УДК 004.932

О.В. Бісікало

В.В. Дадиверін

ДОПОВНЕННЯ ТА ПОГЛИБЛЕННЯ МЕТОДУ ПОШУКУ КЛЮЧОВИХ СЛІВ

Вінницький національний технічний університет

Анотація

Дослідження присвячено тематиці пошуку ключових слів у текстовій інформації. У роботі розкривається ідея застосування інструментів, що можуть покращити якість вихідного результату існуючого гібридного методу пошуку ключових слів. Запропонована модифікація дає можливість для додаткової оцінки влучності відфільтрованої інформації та дозволяє врахувати часові або частотні характеристики підібраних слів.

Ключові слова: *ключові слова, словник синонімів, частотність, розподіл, зв'язки, слово.*

Abstract

The research is devoted to the topic of keyword search in text information. The paper reveals the idea of using tools that can improve the quality of the output of the existing hybrid keyword search method. The proposed modification allows for additional assessment of the accuracy of the filtered information and allows taking into account the time or frequency characteristics of the selected words.

Keywords: *keywords, thesaurus, frequency, distribution, connections, word.*

Вступ

Кожного дня світ продукує неймовірну кількість текстової інформації різного роду направлення, що в свою чергу підвищує важливість аналізу та категоризації цих вхідних даних. Задача пошуку за ключовими словами є максимально актуальною як для систем загального пошуку інформації (наприклад, Google Search), так і для застосунків для оптимізації роботи спеціальних систем (наприклад, внутрішня документація для агрегатів автомобілів Subaru за темами).

Використання ключових слів дозволяє автоматизувати та пришвидшити роботу електронного приладу з інформацією, адже ключові слова слугують стислим та якісним переказом, який дає можливість не витрачати зайвий час на повторну обробку того самого тексту.

Результати дослідження

За основу дослідження взято гібридний метод пошуку ключових слів на основі парсингу англомовних текстів, що був представлений О.В. Яхимовичем у свої дисертаційній роботі [1]. Цей алгоритм складається зі статистичних та словникових методів і заявляє приріст до 14.3% за показником абсолютної точності в порівнянні з подібними існуючими доробками.

Оригінальний алгоритм виглядає наступним чином:

1. Синтаксичний аналіз тексту і отримання даних про зв'язки між парами слів і частини мови, до яких належать слова тексту [1].
2. Отримання з тексту набору всіх виразів з типами зв'язків flat та compound [1].
3. Фільтрування пар слів, зв'язки між якими належать до переліку неінформативних [1].
4. Заміна займенників у парах слів відповідними іменниками [1].

5. Відсіювання слів, які під час синтаксичного аналізу було віднесено до неінформативних частин мови [1].
6. Фільтрування стоп-слів [1].
7. Визначення кількості зв'язків для кожного слова з пари [1].
8. Прийняття перших n слів з найбільшою кількістю зв'язків як ключові (де n - бажана кількість шуканих ключових слів) [1].

Розглянемо можливі рішення для доповнення базового алгоритму:

1. Використання спеціальних словників синонімів.

Використання такого роду словників дозволяє уникати повторення за сенсом слів та допомагає у правильному виборі ключових слів, коли їхня кількість серйозно обмеження. Наприклад, якщо в тексті зустрічаються слова "медик" та "лікар", то вони будуть об'єднані в одне слово "лікар".

Бібліотеки, що підтримують специфічні словники: NLTK [2], SpaCy [3].

Рідше такого роду словники використовують для визначення правильного значення слова. Зазвичай для такого аналізу використовують векторні моделі, за типом GloVe [4].

2. Врахування розподілу та частотності слів.

Потрібно відсіювати слова, що зустрічаються у тексті з частотою раз на п'ять слів або їхня кількість не переважає мінімального кількісного значення (наприклад, слово "лікар" зустрічається менше 3 разів).

3. Застосування часових характеристик.

Потрібно підвищувати вагу слів, які рівномірно розподілені по тексту або з'являються у критичних моментах тексту (наприклад, вступ та висновки) У випадку рівномірного розподілення текст розділяється на частини (найчастіше це абзаци) і обчислюється залученість вибраного слова у всіх цих розділах. У випадку визначення вагу слів за їх знаходженням у критичних моментах треба застосовувати наступну формулу [5]:

$$\text{Weight(time)} = \text{TF-IDF(time)} \times \log(1 + \text{section position}).$$

Інтеграція рішень в існуючий алгоритм:

1. Нормалізація за синонімічним словником варто додати після кроку під номером 3 (видалення неінформативних слів).
2. Врахування частотності та розподілу варто додати після кроку під номером 7 (розбиття слів на пари та підрахунок зв'язків).
3. Врахування часової динаміки варто додати до кроку під номером 8 (вибір n ключових слів).

Висновки

Проведено огляд існуючих підходів та аналіз методу пошуку ключових слів в текстових даних, який взятий за прототип. У результати знайдено рішення, які можуть покращити точність його роботи. Запропоновано модифікації, такі як нормалізація за синонімічним словником, аналіз частотності та розподілу, адаптація TF-IDF з часовою динамікою. Також запропоновано інструкцію до інтеграції модифікацій в існуючий алгоритм – цей підхід відкриває нові можливості для автоматизації текстових досліджень, покращення якості інформаційного пошуку та побудови інтелектуальних систем.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Яхимович О. В. Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів [Текст] : автореф. дис. ... канд. техн. наук : 05.13.06 / Олександр Вікторович Яхимович ; Вінницький національний технічний університет. – Вінниця, 2021. – 25 с. – Бібліогр. : с. 18-20 (21 назва).

2. Національний корпус природної мови. Робота з WordNet [Електронний ресурс] // NLTK Documentation. – Режим доступу: https://www.nltk.org/howto/wordnet.html, вільний.
3. spaCy: Usage Documentation [Електронний ресурс] // spaCy. – Режим доступу: https://spacy.io/usage, вільний.
4. GloVe: Global Vectors for Word Representation [Електронний ресурс] // Stanford NLP Group. – Режим доступу: https://nlp.stanford.edu/projects/glove/, вільний.
5. Manning, C., Raghavan, P., & Schütze, H. Introduction to Information Retrieval. Розділ 6: Scoring, Term Weighting [Електронний ресурс]. – 2008. – Режим доступу: https://nlp.stanford.edu/IR-book/information-retrieval-book.html, вільний.

Бісікало Олег Володимирович – д-р техн. наук, професор, завідувач кафедри АІІТ, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@vntu.edu.ua

Дадиверін Віталій Валерійович – аспірант групи 126-24а, факультет автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: vetaldadyverin@gmail.com

Bisikalo Oleg V. – Dr.Sc. (Eng.), Professor of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia, e-mail: obisikalo@vntu.edu.ua

Dadyverin Vitalii V. – Department of Automation and intelligent information technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: vetaldadyverin@gmail.com