УДК 004.4’2

В.О. Коротких
Т.О. Філімонова

РОЗРОБКА АВТОМАТИЗОВАНОЇ СИСТЕМИ ПАРСИНГУ ДАНИХ

Київський державний торговельно-економічний університет

Анотація
У статті досліджено актуальність розробки автоматизованої системи парсингу даних. Проаналізовано алгоритм, принципи роботи, переваги застосування автоматизованої системи парсингу даних.
Ключові слова: автоматизована система, інформація, парсинг.

Abstract
The article investigates the relevance of the development of an automated data parsing system. The algorithm, principles of work, advantages of application of automated data parsing system are analyzed.
Keywords: automated system, information, parsing.

Вступ

У сучасному бізнес-середовищі існує досить великий потік публічної інформації. Завдяки мережі Інтернет, що об’єднує людей по всьому світу, це нескінченний потік даних. Легкий доступ до цінних знань створює великі можливості для освіти та інновацій. Велика кількість даних привносить у наше життя освітній контент, розваги та різноманітні зручні інструменти, коли кожен може миттєво знайти потрібну інформацію. Додаткові можливості дозволяють значно економити час для пошуку того чи іншого матеріалу або його перевірки. Саме тому, розробка автоматизованих систем обробки даних на сьогоднішній день є досить актуальним питанням.

Виклад основного матеріалу

Для збору інформації в мережі Інтернет використовують спеціальні програми, які ще називаються веб-роботами або парсерами (краулерами) [1, с. 151].

Парсинг – це процес збирання будь-яких неструктурованих даних, їх упорядкування та подальший аналіз. Цей метод актуальний у випадках, коли масив інформації занадто великий і піддається ручної обробці. У такому разі використовують спеціальну програму – парсер, написаний на Delphi, PHP, C++ або будь-якій іншій мові з підтримкою регулярних виразів. Програмний інструмент відповідає за збирання даних та їх аналіз. Процес здійснюється автоматично та значно економить час спеціалісту [2].

Поширене питання, яке постійно виникає під час обробки документів в організаціях, полягає в тому, чи варто створювати власний аналізатор даних. Спеціальне програмне забезпечення для синтаксичного аналізу тексту, створене для внутрішніх команд, безумовно, спеціально розроблене для відповідності конкретним вимогам розбору в організації. Однак недоліком є те, що весь персонал повинен бути навчений тому, як ним користуватися. Витрати на створення спеціальної програми аналізу можуть бути значними, оскільки потрібно більше часу та ресурсів. Крім того, ці рішення вимагають ретельного планування та потребують власних виділених серверів для швидшого аналізу. Якщо ви переносите системи, вони можуть бути несумісними з новими технологіями і потребуватимуть оновлення [3].

Для розробки автоматизованої системи парсингу даних необхідно розуміти алгоритм роботи парсера, що може різнитися у різних реалізаціях, але основний принцип залишається незмінним. Тому, пропонуємо розглянути дані принципи:

1. Програма сканує дані, що надходять на вхід, будь то текст, веб-сторінка або інший набір інформації, і відокремлює деякі елементи.

2. Що саме виділятиме парсер із масиву даних – залежить від конкретного завдання. Зазвичай можна налаштувати програми таким чином, щоб отримувати потрібні результати.

3. Правила пошуку найчастіше задаються регулярними висловлюваннями – рядками, складеними за певними правилами та дають програмі пояснення, що і як шукати.

4. На основі зібраної інформації формується звіт або таблиця, в якій відображено всі отримані результати.

Знаючи, що це таке парсер, можна не тільки серйозно прискорити та оптимізувати роботу, а й розробити новий аналіз даних. Розробка парсингу має ряд переваг у застосуванні, а саме:

– автоматизує процеси та розвантажує працівників;
– висока швидкість (якісний парсер може обробляти тисячі сторінок за хвилину);
– широкі можливості: обсяги, які може пропустити крізь себе програма, незрівнянні з тими, що може проаналізувати людина.

Єдиний мінус, про який можна говорити в контексті розробки інтернет-оптимізації, – неунікальність отриманих даних. Однак при грамотному аналізі результатів пов’язані з цим проблеми зведуть до мінімуму [4].

Плюси парсингу очевидні у порівнянні з ручним збором та сортуванням даних, а саме:

– дані отримуються дуже швидко;
– можна ставити десятки параметрів для складання вибірки;
– у звіті не буде помилок;
– парсинг можна налаштувати з певною періодичністю, наприклад, збирати дані щопонеділка;
– багато парсерів не тільки збирають дані, але й радять, як виправити помилки на сайті.

Розробка автоматизованої системи парсингу даних здебільшого використовується для наступних цілей:

1. Дослідження ринку. Парсинг дозволяє швидко оцінити, які товари та ціни у конкурентів.

2. Аналіз динаміки змін. Парсинг можна проводити регулярно, щоб оцінювати, як змінювалися якісь показники. Наприклад, зростали або падали ціни, змінювалася кількість онлайн-оголошень або повідомлень на форумі.

3. Усунення недоліків своєму ресурсі. Виявлення помилок у мета-тегах, битих посилань, проблем з редиректами, дублюючих елементів тощо.

4. Збирання посилань, що ведуть на ваш майданчик. Це допоможе оцінити роботу підрядника з лінкбілдингу. Як перевіряти зовнішні посилання та якими інструментами це робити.

5. Наповнення каталогу інтернет-магазину. Зазвичай у таких сайтів величезна кількість позицій і витрачається багато часу, щоб скласти опис для всіх товарів. Щоб спростити цей процес, часто аналізують зарубіжні магазини і просто перекладають інформацію про товари.

6. Упорядкування клієнтської бази. В цьому випадку аналізуються і обробляються контактні дані, наприклад, користувачів соцмереж, учасників форумів і т. д. Але тут варто пам’ятати, що збір інформації, якої немає у відкритому доступі, незаконний.

7. Збір відгуків та коментарів на форумах, у соцмережах.

8. Створення контенту, що будується на вибірці даних. Наприклад, результати спортивних змагань, інфографіки щодо зміни цін, погоди тощо [5].

Висновки

Розробка автоматизованих систем парсингу даних дозволяє здійснити цілісний розбір даних та робить інформацію доступною для організацій і полегшує її читання. В подальшому це дозволяє перетворити дані на такі, що можуть ефективно обмінюватися між клієнтами, а аналізатори розробляються для того, щоб зробити бізнес-операції гнучкими та масштабованими за своєю природою. Завдяки ефективній розробці автоматизованих систем парсингу даних велика частина ручної роботи, пов’язаної з вилученням та очищенням даних, автоматизується, і її важливість неможливо недооцінити.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ:

1. Мельник К. В., Мельник В. М., Григоришин А. М. Автоматичний збір інформації (парсинг) в мережі. Науковий журнал «Комп'ютерно-інтегровані технології: освіта, наука, виробництво» Луцьк, 2020. Випуск № 39. С. 151−156.
2. Парсинг [Електронний ресурс]. − Режим доступу: https://blondinka.ru/reading/encyclopedia/parsing/
3. Introduction to Data Parsing : Definition, Overview, and Scope of Data Parsing [Електронний ресурс]. − Режим доступу: https://docsumo.com/blog/what-is-data-parsing
4. Що таке парсинг, і як працює ця технологія [Електронний ресурс]. − Режим доступу: https://vikna.if.ua/cikavo/98479/view
5. Для чого потрібен парсер [Електронний ресурс]. − Режим доступу: https://www.centum-d.com/dlya-chogo-potriben-parser/

Коротких Віталій Олександрович *− студент групи 4-8, факультет інформаційних технологій, Київський державний торговельно – економічний університет, Київ.*

Філімонова Тетяна Олегівна *− кандидат фізико − математичних наук, доцент кафедри комп'ютерних наук та інформаційних систем, Київський державний торговельно – економічний університет, Київ, e-mail: tatyana0377@gmail.com*

Korotkykh Vitaliy Oleksandrovych *− student of the group 4-8, Department of Information Technologies, Kyiv State University of Trade and Economics, Kyiv.*

Filimonova Tetyana Olehivna*− Candidate of Physical and Mathematical Sciences, Associate Professor at the Department of Computer Sciences and Information Systems, Kyiv State University of Trade and Economics, Kyiv, e-mail: tatyana0377@gmail.com*