УДК 004.4

Ліміна В.Ю.
Бабюк Н.П.

ПРОГОРАМНІ ЗАСОБИ ДЛЯ АВТОМАТИЗАЦІЇ ЗБОРУ І ОБРОБКИ ІНФОРМАЦІЇ

Вінницький національний технічний університет

Анотація

Проведено класифікацію програмних засобів для автоматизації збору та оброблення інформації із мережі Internet. Розглянуто алгоритм та схему роботи програми для web-скрейпінгу.

Ключові слова: парсер, web-кроулінг, web-скрейпінг.

Abstract

The classification of software for automation of collection and processing of information from the Internet is carried out. The algorithm and scheme of the program for web-scraping are reviewed.

Keywords: parser, web crawling, web scraping

.

Вступ

В сучасному світі інформація є найціннішим ресурсом, а її найбільшим джерелом є мережа Internet. В тих випадках, коли йдеться про величезні обсяги даних, обробити їх вручну практично дуже важко. Тому автоматизація збору і оброблення інформації є досить корисним і важливим процесом.

Зазвичай, для швидкої обробки інформації використовуються парсери. Парсер – це програмний засіб, що здатний швидко обробляти різні дані згідно з заданим алгоритмом і повертати потрібний результат. Наразі парсери є найефективніше рішенням для автоматизації збору і оброблення інформації. Парсер здатний швидко обходити велику кількість web-сторінок, знаходити потрібні дані і повертати її у визначеному форматі.

Результати дослідження

Найпоширенішими парсерами є пошукові роботи, які використовують пошукові системи. Вони проводять аналіз сторінки, збереження інформації про них в базі даних і далі, під час пошукового запиту користувача, повертають йому всю актуальну інформацію. Для розробки програм-парсерів використовуються багато мов програмування, серед яких: Python, JavaScript, Java, C#, Perl, Ruby, Go та інші.

Переваг у застосуванні парсерів безліч, серед яких є і автоматизація процесів, висока швидкість і висока продуктивність. Комп'ютерна програма-парсер допоможе опрацювати за короткий час тисячі web-сторінок певної тематики, деталізовано виокремить технічну інформацію із наборів загальних даних, таким чином виділить потрібне і видалить зайве, а також ефективно сформує кінцеві дані в необхідному форматі.

Розрізняють два види парсингу – кроулінг та web-скрейпінг [1].

Кроулінг (crawling) – це процес сканування сайту автоматизованою системою. Один із засобів кроулінгу – пошуковий робот (web crawler) – програма, яка є складовою частиною пошукових систем та застосовується для обходу сайтів в мережі Internet для занесення інформації про них за допомогою ключових слів до бази даних. Він здійснює загальний пошук інформації в мережі Internet та виводить звіт про зміст знайденого документу, індексує його й вишукує підсумкову інформацію.

Web-скрейпінг (scraping) – це процес перетворення у структуровану форму інформації з web-сторінок, які призначені для перегляду користувача через браузер. Як правило, скрейпінг виконується за допомогою додатків, що імітують поведінку людини в Інтернеті через web-сервер напряму, або через повноцінний web-браузер.

Зазвичай web-скрапери працюють за таким послідовним алгоритмом:

1. Підготовка механізму отримання HTML-коду за запитом типу GET.

2. Аналіз DOM-структури потрібного Internet-ресурсу.
3. Визначення вузлів з потрібними даними.
4. Налаштування обробника вузлів.
5. Виведення даних в нормалізованому вигляді (наприклад, в форматі JSON).

Схема роботи зображена на рисунку 1.

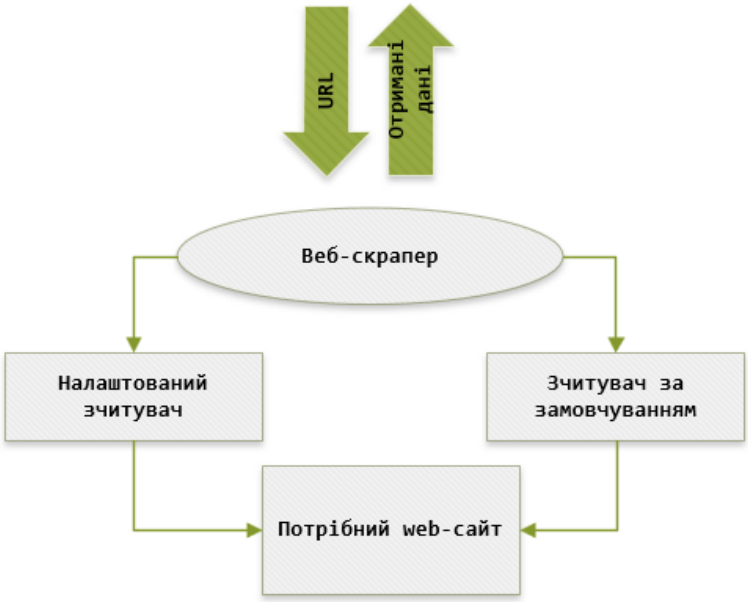

Рис.1 Схема роботи web-скрапера.

На вході система отримує URL потрібної сторінки, а на виході віддає нормалізовані дані (наприклад, в форматі JSON). Отримавши URL, система визначає, якому зчитувачу потрібно направити сторінку на обробку. У випадку, якщо система знає архітектуру web-ресурсу, URL отримує спеціально налаштований зчитувач, в іншому випадку – сторінку починає аналізувати зчитувач, який використовується за замовчуванням. Як правило, в таких випадках використовується найбільш стабільний зчитувач [2].

Оскільки компанії пропонують різноманітні інструменти та послуги для веб-скрейпінгу, може бути важко визначити, які з них є ефективнішими та надійнішими за інші.

1. Розширення для веб-браузерів. Розширення для браузера є чудовим інструментом для вилучення невеликих фрагментів даних, воно дозволяє вам встановити його та вибрати, як ви хочете отримати дані з веб-сайту на ваш вибір (наприклад, CSV, JSON) або в будь-якому іншому зручному форматі.
2. Готові програмні рішення. Добре зарекомендували себе веб-скрейпери використовують найпопулярніші та прості у використанні формати для зберігання зібраних даних, такі як JSON, CSV.
3. Інструменти на основі хмарних технологій. Порівняно з іншими інструментами, які потребують людського втручання на певному етапі веб-скрейпінгу, хмарні інструменти веб-скрейпінгу спрощують процес підтримки веб-скрейпінгу та роблять його повністю автоматизованим і безпроблемним. [3].

Висновки

Розглянуто класифікацію різноманітних засобів, що дозволяють автоматизувати обробку великих об’ємів даних. Ці програмні інструменти дають можливість повноцінно або без втручання людини отримувати та опрацьовувати необхідну інформацію, отриману із мережі Internet.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Web Content Mining Techniques: A Survey [Електронний ресурс]. – 2021 – Режим доступу до ресурсу: https://www.ijcaonline.org/archives/volume174/number24/31826-31826-2021921155
2. What are the Different Scraping Techniques [Електронний ресурс]. – 2019 – Режим доступу до ресурсу: https://www.shieldsquare.com/what-are-thedifferent-scraping-techniques/

3. Types of Web Scraping Tools [Електронний ресурс]. – 2018 – Режим доступу до ресурсу: https://medium.com/prowebscraper/types-of-webscraping-tools-940f824622fb

Ліміна Вероніка Юріївна — студентка групи 1ПІ-18б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: limina.nika@gmail.com

Науковий керівник: ***Бабюк Наталя Петрівна*** – доцент кафедри програмного забезпечення, Вінницький національний технічний університет, Вінниця , e-mail: babiuk@vntu.edu.ua

Limina Veronika Y. – student of the group 1PI-18b, Department of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: limina.nika@gmail.com

Supervisor: ***Babiuk Natalia P.*** – Associate Professor of Software, Vinnytsia National Technical University, Vinnytsia