УДК 004.8

І. О. Миколюк
О. І. Суприган

ДОСЛІДЖЕННЯ ТА ОЦІНКА ОСНОВНИХ МЕТОДІВ ДЛЯ ЗАДАЧІ АНАЛІЗУ АНГЛОМОВНОГО ТЕКСТУ НА НАЯВНІСТЬ СТАЛИХ МОВНИХ КОНСТРУКЦІЙ

Вінницький національний технічний університет;

Анотація

В статті приведено розгляд методів, які використовуються для задач інтелектуального аналізу текстової інформації. Виділено переваги та недоліки таких методів, як метод Байєса, метод k найближчих сусідів, метод дерева рішень. Запропоновано послідовне використання методу підрахунку TF-індексу та методу Байєса.

Ключові слова: аналіз тексту, метод Байєса, метод опорних векторів, метод k найближчих сусідів, метод підрахунку TF-індексу.

Abstract

The article gives an overview of the methods used for the intellectual analysis of text information. The advantages and disadvantages of such methods as the Bayes method, the method of k nearest neighbors, method of reference vectors are singled out. A consistent use of the method has been proposed for evaluation of the TF-index and the Bayes method.

Keywords: text analysis, Naive Bayes, method of reference vectors, k method of nearest neighbors, TF index calculation method.

Вступ

Складність задач семантичного аналізу текстової інформації вважається однією з головних перешкод на шляху побудови штучного інтелекту в цілому та розв'язання з належною якістю значної частини задач комп'ютерної лінгвістики зокрема.

Аналіз текстів на природній мові був актуальним практично з моменту їх появи. При такому аналізі необхідно визначити правила, за допомогою яких, на думку фахівців, «формальна система перетвориться в систему змістовну». Аналіз тексту використовувався і продовжує використовуватися для класифікації текстів за їх словами і словосполученнями, анотування та реферування текстів, проведення семантично орієнтованого пошуку текстів по заданих концептам, визначення авторського права претендента на відповідний текст та ін. [1].

Результати дослідження

Розрізняють кількісні і статичні методи для задач аналізу текстової інформації. Кількісні зводяться до підрахунку частоти вживання мовних одиниць. Статистичні методи передбачають використання різних формул для виявлення правил розподілу мовних одиниць у мовленні, становлення зв'язків між мовними елементами [2].

Основними методами для інтелектуального аналізу англомовного тексту є:

- метод Байєса;
- метод k найближчих сусідів;
- метод опорних векторів;
- метод лінійної регресії.

Переваги кожного підходу залежать від типів та обсягу аналізованих текстів та тих питань, які аналітик повинен вирішувати. Навіть при наявності єдиного підходу можливі варіації щодо його застосування. Немає єдиної методики, яка найбільш підходить для всіх видів текстових аналізів. Однак для підрахунку частотності обрано модель, що базується на показнику TF-IDF. Так як метою

дослідження визначено підвищення швидкості аналізу тексту, то для класифікації метод Байєса можна вважати оптимальним. Адже для методу k найближчих сусідів характерний недолік – велика тривалість роботи через необхідність повного перебору навчальної вибірки; а для методу опорних векторів та методу логістичної регресії притаманний недолік, такий як складна інтерпретуємість параметрів алгоритму і нестійкість по відношенню до викидів у вихідних даних.

Для пошуку слів, які можна рекомендувати проектувальнику в якості сутностей аналізу, можна використовувати статистичний підхід (метод підрахунку TF-індексу) в поєднанні з методами синтаксичного аналізу текстів для моделювання зв'язків між сутностями та атрибутами [3].

Показник TF (англ. term frequency – частота слова) – статистична міра, яка використовується для оцінки важливості слова в контексті документа Doc_k (W,W_d). Вона визначається як відношення ня n_i деякого слова W_i до загальної кількості слів документа.

$$TF(W_i, Doc_k) = \frac{n_i}{\sum_k n_k}$$

Атрибутами БД виступають слова та словосполучення, що не повинні мати високе значення TF. Для оцінки важливості сутності також можна використовувати показник TF-IDF (від англ. TF – term frequency, IDF – inverse document frequency) – статистична міра, яка використовується для оцінки важливості слова в контексті документа, що є частиною колекції документів або корпусу.

Після проведення обрахунку частоти зустрічаємості сутностей необхідно віднести текст до певної категорії. Для цього доцільно розробити алгоритм класифікації на основі методу Байєса [4].

Далі всі ймовірності підраховуються за методом максимальної правдоподібності.

Висновки

В даній статті проведений аналіз існуючих методів та моделей, що застосовуються для вирішення задачі інтелектуального аналізу англомовного тексту, а також висвітлено недосконаліть існуючих методів, що обумовлює необхідність розроблення моделі на основі комбінації наявних методик та алгоритмів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Большакова Е. И. Автоматическая обработка текстов на естественном языке и компьютерная / Е. И. Большакова, Е. С. Клишинский. – М. : МИЭМ, 2011. – 272 с.

2. Шалак В. И. Современный контент-анализ. Приложения в области: политологии, психологии, социологии, культурологии, экономики, рекламы. — М.: Омега - Л, 2009. — 272 с.

3. Кравець Р.Б. Застосування багатозначної логіки для інтелектуального аналізу даних [Текст] / Кравець Р.Б., Шаховська Н.Б. // Вісник Національного університету «Львівська політехніка». – Л.: Вид-во Національного ун-ту «Львівська політехніка», 2002. – №468: Комп'ютерна інженерія та інформаційні технології. – С. 58–65.

4. Наївний баєсів класифікатор [Електронний ресурс] – режим доступа: https://uk.wikipedia.org/wiki/ Наївний_баєсів_класифікатор

Миколюк Ірина Олександрівна *– студентка групи 1КН-18м, факультет інформаційних технологій і комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: 2kn14b.mykoliuk@gmail.com;*

Суприган Олена Іванівна *– к. т. н., доцент кафедри комп'ютерних наук ВНТУ, Вінницький національний технічний університет, м. Вінниця.*

Iryna O. Mykoliuk *– Student, ComputerScienceDepartment, VinnytsiaNationalTechnicalUniversity, Vinnytsia, email: 2kn14b.mykoliuk@gmail.com;*

Olena I. Suprygan *– Associate Professor of the Computer Sciences Chair, Vinnytsia National Technical University, Vinnytsia.*