УДК 004.62

О. Бойко
Д. Волошин

Аналіз методів автоматичного сканування веб-ресурсів

Вінницький національний технічний університет

Анотація

У доповіді розглянуто тему аналізу методів автоматичного сканування веб-ресурсів, а саме веб-скрейпінгу, сучасний стан цієї теми, актуальність, основні положення. Досліджені види, методи. Розглянуті сучасні веб-додатки з сканування сайтів та їх порівняння, переваги та недоліки. Розроблений власний додаток для автоматичного сканування веб-ресурсу.

Ключові слова: *сканування веб-русурсів, web-scraping.*

Abstract

The report includes analysis methods for automatic scanning of web resources, such as web-scraping, the current state of the topic, the relevance of basic provisions. Researched types and methods, the modern web applications with scanning sites and compare their advantages and disadvantages. Developed application for automatically scanning a web resource.

Keywords*: web-resource scan, web-scraping.*

Різні методи і процеси були створені і розвивались протягом довгого часу для збору і аналізу даних. Можна просто визначити веб-скрепінг як процес збору даних з широкого спектра різних веб-сайтів і баз даних. Процес може бути досягнуто або вручну, або з використанням програмного забезпечення. Веб-скрепінг дозволяє збирати в автоматичному режимі вільно доступні дані практично будь-якого виду з їх подальшим аналізом. Розуміючи, як працює ця технологія, можна заощадити час на обробці рутинної інформації, систематизації даних, які зберігаються на певних ресурсах, наприклад, для онлайн порівняння цін, зчитування контактної інформації, моніторингу даних про погоду, виявлення зміни веб-сайту, наукового і дослідження, веб-колажів, інтеграції веб-даних[1].

Веб-сторінки побудовані з використанням тексту на основі мов розмітки (HTML і XHTML), і часто містять безліч корисних даних в текстовій формі. Проте, більшість веб-сторінок призначені для людини, а не для використання автоматизованими програмами. Через це, були створені набори інструментів, які сканували веб-контент. Web scraping являє собою API для отримання даних з веб-сайту. Такі компанії, як Amazon AWS і Google надають веб-скрепінг інструменти, послуги і загальнодоступні дані, доступні безкоштовно для кінцевих користувачів.[2]

Веб-скрепінг (з англ. Web scraping - веб-вишкрібання, веб-скрепінг, веб-витяг) – це технологія аналізу і вилучення веб-даних, яка використовується для технічного комп'ютерного програмного забезпечення з метою отримання інформації з веб-сайтів. Як правило, це програмне забезпечення моделює дослідження людиною мережі Інтернет і реалізується в низькорівневих протоколах передачі гіпертексту (HTTP). Також воно може бути вбудовано в повноцінний веб-браузер, наприклад, такий як Mozilla Firefox. Метод спирається на взаємодію з веб-сторінками, які створюються з використанням текстових мов розмітки (HTML і XHTML), і часто містять безліч корисних даних в текстовому вигляді. Проте, більшість веб-сторінок призначені для людини з точки зору кінцевого користувача, а не для машини. З цієї причини були створені набори інструментів для скрепінга веб-контенту.[3]

Розглянуті сучасні методи аналізу веб-ресурсів, їхні переваги та недоліки.[4]

В доповіді було розглянуто тему веб-скрепінгу даних. Вона була актуальною та продовжую набирати оберти у наш час. Слідкування, збирання, оброблення, зберігання, порівняння інформації з інтернету є важливою тему для багатьох підприємств, тому цей напрям отримав широку популярність. Технологія вилучення даних знаходить активне застосування у компаній-розробників, сприяє зменшенню приросту великої кількості неконтрольованих даних і зберігає зусилля, спрямовані на створення раніше реалізованого контенту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Data scraping [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Data_scraping
2. Web scraping [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Web_scraping
3. Scraping wiki [Електронний ресурс]. – Режим доступу: https://www.scrapesentry.com/scraping-wiki/
4. 10 Web Scraping Tools to Extract Online Data [Електронний ресурс]. – Режим доступу: http://www.hongkiat.com/blog/web-scraping-tools/

Волошин Дмитро – студент групи 1СІ-13б, кафедра автоматики та інформаційно-вимірювальної техніки, Вінницький національний технічний університет, м. Вінниця.

Науковий керівник: ***Бойко Олексій Романович –*** канд. тех. наук, доцент кафедри автоматики та інформаційно-вимірювальної техніки, Вінницький національний технічний університет, м. Вінниця.

Voloshyn Dmytro – student of group 1SI-13b, Department of Automation and Information-Measuring Devices, Vinnytsia national technical university, Vinnytsia.

Supervisor: ***Boyko Oleksiy R. –*** Ph.D. (Eng.), Docent of Department of Automation and Information-Measuring Devices, Vinnytsia national technical university, Vinnytsia.