

Аналіз проблем автоматичного пошуку веб-даних в системі технічного регулювання в будівництві

Сергій Буряков, магістр. Андрій Клочко, аспірант. Олександр Федоров, аспірант. Світлана Теренчук, доцент, к.фіз.-мат.н., доцент

Київський національний університет будівництва і архітектури, Київ, Україна

АНОТАЦІЯ

Представлено аналіз проблем автоматичного збору веб-даних. Показано перспективу використання засобів парсинга для пошуку нормативних документів галузі будівництва та промисловості будівельних матеріалів з метою підвищення ефективності пошуку веб-документів на запит користувача та рівня структурованості системи технічного регулювання в галузі будівництва.

Ключові слова: веб-скрапінг, парсер, пошукова система.

1. ВСТУП

Вирішення проблеми пошуку інформації в мережі Інтернет часто зводиться до розробки таких алгоритмів роботи пошукових систем, як: збирання та індексація інформації з веб-ресурсів; пошук веб-сторінок, які відповідають ключовим словам пошукового запиту; ранжування отриманих результатів в порядку релевантності пошуковому запиту.

Для автоматичного вилучення даних з недостатньо структурованої системи електронного документообігу і збереження структурованих даних у таблицях або базах даних, використовують скрапінг або парсинг [1].

Веб-скрапінг (web-scraping) – процес вилучення даних із сторінок веб-сайтів в інтернеті за допомогою автоматизації звернень до веб-сайту [2].

Парсер (parser) або синтаксичний аналізатор, являє собою частину програми, яка перетворює вхідні дані в структурований формат [3].

Програма-парсер призначається для виокремлення зі сторінки тільки тієї інформації, яка потрібна користувачу. Для цього створюється математична модель порівняння лексем з формальною граматиною, що описана однією з мов програмування. Наприклад, PHP, Perl, Ruby, Python. При цьому програми звертаються до відкритих сторінок цільового сайту, отримують код, розбирають його на складові, шукають потрібні складові та зберігають їх в своїй базі даних. Таким чином, парсер здатен виконувати синтаксичний аналіз текстів. При цьому вирішується актуальна задача отримання «чистої» копії даних із неструктурованої інформації.

2. МЕТА РОБОТИ

Метою роботи є дослідження можливості використання програм-парсерів для пошуку нормативних веб-документів в системі технічного регулювання в будівництві.

3. ОСНОВНИЙ МАТЕРІАЛ

В Інтернеті для відображення даних використовується HTML-код. Проте, ця мова гіпертекстової розмітки не надає дієвого механізму, що здатен полегшити автоматичний аналіз документів. Це обмеження не дозволяє відрізнити контент від макета і семантики даних.

Такі стандарти, як RDF, RDFS і OWL, розроблені для забезпечення загального синтаксису при визначенні моделей даних. Ці стандарти дозволяють визначати онтології, які підтримують запити, але не зрозумілі для розробників, які на початковому етапі ігнорують процес семантичної анотації при розробці HTML-сторінок.

3.1. Аналіз проблем автоматичного збору веб-даних

Інформація, що міститься в базі даних, не може бути проіндексована пошуковою системою, оскільки ця функція передбачає [4]:

- 1) взаємодію програмного рушія з формою пошуку;
- 2) здатність задавати параметри пошуку;
- 3) розуміння семантики даних, що повертаються.

Ця проблема вирішувалася проектом MEMEX DARPA [5], в якому показана спроба індексації інформації, що міститься в DeepWeb; підходом Schema [6], що визначає словник понять, та дозволяє анотувати дані, що містяться в документі HTML.

Структурування електронної бази даних, передбачає пошук документів в різних джерелах та ідентифікацію таких основних елементів в задачі екстракції знань:

- 1) джерело даних, що містить відповідну інформацію;
- 2) базу даних, яка призначена для зберігання даних;
- 3) експерта, який може визначити, як саме перетворювати інформацію.

Навіть якщо перетворення між джерелом і базою даних виконується автоматично, експерт відповідає за визначення правильності вилучених даних. До того ж, незалежно від рівня автоматизації системи документообігу, в функції експерта входить: формалізація вихідної семантики ще до початку екстракції; подолання невизначеності, яка може виникати при виключенні деякого конфліктного параметру технічного регулювання з одного документу [7].

Вилучення конфліктних параметрів із нормативних документів не можна здійснювати без аналізу умов і рекомендацій, які були наявні в менш пріоритетних документах. Такий аналіз запобігає необґрунтованому переміщенню вказівок до більш пріоритетного документу, яке може призвести до втрати змістовності чи точності інформації. При цьому, корегування обох конфліктних документів, може призвести до появи суперечливостей саме в більш пріоритетному документі.

3.2. Проблеми розробки модуля для автоматичного збору веб-даних

Процес вилучення інформації з окремої веб-сторінки можна розділити на такі етапи:

- 1 Побудова запиту для отримання інформації.
- 2 Виконання запиту і отримання відповіді.
- 3 Обробка відповіді, вилучення та структурування необхідної інформації.
- 4 Передача отриманої інформації для подальшої обробки.

Процес вилучення інформації може бути реалізований різними шляхами.

Один з них здійснюється невеликою кількістю таких дій як: завантаження URL, зчитування інформації та надання її користувачу. Другий шлях складається з авторизації в системі, конструювання запиту за інформацією з заголовків і значень JavaScript-змінних на сторінку, ім'я яких може змінюватися з кожним запитом, а JS-код перебувати в мініфіційованому або обфужированому вигляді.

Якщо в перший шлях є порівняно простим, то другий шлях розробки парсеру за короткий час потребує: використання headless-браузерів (без графічного інтерфейсу) з підтримкою сценаріїв, на кшталт PhantomJS для вилучення даних, оптимізації часу на вивчення взаємодії сайту з бекендом. Для цього також підходить Selenium WebDriver з одним із реальних браузерів.

Однією з основних проблем, що супроводжують створення парсерів є забезпечення ефективності на значних об'ємах інформації [8]. Оскільки сконструйований парсер повинен обробляти велику множину типів/форматів веб-сайтів, але не може враховувати специфіку кожного з них, то отриманий код може бути не оптимальним і неефективним. Такі обмеження значно впливають на швидкість отримання даних і робить парсери більш вразливими.

Зайве збільшення навантаження на сервера цільових сайтів веде до збільшення навантаження на інфраструктуру парсинга. Це призводить до нерентабельності витягу даних в стилі терміни.

Згідно з [9] для забезпечення ефективності засоби веб-скрапінгу потрібно використовувати з огляду на специфіку вихідного джерела та системи в якій він застосовується.

Засоби парсинга, в яких кожен веб-документ, що аналізується, зберігається в пам'яті у вигляді ієрархії його елементів (вузлів), при повторному зверненні до різних частин документів програмами витрачають менший час на обробку всього веб-документу, оскільки зникає потреба завантажувати і будувати дерево. Такі засоби надають найбільш зручний спосіб доступу до елементів веб-документа за допомогою запитів (CSS і XPath) і підходять для будь-яких завдань багаторазової вибірки інформації з веб-документа.

Засоби парсинга з потоковою обробкою застосовні в задачах, де повторне звернення не потрібно. В такому випадку відбувається послідовний перебір вузлів до потрібного і потім завантаження в пам'ять тільки поточного вузла. Дані кошти можуть аналізувати дуже великі документи і підходять для таких завдань, як індексація або перетворення в інші формати.

Засіб парсинга, що представлені регулярними виразами можна застосувати для задач, що потребують екстракції частини документа із задалегідь відомою структурою. Швидкий пошук відповідників з шаблонами, написаними за допомогою регулярних виразів, гарантує ефективний

результат за умови надійно складених виразів. Синтаксис регулярних виразів складніший, ніж у CSS селекторів і XPath запитів, а також чутливіший до невеликих змін в розмітці. Такий засіб парсинга доцільніше використовувати там, де стандартні розширення бібліотеки libxml, не здатні вирішити такі задачі, як добування коментарів в коді або фрагменті коду перед розбором будь-якими іншими засобами.

4. ВИСНОВКИ

Використання програми-парсеру дозволить підвищити:

- 1) ефективність пошуку веб-документу в електронній базі даних нормативних документів в будівництві та промисловості будівельних матеріалів на запит користувача;
- 2) рівень структурованості системи документообігу галузі будівництва та промисловості будівельних матеріалів.

Список літератури

- [1] Zosimov V., Stepashko V., Bulgakova O. Inductive Building of Search Results Ranking Models to Enhance the Relevance of Text Information Retrieval. Database and Expert Systems Applications, Valencia, Spain / Ed. by Markus Spies et al. Los Alamitos: IEEE Computer Society, 2015. – 316 p. P. 291 – 295.
- [2] Москаленко А.А., Лапонина О.Р., Сухомлин В.А. Система управления доступом к ресурсам веб-приложений на основе анализа поведения пользователя. International Journal of Open Information Technologies ISSN: 2307-8162, vol. 8, no.9, 2020. P. 30 – 35.
- [3] Правда про парсинг сайтов, или «все интернет-магазины делают это» [Электронный ресурс] – Режим доступа: <https://habr.com/ru/post/446488/>.
- [4] Raghavan S., Garcia-Molina H. Crawling the hidden web. Proceedings of the 27th International Conference on Very Large Data Bases, VLDB'01. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, 2001. P. 129–138.
- [5] Memex. Domain-Specific Search [Электронный ресурс] – Режим доступа: www.darpa.mil/program/memex.
- [6] Guha R., Brickley D., Macbeth S. Schema.org: Evolution of structured data on the web. ACM.2008, 59(2). P. 44–51.
- [7] Klochko A., Terenchuk S., Chernyshev D., Zapryvoda V. Using Deep Structured Semantic Model to Analyze Text Documents in the Building Normative Base. 2020 IEEE International Conference on Problems of Infocommunications Science and Technology, PIC S&T'2020, October 6-9, 2020 Kharkiv, Ukraine.
- [8] Кульгин М. Конструкторы для парсинга цен и ассортимента конкурентов: что делать, если они не справляются? [Электронный ресурс] – Режим доступа: <https://vc.ru/marketing/77766-konstruktory-dlya-parsinga-cen-i-assortimenta-konkurentov-chto-delat-esli-oni-ne-spravlyayutsya>.
- [9] Лыгина Н.И., Пудич А.С. Исследование правильности и эффективности средств парсинга информации на веб-ресурсах // Инновационная наука. 2017. №3-1.