

РОЗРОБКА МОДЕЛІ КЛАСИФІКАЦІЇ ВЕБ-КОНТЕНТУ ДЛЯ ПІДВИЩЕННЯ ТОЧНОСТІ ПОШУКУ

Тоніца О. В., Ушаков Д. Р., Єльчанінов Д. Б.

***Національний технічний університет
«Харківський політехнічний інститут», м. Харків***

Для користувачів Інтернету перевантаженість веб-сайтів інформацією є величезним тягарем. Інформація, що передається без розбору, містить невідповідні, шкідливі веб-сайти та комерційну рекламу, які є зайвими або навіть шкідливими для підлітків. Через рекламні банери, якими переповнений Інтернет, швидкість передачі даних у мережі послаблюється, мережеве середовище захаращується. З іншого боку, хоча цифрові рекламодавці можуть розміщувати свою рекламу всюди, вони все одно стикаються з проблемою низької ефективності бізнесу через жорстку конкуренцію з іншими. Вони не можуть зосередитися на цільових клієнтах, але можуть дозволити собі високі витрати на цифрову рекламу.

Ціллю роботи є дослідження способів класифікації веб-сторінок на основі методів інтелектуального аналізу даних, також покращення цих методів, щоб підвищити точність та побудова моделі, що може класифікувати веб-сторінки за кількома категоріями.

У роботі було проведено аналіз існуючих методів класифікації веб-сторінок. В результаті чого було виявлено, що оптимальним методом є класифікація на основі ієрархічної моделі, яка використовує бінарні моделі класифікації з рефері [1].

Під час класифікації важливо використовувати якісну вибірку без порожніх або неправильно маркованих веб-сторінок. Ефективність такого підходу було підтверджено під час навчання моделей на відфільтрований вибірці, де всі сторінки були коректно класифіковані.

На основі цього дослідження була розроблена модель для класифікації веб-сторінок, яка забезпечує точність мультикласової класифікації у 96% за мікроусередненням.

Виявлено, що найбільш складно класифікувати веб-сторінки, що не містять текстового контенту, оскільки зазвичай користувачі оцінюють вміст за допомогою зображень. Саме тому наявність таких сторінок має велику кількість. Можливі покращення моделі включають додавання атрибутів, які враховують цю особливість, для підвищення якості класифікації у майбутньому.

Література:

1. Qi X., Davison B.D. Web page classification: Features and algorithms // Journal ACM Computing Surveys. 2009.