УДК 681.12

В. Саранчук[1]

ІНТЕЛЕКТУАЛЬНА КОМП'ЮТЕРНО-ІНТЕГРОВАНА ТЕХНОЛОГІЯ КЛАСИФІКАЦІЇ ВЕБ-РЕСУРСІВ

[1] Вінницький національний технічний університет;

Анотація

Проведені дослідження показують, що найбільш простим і ефективним способом класифікації контенту веб-сторінок є класифікація на основі ієрархічної моделі з використанням бінарних моделей класифікації з залученням ріних наборів атрибутів. Варто зауважити, що, як показала практика, в процесі класифікації одним з найважливіших факторів є використання якісної вибірки, яка не буде містити порожніх або невірно промаркованих веб-сторінок.

Ключові слова: класифікація контенту веб-сторінок, інформаційні системи.

Abstract

Studies show that the simplest and most effective way to classify the content of web pages is to classify on a hierarchical model using binary classification models involving different sets of attributes. It is worth noting that, as practice has shown, one of the most important factors in the classification process is the use of a quality sample that will not contain blank or incorrectly marked web pages.

Keywords: classification of web page content, information systems.

Вступ

Інформація в Інтернеті відрізняється високою динамікою: створення нового контенту, його редагування та видалення займають кілька секунд. З огляду на кількість користувачів, які можуть створювати небажаний контент, використання традиційних методів виявлення та класифікації подібної інформації стає незручним.

Визначення тематики контенту веб-сторінок є однією з найважливіших задач багатьох інтернет-компаній. Наприклад, за умови коректної класифікації можна пропонувати користувачеві більш точну підбірку рекламних блоків, що в свою чергу дозволить підвищити продаж як місць розміщення рекламних банерів, так і рекламованого товару. Крім того, захист від небажаної інформації також є однією з основних можливих сфер застосування класифікації контенту.

Для автоматизації перевірки і класифікації веб-контенту, а також для виявлення небажаних для перегляду веб-сторінок і веб-сайтів, можна використати методи інтелектуального аналізу даних [1-3]. Завдання технології інтелектуального аналізу даних - виявити структури даних і знайти закономірності в слабоструктурованих даних. Зважаючи на точність класифікації, що дають існуючі методи, можна зробити висновок, що такі методи потребують модифікації.

Метою є дослідження способів класифікації веб-сторінок за допомогою існуючих моделей, методів і алгоритмів інтелектуального аналізу даних, модифікація цих методів та підвищення їх точності.

Результати дослідження

Одним з найважливіших етапів в підготовці даних для класифікації веб-сторінок є векторизація. Основними параметрами в цьому випадку виступають алгоритм отримання атрибутів, кількість слів (n-gram) і максимальна кількість атрибутів. У бібліотеці sklearn присутні кілька реалізацій векторизації даних:

- HashingVectorizer (HV) - перетворює вхідний текст в вектор зі значеннями кількості входження слова в текст.
- TfidfVectorizer (TF) - перетворює в вектор зі значеннями відношення числа входження слова до загальної кількості слів документа.

В якості додаткового параметра в TF-IDF можна використовувати сублінійну функцію (SB, sublinear), яка замінює стандартний підрахунок TF і дозволяє прибрати «звичайні» слова з розрахунків. На практиці виявилось, що найвищу точність дає використання векторизації з TF-IDF+ sublinear з 5000 атрибутів і n-gram (1, 2).

При скачуванні даних з мережі Інтернет існує проблема в тому, що у нас досить багато даних і визначити, чи вірно були промарковані веб-сторінки, чи не скінчилася оренда домену і т.д., досить складно. Тому було вирішено провести фільтрацію веб-сторінок за «правильно класифікованими». Для цього векторизуємо всю вибірку, а потім навчаємося на отриманому векторі і робимо прогноз по ньому ж. У цього методу є шанс того, що буде перенавчання, але наявність невірно передбачених менше 5%. Тому приймемо їх за недоступні веб-сайти або категорії, які були спочатку з помилкою і відкинемо їх з вибірки на навчання. Також для класифікації веб-сторінок підходить метод PCA (метод головних компонент). Це дозволяє виконати зменшення аналізованої вибірки даних до розміру, який буде оптимальним з точки зору розв'язуваної задачі. Даний метод використовується в якості підготовки даних перед класифікацією.

Найважливішим фактором в інтелектуальному аналізі даних є правильне використання атрибутів. На веб-сторінках можна як атрибути вибрати теги, наприклад, «title», «a», «meta» і блоки «header», «content» і «footer», які були отримані селекторами по тегу, id і class. Так як на цей текст був акцент з боку розробників веб-сайту, то можливо роль їх набагато вища. Використання окремих моделей класифікації для кожного з атрибутів не дає підвищення точності. Дослідним шляхом обчислено, що теги «title» і «meta» (description, keywords) збільшують точність.

Для підвищення класифікації також використаємо URL веб-сторінки як атрибут. В мережі Інтернет для веб-сайтів використовується спрощений формат запису URL: <схема>: // <хост>: <порт> / <шлях>? <параметри> # <якір>. Кожна веб-сторінка використовує свій унікальний URL, при цьому у розробників з'явилося неформальне правило робити зрозумілі для людини адреси, що містять зрозумілі слова. Зазвичай адреса складається з кількох слів без поділу, тому скористаємося символьним n-gram. Так як адреси можуть складатися з абревіатур, скорочень або вигаданих слів, скористаємося ще однією властивістю веб-сторінок - заголовком, в ньому зберігається в текстовій формі основна ідея вмісту. Об'єднаймо атрибути, тим самим додамо ключові слова окремих атрибутів, яких не вистачало.

Ще одним способом отримання нових атрибутів є метод з використанням word2vec. Дана бібліотека представляє слова у вигляді числового вектора, де мінімальна відстань між векторами буде у найбільш схожих за змістом слів. Для класифікації тексту можна використати в якості атрибута середнє арифметичне (average vector) або набір центроїдів (bag of centroids). В даному випадку точність залежить від розміру вибірки навчання. Можливо, за умови використання більших баз даних, результати можуть бути кращими.

Висновки

Проведені дослідження показують, що найбільш простим і ефективним способом класифікації контенту веб-сторінок є класифікація на основі ієрархічної моделі з використанням бінарних моделей класифікації з залученням ріних наборів атрибутів. Варто зауважити, що, як показала практика, в процесі класифікації одним з найважливіших факторів є використання якісної вибірки, яка не буде містити порожніх або невірно промаркованих веб-сторінок. Такий висновок зроблено після навчання моделей класифікації на відфільтрованої вибірці по «правильно передбаченим».

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Барский А.Б. Нейронные сети: распознавание, управление, принятие решений. — М.: Финансы и статистика, 2004. – 176 с.

2. Ясницкий Л.Н. Введение в искусственный интеллект. — Учебное пособие для вузов. 2-е издание, испр. — М.: Академия, 2008. — 176 с. — ISBN 978-5-7695-5390-5..

3. Храмов В.В. Интеллектуальные информационные системы. Интеллектуальный анализ данных. Часть 2. — Учебное пособие. — Ростов н/Д.: Ростовский государственный университет путей сообщения, 2012. — 134 с.

Саранчук Вадим — студент групи 1АКІТ-20м, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця, e-mail: saranchukvntu@gmail.com

Науковий керівник: ***Софина Ольга Юріївна*** — к.т.н, доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця

Saranchuk Vadym – student of group 1AKIT-20m, faculty of computer systems and automation, Vinnytsia National Technical University, Vinnytsia, e-mail: saranchukvntu@gmail.com

Supervisor: ***Sofina Olga Y.*** – Ph. D., Assistant Professor of the Automation and Intelligent Information Technologies Department, Vinnytsia National Technical University, Vinnytsia