистему обробки невизначеності* S_{FUZZY}: використовує нейро-нечіткі моделі (neuro-fuzzy) для агрегації суперечливих або неповних даних, отриманих від S_{CV} та $S_{NLP/SYM}$. Це дозволяє системі приймати рішення в умовах нечітких вхідних сигналів (наприклад, «блок схожий на заголовок» та «тег ймовірно є ціною»).

Ключовим елементом пропонованої архітектури є інтегруючий модуль пояснюваності I_{XAI}, який не просто об’єднує виходи підсистем, але й генерує пояснення R_{reason} для кінцевого рішення.

Формально, вихідний оператор F_{hybrid} є складною композицією:

$$F_{hybrid} = I_{XAI}(S_{FUZZY}(S_{CV}(p), S_{NLP/SYM}(p))).$$

Очікується, що така гібридна архітектура (рис. 3) дозволить вирішити ключові системні проблеми: S_{CV} покликана боротися з динамічністю T_{DOM}, $S_{NLP/SYM}$ – з гетерогенністю C_{data}, S_{FUZZY} – з невизначеністю, а I_{XAI} – з проблемою спостережуваності та інтерпретації («чорної скриньки»).

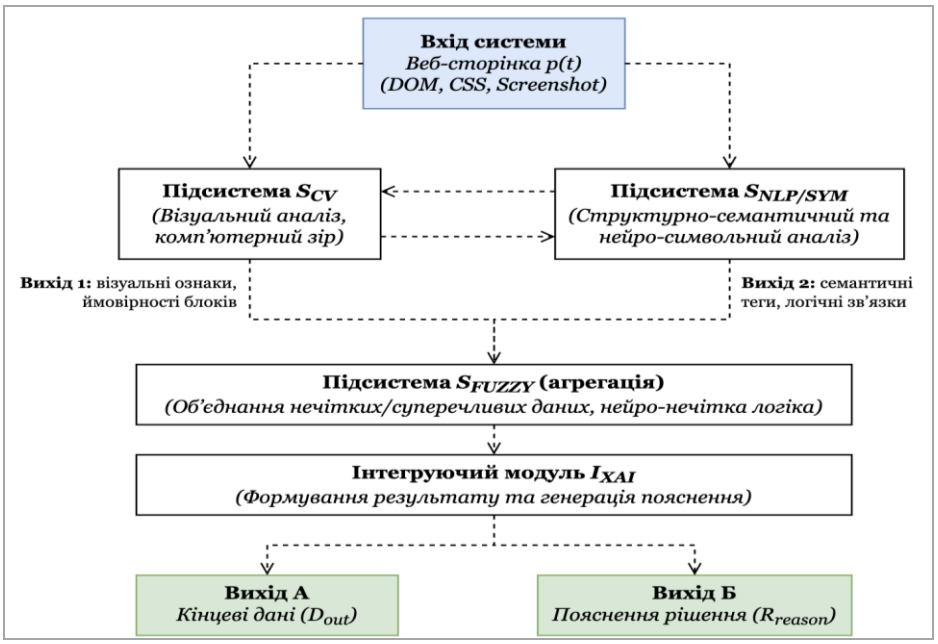

Рисунок 3 – Архітектура пропонованої гібридної системи S_{HYBRID}

Перевагою запропонованої системної архітектури є пояснюваність, яка реалізується через спеціалізований інтегруючий модуль I_{XAI}. Цей модуль забезпечить необхідний рівень спостережуваності за внутрішнім станом системи, дозволяючи не лише фіксувати помилки, але й розуміти їхні причини, що є базисом для побудови адаптивних та робастних інформаційних систем екстракції даних.

Подальші дослідження будуть спрямовані на практичну реалізацію та експериментальну валідацію описаної гібридної моделі, оцінку її робастності порівняно з існуючими підходами та розробку формальних метрик оцінки якості пояснюваності R_{reason} у контексті задачі екстракції веб-даних.

Література:

1. Web Scraping Techniques and Applications: A Literature Review / C. Lotfi та ін. SCRS CONFERENCE PROCEEDINGS ON INTELLIGENT SYSTEMS. 2021. С. 381–394. URL: https://doi.org/10.52458/978-93-91842-08-6-38 (дата звернення: 30.10.2025).
2. Web data extraction, applications and techniques: A survey / E. Ferrara та ін. Knowledge-Based Systems. 2014. Т. 70. С. 301–323. URL: https://doi.org/10.1016/j.knosys.2014.07.007 (дата звернення: 31.10.2025).
3. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI / A. Barredo Arrieta та ін. Information Fusion. 2020. Т. 58. С. 82–115. URL: https://doi.org/10.1016/j.inffus.2019.12.012 (дата звернення: 31.10.2025).
4. Dori D. ViSWeb - the Visual Semantic Web: unifying human and machine knowledge representations with Object-Process Methodology. The VLDB Journal. 2004. Т. 13, № 2. С. 120–147. URL: https://doi.org/10.1007/s00778-004-0120-x (дата звернення: 02.11.2025).
5. Artur Garcez. Neurosymbolic AI is the answer to large language models' inability to stop hallucinating. The Conversation. URL: https://theconversation.com/neurosymbolic-ai-is-the-answer-to-large-language-models-inability-to-stop-hallucinating-257752 (дата звернення: 02.11.2025).