УДК 004.9
Миронова Н.О.[1], Волков П.С.[2]
[1] канд. техн. наук, НУ «Запорізька політехніка»
[2] студ. гр. КНТ-216 НУ «Запорізька політехніка»

ПРОГРАМНА РЕАЛІЗАЦІЯ ЗАСОБІВ СИНТАКСИЧНОГО АНАЛІЗУ ЗОВНІШНІХ ДАНИХ

На сьогодні збір даних стає більш складним, і ємним процесом. З цим зіштовхнулися майже всі, користувачі інтернету. Це не завжди пошук реферату чи якоїсь інформації. Багато хто, хоче купити щось дешевше аніж у магазині та на допомогу приходять різноманітні сервіси для пошуку товарів з порівнянням цін. Цим займаються великі сервіси, які використовують синтаксичний аналіз усіх веб-сайтів, які можуть продавати цей товар. Найбільш активно парсинг використовується у пошукових системах, таких як: Google, Yahoo, Bing тощо. Пошукова система, по запиту користувача шукає потрібну йому інформацію. Також парсинг використовують для швидкого заповнення товарів для інтернет магазинів, чи для аналізу ринку або ж великого та середнього бізнесу.

Тема синтаксичного аналізу чи парсингу існує давно, але набирати свою популярність почала нещодавно. Парсинг дозволяє за лічені хвилини зібрати велику кількість інформації, чи проаналізувати щось на плагіат, помилку в синтаксисі програми, чи в орфографічному правописі. Тобто це набір певних правил, які повинні виконуватись за для збору та аналізу даних.

Алгоритм парсеру потребує швидкої обробки інформації за певними правилами, тобто регулярні вирази. Щоб парсер розумів регулярні вирази, він повинен бути написаний, на мові, що підтримує роботу з рядками. Така можливість є у Php, Ruby тощо. Регулярні вирази, описуються синтаксисом UNIX, який хоч і вважається застарілим, але широко використовується завдяки властивості зворотній сумісності.

Для розробки парсеру необхідно обрати певну структуру даних. В нашому випадку це інтернет-всесвіт, який будується за певною структурою. Ця структура написана мовою HTML(HyperText Markup Language). Структура схожа на XML, та є похідною від неї. Парсинг цієї структури, виконується по відкритому та закритому тегу <назва тегу>. Дані, які містяться у тегу </назва тегу>. По цьому можна аналізувати та збирати необхідні дані.

В роботі необхідно було виконати збір даних з сервісу перегляду серіалів за такими складовими: назва, жанр, рік випуску, статус серіалу.

Отже, був виконаний аналіз сервісу для перегляду серіалів, тобто його DOM дерево. Також було проаналізовано внутрішні запити цього сервісу для

більш детального аналізу та створення правил.

Для написання алгоритму було використано мову програмування php та бібліотеку для роботи з DOM деревом. Після парсингу, усе було об’єднано в зручні масиви. (рис. 1.)

Рисунок 1 – Дані розсортовані по масивам

Отже, для написання такого парсеру дозволяє спростити та автоматизувати пошук потрібної інформації.