АНАЛІЗ МЕТОДІВ КЛАСИФІКАЦІЇ ТЕКСТІВ ЗА ЛІТЕРАТУРНИМИ ЖАНРАМИ

Чухненко М.В., Бабкова Н.В., Угольнікова Н.С.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Автоматична класифікація текстів досить складне завдання. Способи та методи класифікації включаються у напрям Text Mining. Зараз Text Mining активно розвивається: тут проводять дослідження, запускаються проекти і конкурси на виявлення кращих по точності алгоритмів.

Мета текстової категоризації – класифікація текстових документів у певні заздалегідь задані категорії. Останнім часом, із розвитком машинного навчання та методів обробки природної мови, автоматичні методи класифікації тексту забезпечують нові підходи до більш складних проблем літературного аналізу тексту наприклад, для жанрового аналізу п'єс Шекспіру, для аналізу сентименталізму в ранніх американських романах, тощо.

Жанрова класифікація дозволить користувачам сортувати результати інформаційного пошуку безпосередньо за їх інтересами. Люди які заходять в книжковий магазин або бібліотеку, звичайно не просто шукають інформацію на якусь тему, але вони також мають окремі вимоги до жанру книги: вони шукають наукові статті про гіпнотизм, романи про французьку революцію, редакційні статті про суперколайдер, тощо.

В процесі класифікації можуть виникнути деякі проблеми. Перша проблема – попередня обробка вхідних даних. Складність даного етапу полягає і в розмірі даних – документи містять десятки тисяч різних слів, кількість класів так само може досягати тисячі – і це все при недостатньому описі класів (по кілька документів на клас) і невеликій кількості рубрик у кожного документа (зазвичай не більше 5–8). Також в текстах міститься безліч «шуму», який не дає нам уявлення про приналежність документа до класу. Завдання попередньої обробки зводиться до вилучення з тексту тільки необхідних, властивих класу слів. Цього можна досягти шляхом видалення синонімів і однокореневих слів.

Друга проблема – вибір методу класифікації. На сьогоднішній день більшість методів показують дуже малу точність класифікації текстів (~ 50%). Також існує проблема ресурсоємності. При наявності великого кількості класів та ще більшої кількості слів всередині класів, необхідні вельми великі обчислювальні потужності. Кожен з етапів класифікації займає багато часу, що в сумі дає досить довгий час роботи навіть для однієї книги. І, власне, може бути неможливо класифікувати спеціальні тексти (наприклад, математичні, які містять багато формул).

Особливість задачі, яка розглядається у цій роботі, полягає у тому, що для її ефективного рішення необхідно аналізувати велику кількість характеристик текстів, які потрібно врахувати в явному вигляді. Актуальним є застосування методів машинного навчання, оскільки вони дозволяють будувати ефективні програмні системи, які враховують такі неявні закономірності.