УДК 004.93’1

# МЕТОДИ ПОШУКУ КЛЮЧОВИХ СЛІВ

Лебідь В.М.
Науковий керівник – к.т.н., доц., доц. каф. КІТС Сердюк Н. М.
Харківський національний університет радіоелектроніки, каф. КІТС,
м. Харків, Україна
тел. +38(095) 515-37-39, e-mail: vadym.lebid@nure.ua.

This work is devoted to methods for extracting keywords. Often in scientific papers and publications there is a large amount of text that is very difficult to read quickly and at the same time not to lose important information. In this case, it becomes necessary to reduce the volume of the document by highlighting the most significant parts of the text, called the abstract. Compiling keywords is a complex and time-consuming job. This task requires additional staff and therefore it is more expedient to use systems for automatically extracting keywords from text.

Ключові слова – це ідеї та теми, які визначають зміст контенту. З погляду SEO, це слова та фрази, які користувачі вводять у пошукові системи, також звані пошуковими запитами. Якщо звести все на своїй сторінці – всі зображення, відео, текст тощо – до простих слів та фраз, це будуть основні ключові слова. За допомогою ключових фраз кінцевий користувач може отримати максимально повну інформацію про основні думки або теми конкретного тексту. Це завдання постійно повинні вирішувати пошукові системи, агрегатори новин і ресурсів, які аналізують думку користувачів соціальних мереж [1].

Основні види методів і моделей автоматичного вилучення ключових слів можна розділити на лінгвістичні, статистичні, спектральні, та гібридні.

Етапи вилучення ключових слів з тексту практично ідентична всім використовуваним методам з чотирьох етапів:

1. Етап попередньої обробки текстів. Видалення слів із тексту з метою його спрощення, виключення елементів маркування, приведення слів до словникової форми та видалення стоп-слів, що не несуть смислового навантаження.

2. На основі цих даних складається список кандидатів в ключові слова, за якими проводитиметься пошук.

3. Виявлення найважливіших ознак кожного з кандидата.

4. Відбір ключових слів з-поміж кандидатів для подальшого використання.

Лінгвістичні методи базуються на значеннях слів, використовують онтології та семантичні дані про слово. З цим методом пов'язана низка проблем. По-перше, це трудомісткість виконання операцій аналізу текстів, що виконуються вручну. Крім того, цей метод вимагає від дослідника

великої кількості вільного часу для проведення досліджень, а також може виникнути значна кількість помилок та неточностей.

На основі чисельних даних про зустрічальність слова будується статистичний метод дослідження тексту. Головною перевагою статистичних методів є універсальність алгоритмів вилучення ключових фраз, відсутність необхідності у трудомістких процедурах побудови лінгвістичних баз знань, простота реалізації [2].

При використанні гібридних методик статистичні методи обробки доповнюються однією-двома лінгвістичними процедурами та лінгвістичною базою знань різної глибини. Цей підхід є комбінованим, оскільки він поєднує лінгвістичний та статичний методи.

Для того, щоб отримати спектральний аналіз, використовується перетворення Фур'є та карти усереднення – подібно до частотно-часового представлення сигналу від джерела, отриманого за допомогою вейвлет-перетворення. А це означає, що цей метод може працювати зі спектрограмою тексту, що містить новий якісний вигляд вмісту текстового файлу. У цьому методі відсутній ряд недоліків, властивих іншим методам. Наприклад нема потреби проводити складні процедури лексичного та морфолого-лінгвістичного дослідження, а також можна визначити місце ключового слова в тексті [3].

Отже, розглянуто чотири метода які мають як переваги, так і недоліки. Ключові слова є одним із найважливіших етапів у процесі класифікації та кластеризації документів, а також виявлення загальної теми документа. В останні роки розроблено різні підходи автоматизованого вилучення ключових слів та фраз, в основі яких лежать лінгвістичні чи статистичні методи. У їхньому аналізі зазначається, що створення ефективних екстракторів стане й надалі одним із ключових трендів комп'ютерної обробки текстів.

Список використаних джерел:

1. Лебідь В.М. Основні проблеми пошуку необхідного контенту при виборі сайтів (2021, 5 листопада) https://ojs.ukrlogos.in.ua/index.php/liga/issue/view/inter-05.11.2021/629

2. Мазурець О. В. Інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів / О. В. Мазурець // Вісник Хмельницького національного університету. Серія: Технічні науки. – 2018. – № 3. – С. 223–230.

3. Современные методы автоматизированного извлечения ключевых слов из текста (б.д.). Взято 23 лютого 2022 з https://www.academia.edu/30459750