Кластеризація сторінок веб-ресурсу із застосуванням об'єктної моделі

Станіслав Диханов, Наталія Гук

dykhanovstas@gmail.com, natalyguk29@gmail.com

Дніпровський національний університет імені Олеся Гончара

На сьогодні мережа Інтернет широко застосовується для розташування різноманітної інформації, за допомогою мережи розповсюджуються новини, здійснюється продаж та реклама товарів, надаються освітні послуги тощо.

Тому актуальною є розробка нових моделей та методів, які дозволяють швидко надавати релевантну до запитів користувача інформацію. Подібні задачі виникають під час розробки рекомендаційних систем, при аналізі структури веб-сайтів з метою покращення зручності пошуку необхідної інформації.

Кластеризація сторінок веб-сайтів є інструментом для перевірки його структури, оскільки отримане розбиття сторінок на кластери за деякими ознаками дозволяє зрозуміти, чи вірно розташовані сторінки, чи вірно налагоджені зв'язки між ними.

В роботі для здійснення процедури кластеризації пропонується застосування об'єктної моделі веб-ресурсу – DOM (Document Object Model). Зазначена модель будується з HTML-тегів за допомогою спеціальних запитів. DOM модель веб-ресурсу є деревом, його коренем <html> є заголовок HTML-документу (посилання на сайт). Ліве піддерево <head> зберігає мета-теги для браузерів та пошукових систем, назву документу, скрипти та стилі, а праве піддерево <body> зберігає контент веб-сторінки (текст, зображення, медіа-файли), тобто інформацію, яка відображається у вікні браузера.

Зазначена модель підтримує об'єктно-орієнтоване представлення вебсторінки та дозволяє порівнювати структури веб-ресурсів між собою.

Для порівняння структур двох дерев застосується метод вирівнювання дерев з використанням поняття відстані редагування дерев (Tree Edit Distance) в якості метрики [1].

Відстань між деревами T1 та T2 дорівнює кількості операцій, які необхідно виконати, щоб перетворити дерево T1 на дерево T2. Множина операцій, які можна виконувати над деревом, складається з операцій перейменування вузла (Remove), видалення вузла (Remote) та додавання вузла (Update).

Значення «вартості» пов'язане з послідовністю виконаних операцій по зміні дерева, необхідних для перетворення його із початкового стану до «вирівняного». Оскільки кількість редагувань дерева не обмежена, то здійснюється нормалізація «вартості» шляхом встановлення критерію

максимуму, а структурна схожість двох DOM дерев обчислюється у такий спосіб:

$$S_{struct} = 1 - \frac{TED(T1, T2)}{y_{max}(|T1| + |T2|)} \quad (1)$$

де y_{max} – максимальна кількість операцій Remote, Remove та Update.

Зазначену метрику використано для здійснення кластеризації сторінок веб-ресурсу. Для обчислення відстані між деревами застосовується алгоритм [1].

Для практичної реалізації запропонованого підходу було застосовано мову програмування Python, спеціальні бібліотеки та методи. Бібліотеки Lxml та urllib використовувалась для обробки HTML файлів, звернення до URL-адрес сайтів, парсингу веб-сторінок, бібліотека Pandas для реалізації методів очищення даних, модуль difflib застосовано для пошуку й обробки розбіжностей у послідовностях, бібліотеку sklearn для реалізації методів кластеризації.

Розв'язано задачі кластеризації веб-сторінок сайтів інтернет-магазину та закладу освіти за їх DOM моделями. Сайти мають деревовидну структуру, інтернет-магазин має головну сторінку, сторінки з категоріями товарів, картки товарів. В результаті кластеризації було отримано розбиття за типами сторінок: утворились кластери з головної сторінки, сторінок категорій товарів, сторінок товарів. Подібність між елементами одного кластеру є надзвичайно високою та становить 95-97%. Це обумовлено тим, що веб-сторінки одного сайту мають майже ідентичну структуру, вони формуються динамічно, а контент яким вони наповнені, зберігається у базі даних та змінюється у різних HTML-тегах при переході на певне посилання з товаром або категорією товарів.

Сайт закладу освіти утворюється з головної сторінки, сторінок кафедр, сторінок з інформацією про навчання на різних рівнях освіти, сторінок для абітурієнтів, новин. Розбиття, яке було отримано, відповідає зазначеної структурі сайту, але подібність між елементами одного кластеру виявилась нижчою, ніж у інтернет-магазину, та становила 68-72%. Аналіз сторінок в середині кожного з кластерів показав, що в їх DOM деревах присутні HTML-теги, які відрізняються для веб-сторінок в середині одного кластеру. Такими елементами є додаткові панелі, що з'являються лише на певних веб-сторінках, та додаткові опції, які присутні на деяких сторінках, але відсутні на інших.

Проведений аналіз може бути корисним під час реінжинірингу існуючих сайтів та налаштуванні посилань між сторінками в середині сайту.

1. Kaizhong Zhang, Dennis Shasha Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems December. – SIAM Journal on Computing 18(6). – 1989. – Pp. 1245-1262. DOI:10.1137/021808