Інформаційна технологія рекурсивного семантичного аналізу текстів шляхом дисперсійного оцінювання слів

Сергієва О.О., Мазурець О.В.
Хмельницький національний університет

Семантичний аналіз тексту є етапом у послідовності дій алгоритмів автоматичного розуміння текстів, що полягають у виділенні семантичних відношень, формуванні семантичного подання текстів. Один з можливих варіантів відображення семантичного подання – структура, що складається з текстових елементів. Глибина семантичного аналізу може бути різною, а в існуючих системах найчастіше будується тільки синтаксико-семантичне подання тексту або окремих фрагментів, до яких відносять анотації, реферати та переліки ключових слів [1].

Перелік ключових слів тексту є найбільш семантично стиснутим результатом семантичного аналізу тексту, й пошук ефективних методів автоматизації його формування відкриє надає можливість вирішення багатьох похідних задач.

Метою роботи є розробка інформаційної технології семантичного аналізу текстів шляхом дисперсійного оцінювання слів.

Схему відповідної інформаційної технології подано на рисунку 1. В якості метода семантичного аналізу текстів використовується метод рекурсивного дисперсійного оцінювання [2] для пошуку ключових термінів у текстовому контенті цифрового файлу .docx.

Рисунок 1 – Схема інформаційної технології рекурсивного семантичного аналізу текстів шляхом дисперсійного оцінювання слів

Наведені на рисунку 1 основні етапи автоматизованого аналізу текстового контенту цифрових документів передбачають на першому етапі одержання системою вхідних даних, до яких належать власне цифровий документ, що аналізується, та необов'язкові параметри роботи, якщо існує необхідність їх перевизначення. Після чого проводиться рекурсивний пошук множин ключових термінів (рис. 2), по одній ітерації для кожного варіанту розмірності термінів (1-*n* слів у терміні). На завершальному етапі роботи

системи виконується об’єднання результатів окремих ітерацій пошуку множин ключових термінів, за результатами чого впорядкована й обмежена результуюча множина ключових термінів виводиться в якості вихідних даних.

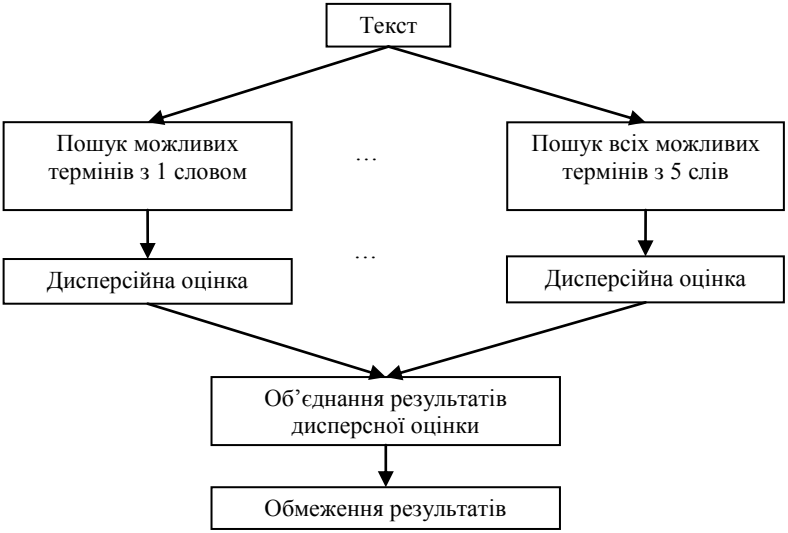


Рисунок 2 – Схема рекурсивної функції методу дисперсійного оцінювання слів в інформаційній технології

Вхідними даними інформаційної технології рекурсивного аналізу текстів, є цифровий документ (файл з розширенням .docx), контент якого містить текст для аналізу; та набір необов’язкових параметрів роботи системи: максимальна кількість слів у терміні n та параметр щільності ключових термінів у тексті P. По замовчуванню використовуються параметри n=5 та P=15. Вихідними даними інформаційної технології є множина ключових термінів тексту.

Дисперсійний аналіз, що використовується для пошуку ключових термінів, є статистичним методом оцінки зв'язку між факторними й результативними ознаками в різних групах, відібраний випадковим чином, заснований на визначенні розходжень значень ознак. В основі дисперсійного аналізу лежить аналіз відхилень всіх одиниць досліджуваної сукупності від середнього арифметичного. Як міра відхилень береться дисперсія – середній квадрат відхилень. Відхилення, викликані впливом факторної ознаки порівнюються з величиною відхилень, викликаних випадковими обставинами. Якщо відхилення, викликані факторною ознакою, більш істотні, ніж випадкові відхилення, то вважається, що фактор впливає на результуючу ознаку. Таким чином, дисперсійна оцінка є оцінкою

дискримінантної сили слів й дозволяє відділити із загальної множини широковживаних у тексті слів слова, що розташовані рівномірно.

На основі запропонованого підходу було створено тестову автоматизовану систему рекурсивного семантичного аналізу текстів шляхом дисперсійного оцінювання слів, яка продемонструвала достатньо високу ефективність, особливо для великих текстів. Розроблена система дозволяє за вхідними даними у вигляді завантаженого файлу цифрового документу з розширенням .docx, контент якого містить текст для аналізу (рис. 3) та наборк параметрів роботи системи одержати множину ключових термінів тексту (рис. 4).

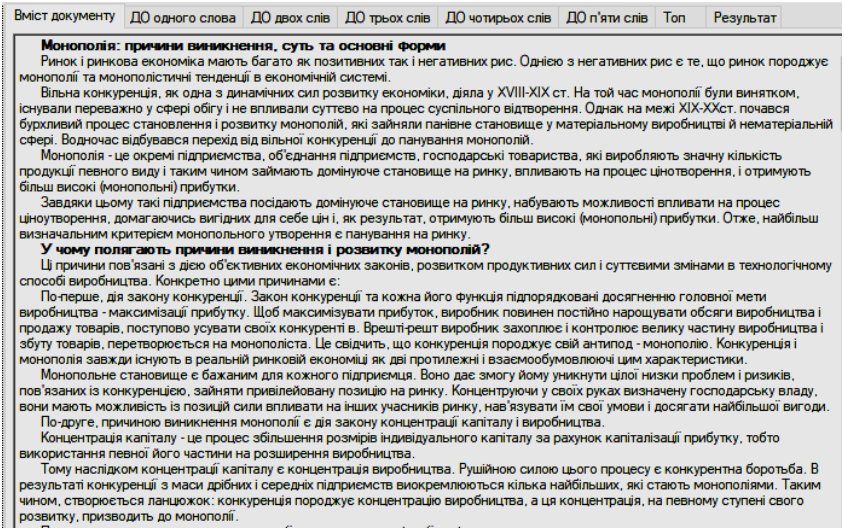
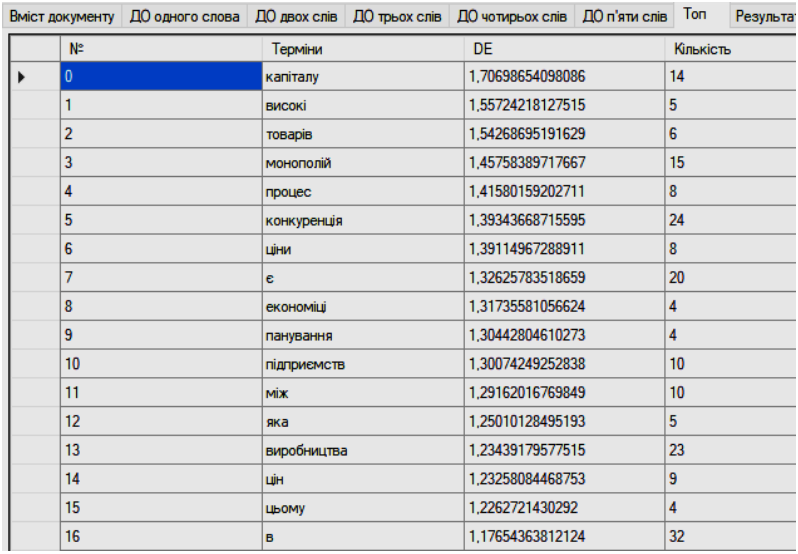


Рисунок 3 – Завантажений у систему файл цифрового документу для аналізу

Розроблена система рекурсивного семантичного аналізу текстів виконує такі функції:

– зчитування .docx файлу та витяг його текстового контенту;
– побудова дерева структури документу;
– вибір елемента дерева структури документу для аналізу;
– формування множини термінів тексту;
– оцінка важливості термінів тексту методом дисперсійного оцінювання;
– сортування й обмеження результуючої множини термінів;
– збереження результатів пошуку в базу даних та перегляд збережених результатів.



Вміст документу | ДО одного слова | ДО двох слів | ДО трьох слів | ДО чотирьох слів | ДО п'яти слів | Топ | Результат

№	Терміни	DE	Кількість
0	капіталу	1,70698654098086	14
1	високі	1,55724218127515	5
2	товарів	1,54268695191629	6
3	монополій	1,45758389717667	15
4	процес	1,41580159202711	8
5	конкуренція	1,39343668715595	24
6	ціни	1,39114967288911	8
7	є	1,32625783518659	20
8	економіці	1,31735581056624	4
9	панування	1,30442804610273	4
10	підприємств	1,30074249252838	10
11	між	1,29162016769849	10
12	яка	1,25010128495193	5
13	виробництва	1,23439179577515	23
14	цін	1,23258084468753	9
15	цьому	1,2262721430292	4
16	в	1,17654363812124	32

Рисунок 4 – Результат визначення множини ключових термінів тексту

Отже, було запропоновано інформаційну технологію семантичного аналізу текстів шляхом дисперсійного оцінювання слів. На основі запропонованої інформаційної технології було створено тестову автоматизовану систему рекурсивного семантичного аналізу текстів шляхом дисперсійного оцінювання слів, яка продемонструвала достатньо високу ефективність, особливо для великих текстів. Подальші дослідження спрямовані на поширення можливостей технології на ефективну роботу із складними реченнями та семантичними конструкціями.

Перевагою даної інформаційної технології рекурсивного семантичного аналізу текстів шляхом дисперсійного оцінювання слів є те, що вона дозволяє проводити пошук ключових термінів без використання спеціальних цифрових словників чи баз даних корпусу слів відповідної мови.

Література

1. Сергієва О. О., Мазурець О. В. Інтелектуальна система автоматизованого стиснення текстів / О. О. Сергієва, О. В. Мазурець // Матеріали VI Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ICST-ODESSA-2017». Одеса – 2017. – С.283-285.

2. Бармак О. В., Мазурець О. В. Методи автоматизації визначення семантичних термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Вісник Хмельницького національного університету. Сер.: Технічні науки. Хмельницький. – 2015, №2(223). – С.209-213.