УДК 004.4

ІНТЕЛЕКТУАЛЬНИЙ МОДУЛЬ ІНТЕРНЕТ-ПАРСИНГУ КОНТЕНТУ ВЕБ-РЕСУРСІВ

Арсенюк Ігор, Кратасюк Віктор

Вінницький національний технічний університет

Анотація

В ході роботи проведено дослідження та аналіз існуючих аналогів на ринку, обґрунтування доцільності розробки інтелектуального модуля. Виконано обґрунтування вибору математичної моделі для інтернет-парсингу контенту веб-ресурсів. Розроблено алгоритм функціонування інтелектуального модуля та обґрунтовано вибір генетичного алгоритму.

Abstract

In the course of the research and analysis of existing analogues in the market, study the feasibility of developing intelligent module. Done choice of mathematical model for Internet content parsing web resources. The algorithm of intelligent modules and the choice of the genetic algorithm.

Вступ

Об’єктом дослідження є процес аналізу та отримання інформації з веб-ресурсів.

Предметом дослідження є моделі та програмні засоби отримання контенту веб-ресурсів.

Метою дослідження є покращення якості і спрощення методів отримання інформації, що розміщюється на веб-ресурсах шляхом розроблення інтелектуального модуля для інтернет-парсингу контенту веб-ресурсів.

В ході дослідження виконано аналіз предметної області інтернет-парсера контенту веб-ресурсів, виконано класифікацію веб-ресурсів в мережі, досліджено основні засоби парсингу контенту на основі аналізу доступних бібліотек та аналогів у мережі, визначено їх переваги та недоліки.

До недоліків, в першу чергу, можна віднести високу вартість платних аналогів та не універсальність безкоштовних продуктів представлених на ринку.

На основі аналізу обґрунтовано актуальність та доцільність розробки інтелектуального модуля для інтернет-парсера контенту веб-ресурсів та окреслено основні задачі, які потрібно розв’язати для досягнення мети.

В ході дослідження розглянуто різні математичні підходи (зокрема метод рекурсивного спуску та повного перебору) до вирішення задачі парсингу та обрано модель, що базується використанні генетичного алгоритму та регулярних виразів для відсіювання потрібної інформації із загального об’єму інформації, що отримується з ресурсу.

Відповідно до обраного методу розв’язання задачі розроблено алгоритм інтернет-парсингу контенту веб-ресурсів, обґрунтовано основні методи та етапи отримання інформації шляхом аналізу HTML коду та поетапної роботи користувача з використанням генетичного алгоритму (ГА).

Під час аналізу розглянуто різні шляхи розв’язання задачі для різноманітних типів вхідного контенту в результаті чого було виявлено, що у випадку коли простір пошуку досить вузький, то розв’язання може бути отримане методом повного перебору, при цьому можна бути впевненим, що отриманий результат є найкращим, тоді як ГА міг з більшою ймовірністю зійтися до локального оптимуму, а не до глобально кращого рішення. Метод швидкого спуску [1] буде більш ефективний, ніж ГА у випадках обробки вхідної інформації. Якщо відносно простір пошуку є деяка додаткова інформація (як, наприклад, простір для добре відомої задачі про комівояжера), методи пошуку, які

використовують евристики, що визначаються простором, часто перевершують будь-який універсальний метод, яким є ГА [2].

За ціль було поставлено розробку універсального модуля, який міг би працювати з різним вхідним конвентом, тому ГА є одним з найкращих варіантів вирішення, оскільки дає можливість обробляти різні за типізацією дані.

У ході роботи було пристосовано ГА для розв'язання поставленої задачі (Рис 1). Суть його роботи полягає в тому, що після отримання даних від користувача про об'єкт парсингу відбувається звернення до блоку реалізації генетичного алгоритму де відбувається підбір популяції хромосом, яка відповідає параметрам поставленої задачі [3]. Якщо така популяція знайдена відбувається парсинг, у протилежному випадку відбувається перехід до кроку 4 та формування нової популяції.

В ході розробки функціональної частини модуля сформовано основні класи, здійснено їх обгрунтування, а також відображено зв'язки між ними. Проведено тестування інтелектуального модулю інтернет-парсингу контенту веб-ресурсів, помилок і невідповідностей не виявлено, а тестування на швидкодію показало кращі результати ніж у аналогів при виконанні подібних задач.

Виходячи з вищепроаналізованого можна зробити висновок, що мета роботи по спрощеню методів отримання інформації веб-користувачами даних з Інтернет-ресурсів шляхом розроблення інтелектуальної системи для інтернет-парсингу контенту веб-ресурсів, була досягнена.

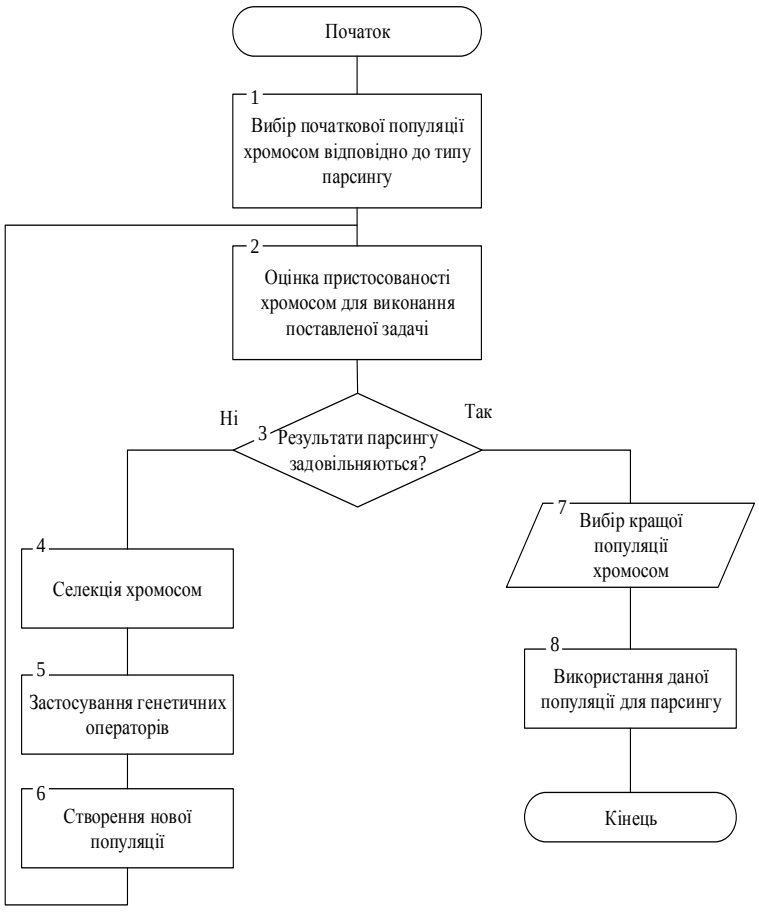

Рисунок 1 – Схема генетичного алгоритму

Список використаних джерел:

1. Рассел С. Искусственный интеллект. Современный подход. 2-ое изд. / С. Рассел, П. Норвиг; [пер. з англ. К.А.Птицина].– М.: Видавничий дім «Вильямс», 2006. – 1409 с. – ISBN 5-8459-0887-6.

2. Люгер Дж.Ф. C# стратегии и методы решения сложных проблем, 4-е изд. / Дж.Ф. Люгер [пер. з англ. К.Д.Протасовой]. – М.: Видавничий дім «Вильямс», 2003. – 864. – ISBN 5-8459-0437-4.

3. Смолін Д.В. Введення у штучний інтелект: конспект лекцій [навчальне видання]/ Д.В. Смолін – М.: ФИЗМАТЛИТ, 2004. – 208 с. – ISBN 5-9221-0513-2