УДК 004.912

МОДЕЛЮВАННЯ КЛАСІВ ТЕКСТОВИХ ДОКУМЕНТІВ ПРИ ВИЗНАЧЕННІ ЇХ ПРИКЛАДНОЇ СПРЯМОВАНОСТІ

Курто О. С., студ. гр. МПІн-181
Науковий керівник: **Скітер І. С.,** кф.-м.н., доцент
Національний університет «Чернігівська політехніка»

Питання визначення тематичної та морфологічної спрямованості тексту стає все більш актуальним. Воно використовується як для пошукових систем, так і для відповідності статті науковому журналу в якому вона публікується. Але якщо в масштабних системах такі модулі є вбудованими, то знайти окремий сайт, систему або застосунок доволі складно. Тому для реалізації подібної програми потрібно коректно обрати метод класифікації тексту, враховуючи обсяг об'єктів аналізу, для визначення тематичної спрямованості, компонуючи його з визначенням стилістичної належності.

Для групування об'єктів зазвичай використовуються кластеризація та класифікація. Кластеризація та класифікація виступають протилежними сторонами відносно участі користувача в процесі. Зазвичай механізм класифікації застосовують на автоматично виявлених кластерах даних. Завдання класифікації визначається як загальною тематикою, так і за наявністю певних дескрипторів або певних умов.

Методи машинного навчання або ж просто автоматизація класифікації текстів передбачають наявність навчальної та тестової вибірки даних. На основі навчальної вибірки будується класифікатор, а тестова використовується для перевірки коректності роботи алгоритму. Дані з тестової вибірки не мають використовуватись при побудові класифікатора.

Для класифікації текстів використовують наступні методи:

- дерева та спектри *N*-грам;
- міра близькості об'єкта та категорії;
- наївний класифікатор Байєса.

Ідея основного алгоритму вибору *N*-грам полягає в тому, що символи послідовно зчитуються [1] і таким чином утворюється накопичуваний рядок. Перевіряється наявність цього рядка в словнику: якщо рядок відповідає певному запису в словнику, то зчитується наступний елемент, в іншому випадку в словник додається отриманий рядок.

Існують модифіковані алгоритми вибору *N*-грам, наприклад, на основі суфіксного масиву або префіксного дерева. Подібні алгоритми більш вигідні з точки зору швидкості обробки та сортування масивів різноманітних списків, масивів суфіксів та складання дерев.

При побудові моделі тексту необхідно інтерпретувати текст документу як послідовність символів без розділення її на окремі слова, тобто не оброблювати будь яким особливим чином пробіли. При цьому *N*-грами необхідно будувати не для кожного слова окремо, а для всього документу як для єдиної послідовності символів [2].

Істотним недоліком методик автоматичної обробки тексту, що використовують *N*-грами є заздалегідь встановлене обмеження на можливі значення *N* та на можливі значення самих *N*-грам. Що стосується дерев, то отримуваний таким чином список *N*-грам мало інформативний, оскільки не відображає зв'язків між *N*-грамами і не дозволяє виділити структуру моделі тексту, тому на його основі пропонується будувати модель тексту в вигляді лісу дерев або одного дерева з фіктивним коренем.

Нехай є алфавіт $An = \{a_1, a_2, \ldots a_n\}$ з n символів, тоді модель тексту буде містити максимум n дерев, кожне з яких також може містити максимум n піддерев і т.д. Вузлами і листям дерев є символи вхідного алфавіту [2, 3].

Описане подання тексту у вигляді дерева *N*-грам є базовою моделлю для створення різних уявлень вихідного документа, що вимагається для вирішення конкретних прикладних задач. Описаний вище спектр моделі з наперед визначеним рівнем деталізації є одним з таких представлень. Створення представлення документа на основі його моделі можна, з одного

боку, розглядати як зниження розмірності простору елементів, і в цьому випадку модель не накладає ніяких обмежень на використовувану техніку зниження розмірності. З іншого боку, можна розглядати цей процес як додаткову фазу індексування, яка повторно проводиться вже не для документа, а для його моделі. В останньому випадку вибір також не обмежений *N*-грамами як індексованими елементами [4].

Важливими поняттями при моделюванні класів текстових документів є міра близькості об‘єкта та категорії. Нехай кожній категорії C_i відповідає вектор C_i (c_{i1} , ...,c_{iN}), де N – це розмірність простору термів. Тоді в якості правила класифікатору використовується скалярний добуток:

$$CSV_i(d) = d \cdot C_i = \sum_{j=1}^{N} c_{ij} d_j.$$

Виходячи з цього кінцева формула виглядає таким чином:

$$CSV_i(d) = \frac{d \cdot C_i}{|d| \cdot |C_i|}.$$

Метод наївного класифікатора Байєса вивчається лише з 50-х років та використовується для класифікації текстів, але в наш час він вже широко використовується різними дослідниками для відшукання інформації. Його базова версія використовує техніку відшукання термінів за їх частотою шляхом підрахунку кількості слів у документах.

Якщо є згадати про морфологічну спрямованість тексту, то слід розглянути структуру функціональних стилів мови, до яких входять розмовний, книжний, публіцистичний, офіційно-діловий, художній та науковий стилі.

Якщо проаналізувати тексти різної стилістичною спрямованості, то можна виявити їх особливості щодо таких параметрів як: мінімальне, середнє та максимальне значення слів речення, букв у слові, кількість груп різних частин речення, співвідношення динамічності та статичності тексту, загальна кількість слів тексту, кількість слів різних частин речення та їх співвідношення. На основі цих параметрів вибудовуються кластери с певними характеристиками, на основі яких текст можна віднести до певного стилю.

Отже, оцінка методів класифікації може бути такою [5]:

- TP (true positive);
- FP (false positive);
- TN (true negative);
- FN (false negative).

Додатково можна оцінити точність

$$Precission = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$$

та повноту системи

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}.$$

Список використаних джерел

1. Cavnar W.B. N-Gram-Based Text Filtering For TREC-2 // Proceedings of the Second Text Retrieval Conference (TREC-2). – NIST. – Gaithersburg. – Maryland. – 1993. – P. 171-180
2. Лингвистический энциклопедический словарь / Гл. ред. В.Н. Ярцева. – М.: Сов. энциклопедия, 1990
3. Ломакина Л.С., Мордвинов А.В., Суркова А.С. Построение и исследование модели текста для его классификации по предметным категориям. // Системы управления и информационные технологии. – 2011. – №1(43) . – С. 16-20.
4. Суркова А. С. Концептуальный анализ, принципы моделирования и оптимизация алгоритмов синтеза текстовых структур : автореферат дис. ... доктора технических наук : 05.13.01 / Суркова Анна Сергеевна; [Место защиты: Нижегор. гос. техн. ун-т им Р.Е. Алексеева]. - Нижний Новгород, 2017. – 39 с.
5. Confusion matrix [Електронний ресурс] – Режим доступу: https://en.wikipedia.org/wiki/Confusion_matrix (дата звернення: 02.04.20). – Назва з екрана.