УДК 004.67

ІНТЕЛЕКТУАЛЬНИЙ МОДУЛЬ АГРЕГАТОРА СПЕЦІАЛІЗОВАНОЇ ІНФОРМАЦІЇ З ВІДКРИТИХ САЙТІВ

Мелікбекян А.А., Токар М.М.
Тернопільський національний економічний університет, магістранти

І. Актуальність проблеми

Існує багато типів веб-сайтів, які орієнтуються на надання певної інформації: новини, блоги, інтернет-магазини, карти, біржі, фото, відео і т.д. З кожним днем кількість таких сайтів зростає, отже, збільшується кількість різної інформації, що створює проблеми для рядового користувача Інтернету в пошуку специфічних продуктів та інформації.

Щоб вирішити дану проблему використовуються агрегатори сайтів. Їх метою є надання зручного для користувача інтерфейсу для пошуку і доступу до даних, які розташовані на інших сайтах [1]. Як приклад можна розглянути такі сервіси: prom.ua – сервіс для пошуку і покупки товарів, getmancar.com.ua – сервіс для пошуку і оренди автомобілів, Slickdeals – сервіс для пошуку і проведення різного виду угод, TheStocks – сервіс пошуку стокових зображень.

Проте, переважна більшість веб-сторінок проектуються з орієнтацією на зручність сприйняття людиною, що часто ускладнює автоматичне агрегування інформації і розуміння структури веб-сторінок машинами.

Зазвичай у сервісів-агрегаторів висока користувацька конверсія (відношення числа відвідувачів сайту, що виконали цільові дії, наприклад, покупка товару або замовлення послуги, до загальної кількості відвідувачів), що є стимулом для агрегаторів сайтів самостійно ділитися даними для підвищення якості надання інформації користувачам і підтримки її актуальності. В іншому випадку, агрегатору сайтів доводиться підлаштовуватися під верстку або API сайту, на якому розміщена інформація, щоб бути впевненими в тому, що ці дані вірні і відповідають один одному (наприклад, ціна, заголовок і опис повинні ставитися до одного і того ж товару на веб-сторінці).

Такий підхід до збору даних має один серйозний недолік: часто сайти, що розглядаються агрегатором, прагнуть змінити шаблон сайту на більш зручний і функціональний, що передбачає зміну верстки веб-сторінок. До того ж із зростанням числа сайтів, зростають витрати часу на освоєння програмістом API або структури веб-сторінок для правильного налаштування пошукового робота.

Тому актуальною є задача: вивчивши існуючі рішення проблеми отримання даних з відкритих веб-сторінок, створити універсальне рішення проблеми збору інформації, яке здатне добувати дані незалежно від структури і стилю веб-сторінок сайтів заданої тематики. Виходячи зі складності та обсягу задачі, було прийнято рішення заздалегідь позначити тематику сайтів та тип добутих даних. Наприклад, сайти фотобанків, які вирізняються високою варіативністю верстки веб-сторінок і різноманітністю контенту. За інформацію, що добувається приймемо назву (заголовок) зображення на веб-сторінці, як найбільш інформативну частину даних, після самого зображення.[2]

ІІ. Мета дослідження

Метою даної роботи є розробка інтелектуального робота, який здійснює добування заголовків продаваних зображень з веб-сторінокфотобанків за допомогою алгоритму машинного навчання. Для досягнення зазначеної мети необхідно вирішити такі завдання:

- вивчення існуючих рішень збору і аналізу даних з веб-сторінок;
- визначення простору ознак;
- вибір і навчання моделі машинного навчання;
- розробка агрегатора;
- проведення тестування якості алгоритму машинного навчання;
- проведення тестування агрегатора.

ІІІ. Загальний опис агрегатора спеціалізованої інформації

Перерахуємо функціональні вимоги до програми:

1) агрегатор повинен здійснювати добування релевантних ознак з текстових об'єктів веб-сторінок;

2) агрегатор повинен використовувати попередньо навчену модель машинного навчання (градієнтне підсилювання (GradientBoosting) дерев рішень – модель машинного навчання, що представляє підсилювання як процес градієнтного спуску. В основі алгоритму лежить послідовне уточнення функції, що представляє собою лінійну комбінацію базових класифікаторів (дерев рішень), з тим щоб мінімізувати функцію втрат), для оцінки ступеня приналежності об'єктів веб-сторінки до заголовку; [3]

3) агрегатор повинен приймати на вхід URL адресу веб-сторінки;

4) агрегатор повинен повертати текст, що міститься в об'єкті веб-сторінки, який отримав найвищу оцінку на етапі класифікації.

Модульну структуру у вигляді діаграми компонентів схематично зображено на рис. 1.

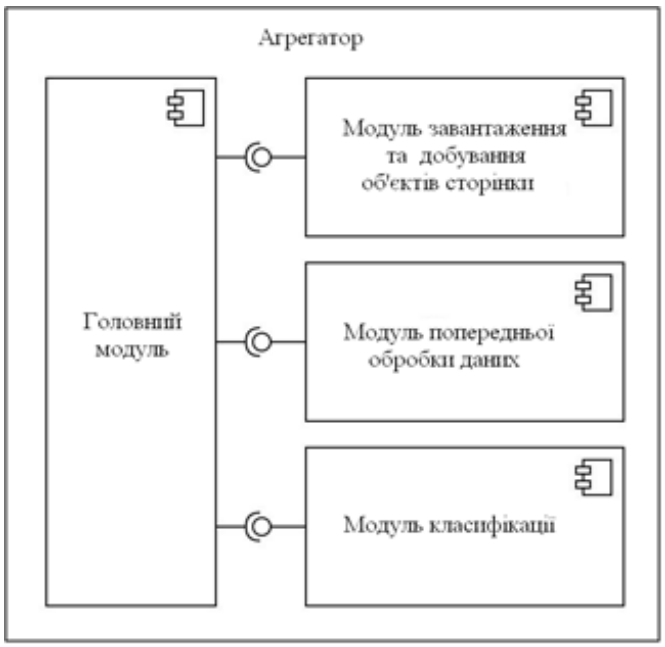

Рис.1. Діаграма компонентів

Модуль завантаження та добування об'єктів сторінки реалізує завантаження і отримання необхідних даних з веб-сторінки. Завдання даного модуля полягає в добуванні інформації про кожний вузол DOM-дерева веб-сторінки, який містить будь-який видимий текст.

Модуль попередньої обробки даних призначений для добування релевантних ознак з текстових об'єктів веб-сторінок;

Модуль класифікації призначений для оцінки ступеня приналежності об'єктів веб-сторінки до заголовку.

Головний модуль об'єднує роботу всіх модулів, здійснює контроль передачі даних між модулями. Даний модуль відповідальний за обробку вхідних (URL адреса веб-сторінки) і вихідних даних (текст заголовка).

Висновки

Отже, в роботі було проведено дослідження особливостей застосування агрегаторів спеціалізованої інформації з відкритих сайтів. Здійснено загальний опис основних модулів реалізації інтелектуального модуля агрегаторів спеціалізованої інформації з відкритих сайтів.

Список використаних джерел

1. Chiou L., Tucker C. Content Aggregation by Platforms: The Case ofthe News Media. [Электроний ресурс] - Режим доступа: http://dx.doi.org/10.3386/w21404
2. Fan J., Luo P., Joshi P. Title identification of web article page susing HTML and visual features. [Электроний ресурс] - Режим доступа:http://dx.doi.org/10.1117/12.876708
3. Robert E. Schapire. The Boosting Approach to Machine Learning: An Overview , MSRI (Mathematical Sciences Research Institute) Workshop on Nonlinear Estimation and Classification – 2003.