DOI 10.31891/2307-5732-2019-273-3-188-196
УДК 004.9

О.В. МАЗУРЕЦЬ, О.О. КОВАЛЬ
Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ РЕКУРСИВНОГО ПОШУКУ КЛЮЧОВИХ ТЕРМІНІВ У ЦИФРОВИХ ТЕКСТАХ

В статті розглянуто інформаційну технологію рекурсивного пошуку ключових термінів у цифрових текстах, яка проводить аналіз текстового контенту із використанням методу дисперсійної оцінки та без використання лексичних баз даних корпусів слів. Характерною рисою запропонованої інформаційної технології є використання рекурсивних складових при пошуку ключових термінів. Процес автоматизованого аналізу цифрового тексту шляхом рекурсивного пошуку ключових термінів із використанням методу дисперсійного оцінювання складається з ряду етапів перетворення інформації, які у сукупності формують інформаційну технологію рекурсивного пошуку ключових термінів. Розроблена інформаційна технологія рекурсивного пошуку ключових термінів була реалізована в тестовому програмному продукті. Вхідними даними для системи є електронний документ із цифровим текстом, а вихідними даними є множина ключових термінів, що відповідна досліджуваному фрагменту текстового контенту електронного документу. За допомогою розробленого тестового програмного забезпечення були проведені дослідження, що підтвердили можливість ефективно автоматизовано формувати множини ключових семантичних термінів текстів із показниками точності пошуку до 89,6% й повноти пошуку до 93,3%.

Результати порівняння ефективності інформаційної технології рекурсивного пошуку ключових термінів у цифрових текстах із аналогічними результатами для технологій, що використовують лексичні бази даних корпусів слів для ідентифікації слів у текстах, є неоднозначними. У 42,3% випадках використання рекурсивного пошуку негативно вплинуло на якість результату, проте в 18,6% випадків такий підхід виявив кращий результат. Перевагами розробленої інформаційної технології рекурсивного пошуку ключових термінів у цифрових текстах, яка проводить аналіз текстового контенту із використанням методу дисперсійної оцінки, є відсутність необхідності використання лексичних баз даних корпусів слів, суттєве прискорення швидкодії, можливість використання для текстів різними мовами, можливість використання для текстів із кількома мовами, кращі результати під час обробки вузькоспеціалізованого контенту. Дана інформаційна технологія може бути ефективно використана для аналізу текстів із невідомими властивостями тематики та мови.

Ключові слова: цифровий документ, ключові терміни, дисперсійна оцінка.

O. MAZURETS, O. KOVAL
Khmelnytskyi National University

INFORMATION TECHNOLOGY FOR RECURSIONAL DEFINITION OF KEY TERMS IN DIGITAL TEXTS

In the article the information technology for recursional definition of semantic key terms in digital texts is considered, which conducts the analysis of text content using the method of dispersion evaluation and without the use of lexical databases of word cases. A characteristic feature of the proposed information technology is the use of recursive components in the search for key terms. The process of automated analysis of digital text through recursional search of key terms using the dispersion evaluation method consists of series of stages of the transformation of information, which collectively form the information technology for recursional definition of semantic key terms. The information technology for recursional definition of semantic key terms has been introduced in the test software product. The input data for the system is an electronic document with digital text, and the output data is a set of key terms that correspond to the investigated fragment of the text content of the electronic document. With the help of developed test software, studies were conducted that confirmed the ability to effectively formulate a set of key semantic terms of texts with search precision up to 89.6% and search recall up to 93.3%. The results of the comparison of the effectiveness of information technology for recursional definition of semantic key terms in digital texts with similar results for technology that use lexical databases of word cases to identify words in texts are ambiguous. In 42.3% of cases, the use of recursional definition negatively affected the quality of the result, but in 18.6% of cases, this approach has shown better result. The advantages of the developed information technology for recursional definition of semantic key terms in digital texts, which conducts analysis of text content using the dispersion evaluation method, are the absence of the need to use lexical database of word cases, significant acceleration of speed, the possibility of using for texts in different languages, the possibility of using for texts in several languages, better results in handling highly specialized content. This information technology can be effectively used to analyse texts with unknown properties of the subject and language.

Keywords: digital document, key terms, disperse evaluation.

Постановка проблеми в загальному вигляді

З інформаційним розвитком суспільства зростає кількість інформації, яку людині потрібно опрацьовувати. Найбільш поширеним є текстовий формат передачі інформації, оскільки він є найбільш звичним для людини. Але при великих обсягах тексту втрачається якість його сприйняття. Крім того, часто можна зустріти текст, який не відповідає тому, за що його видають. Тому зростає потреба в попередньому аналізі такого тексту. Такий аналіз полягає в швидкому отримані сенсу тексту. Для людини дана операція виснажлива і може відбуватися доволі довго в залежності від обсягу текстового контенту.

Семантичний аналіз текстів є складним математичним завданням, рішення якого застосовується у процесі створення штучного інтелекту, при цьому воно ускладнюється необхідністю обробки природної мови. Дані якісного семантичного аналізу також можуть використовуватися в торгівлі для аналізу попиту на товари за отриманими відкликаннями, у системах автоматичного перекладу, пошукових системах тощо. Поняття ключових термінів як носіїв найбільш семантично вагомої інформації про текст активно

використовується в інформатиці, зокрема, завданнях інформаційного пошуку. Для інформаційних технологій даний напрямок не є новим і знайшов своє відображення переважно в SEO-системах. Існує багато програмних продуктів і теоретичних пропозицій, як отримати сенс тексту, але не існує єдиного рішення, яке б остаточно вирішувало дану проблему.

Множина ключових термінів тексту є найбільш семантично стиснутим результатом семантичного аналізу тексту [1], й пошук ефективних методів автоматизованого формування таких множин відкриє можливість розв'язання багатьох похідних задач.

Аналіз останніх досліджень

Семантичний аналіз тексту є етапом у послідовності дій алгоритмів автоматичного розуміння текстів, що полягають у виділенні семантично важливих конструкцій, семантичних відношень, формуванні семантичного подання текстів. Один з можливих варіантів відображення семантичного подання – це структура, що складається із текстових елементів. Глибина семантичного аналізу може бути різною [2], однак в існуючих системах найчастіше будується тільки згорнуте синтаксико-семантичне подання тексту чи його окремих фрагментів, до яких відносять анотації, реферати та переліки ключових термінів [3].

Проблему автоматизованої аналітичної обробки текстової інформації намагаються вирішити багато вітчизняних та іноземних авторів, серед яких можна відзначити роботи Д.В. Ланде, В.Е. Снитюка, В.І. Горькової, Х.П. Луна, В.Є. Берзона, І.П. Севбо, Є.А. Борохова, В.П. Леонова, С.І. Гінді та інших. Дослідження в напрямку автоматизації обробки текстів у Європі та США привертають увагу відомих приватних фірм і державних установ найвищого рівня. Європейський Союз наразі координує ряд програм у галузі автоматичної обробки тексту, зокрема Human Language Technology Sector of the Information Society Technologies (IST) Programme. Основні вишукування присвячені автоматизації процесу синтаксичного аналізу цифрових текстів.

Більшість існуючих програмних систем, створених для автоматизованої аналітичної обробки цифрових текстів, призначені переважно для SEO-аналізу текстів. Наприклад, аналізатор від біржі контенту «Адвего» [4] вираховує кількість слів, кількість граматичних помилок і т.п. Сервіс також дозволяє побачити перелік ключових слів, які вираховуються за методом частотної оцінки, та забезпечує аналіз текстів на плагіат. Багатофункціональна SEO-платформа «Serpstat» [5] забезпечує глибинний аналіз текстового контенту, аналіз пошукових питань, розподіл запитів по дереву сайту, пошук схожих фраз тощо; аналіз ключових слів, зокрема, допомагає використовувати найефективніші ключові слова в контенті й рекламних оголошеннях для розширення присутності у комерційній ніші. Текстовий аналізатор від компанії «Seozor» [6] визначає вагу слів в тексті для складання анкор-листа.

Для автоматизації пошуку ключових слів використовуються різноманітні методи аналізу текстів, таких як частотна оцінка TF, оцінка TFIDF та дисперсійна оцінка DE [3]. Ці методи дозволяють співставити окремим словам або словосполученням тексту деякі певним чином поставлені у відповідність числові вагові значення, що вказують на міру їх важливості в досліджуваному тексті [7]. Попередніми дослідженнями було визначено найбільш ефективним методом аналізу текстів метод дисперсійної оцінки, проте встановлено й фактори, які ускладнюють його застосування для вирішення задачі автоматизованого визначення семантичних термінів в навчальних матеріалах [8]. Зокрема, малий обсяг контенту й вузька семантична направленість елементів аналізу зменшують ефективність наведених методів аналізу текстів.

Суттєвою вадою існуючих методів та систем є необхідність використання баз даних корпусів слів відповідних мов для ідентифікації слів у текстах, що обмежує можливості таких систем та знижує гнучкість аналізу вузькоспеціалізованого контенту. Тому є доцільною розробка нової інформаційної технології, яка із використанням методу дисперсійної оцінки дозволить ефективно й автоматизовано визначати семантичні терміни в цифрових текстах.

Постановка задачі

Метою роботи є розробка інформаційної технології рекурсивного пошуку ключових термінів у цифрових текстах, яка проводить аналіз текстового контенту із використанням методу дисперсійної оцінки та без використання лексичних баз даних корпусів слів.

Викладення основних матеріалів дослідження

Характерною рисою запропонованої інформаційної технології є визначене використання рекурсивних складових при пошуку ключових термінів. Процес автоматизованого аналізу цифрового тексту шляхом рекурсивного пошуку ключових термінів із використанням методу дисперсійного оцінювання складається з ряду етапів перетворення інформації, які у сукупності формують інформаційну технологію рекурсивного пошуку ключових термінів, загальна схема якої показана на рис. 1.

Вхідними даними для обробки, відповідно до схеми інформаційної технології рекурсивного пошуку ключових термінів, є цифровий текст (файл документу з розширенням .docx) для аналізу та набір параметрів пошуку, до яких належать максимальна кількість слів у терміні n та гранична щільність ключових термінів у тексті P. По замовчуванню використовуються параметри $n = 5$ та $P = 15$ [9].

На *підготовчому етапі* (Блок 1) визначаються межі фраз, що обмежують локальні області пошуку окремих термінів. *Формування впорядкованої множини слів тексту* (Блок 1.1) полягає в аналізі структури цифрового документу. Шляхом витягу текстового контенту з цифрового документу формується впорядкована множина M_{TXT}, яка складається зі слів тексту в порядку їх слідування у документі. Кожному слову в множині привласнюється порядковий номер, за допомогою якого в подальшому відбувається пошук

відстаней між однаковими термінами. Оскільки термін може складатися з кількох слів, то в обрахунках його порядковим номером (позицією) приймається номер першого слова.

Рис. 1. Загальна схема інформаційної технології рекурсивного пошуку ключових термінів у цифрових текстах

Блок 1.2 (*Пошук меж текстових контейнерів)* призначений для розбиття текстового контенту електронного документу, що обробляється, на менші фрагменти – контейнери (фрази). Під фразою розуміється семантично цілісний вузол, який виокремлений стилістичним форматуванням тексту чи розділовими знаками, й локалізує місцезнаходження окремих термінів. Відтак, межами контейнеру визначаються:

- зміна стилістичних властивостей тексту;
- розділові знаки;
- абзаци/параграфи.

Результати проведеного аналізу властивостей ключових термінів [8] свідчать, що терміни не можуть виходити за межі таких контейнерів.

В процесі обробки власне розділові знаки (наприклад: !№;%:?*()_+=-,.@#$^&*<>'""|\/—…{}–»«•§) видаляються з контенту елементів множини слів тексту M_{TXT}, за винятком випадків коли вони є частиною слів (зокрема, апострофи та дефіси).

Визначення приналежності слів до контейнерів (Блок 1.3) полягає в привласненні кожному елементу множини слів тексту M_{TXT} окрім порядкового номеру слідування у документі ще й номеру контейнера, до якого воно віднесене. При подальшому аналізі до одного терміну не будуть входити елементи множини слів тексту з різними номерами контейнерів.

Рекурсивний етап (Блок 2), що забезпечує пошук множин термінів розмірності n, включає наступні етапи інформаційної технології рекурсивного пошуку ключових термінів, які реалізують рекурсивний пошук множин ключових термінів; при цьому проводиться по одній ітерації для кожного варіанту розмірності термінів. На початковій ітерації розмірність терміну x (кількість слів у терміні) приймається $x = 1$ й збільшується на одиницю кожну ітерацію до досягнення граничного значення n.

Формування початкової множини термінів в межах контейнерів (Блок 2.1) полягає у формуванні множини всіх можливих термінів розмірності x, які присутні в досліджуваному контенті. До множини можливих термінів M_T включаються всі знайдені неперервні впорядковані послідовності слів, які не

виходять за межі контейнерів.

Компактифікація множини варіантів термінів (Блок 2.2) виключає повтори термінів і дозволяє на основі множини можливих термінів M_{T0} сформувати множину оригінальних термінів M_{T1}, а також співставити кожному з них кількість поява у досліджуваному тексті.

Обрахунок відстаней між термінами в тексті (Блок 2.3) є підготовчим етапом до дисперсійного оцінювання термінів, за якого визначаються для кожного терміну з множини M_{T1} (з кількістю поява у тексті більше одного) всі відстані між сусідніми їх появами. При цьому за відстань береться різниця між меншим порядковим номером наступного терміну й більшим порядковим номером попереднього терміну у множині M_{TXT}.

Визначення дисперсії для кожного терміну (Блок 2.4) полягає у визначенні дисперсійної оцінки для кожного з елементів множини оригінальних термінів M_{T1} за відстанями між термінами у тексті.

Дисперсійний аналіз є статистичним методом оцінки зв'язку між факторними й результативними ознаками в різних групах, відібраний випадковим чином, заснований на визначенні розходжень (розкиду) 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 (фактору) 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. В даній технології дисперсійна оцінка є оцінкою дискримінантної сили термінів й дозволяє відділити із загальної множини широковживаних у тексті термінів терміни, що розташовані рівномірно. Якщо деякий термін T в тексті, що складається з N слів, позначений як T_k^n **A**, де індекс k – номер появи даного терміну в тесті, а n – позиція даного слова в тексті, то інтервалом між послідовними появами терміну при таких позначеннях буде величина $\Delta T_k^m = T_{k+1}^m - T_k^n = m - n$ **ΔA** = **A** - **A** = **m-n**, де на m-й і n-й позиціях в досліджуваному тексті знаходиться термін T, який зустрівся k+1-й і k-й рази. Тоді дисперсійна оцінка [9] розраховується за формулою $\sigma = \sqrt{(\Delta T^2) - (\Delta T)^2} \Big/ (\Delta T)$, де (ΔT) **(ΔA)** – середнє значення послідовності $\Delta T_1, \Delta T_2, \Delta T_k$ **ΔA, ΔA, ΔA**; (ΔT^2) **(ΔA²)** – послідовності T_1^2, T_2^2, T_k^2 **A², A², A²**; K – кількість поява терміну T в тексті.

Рис. 2. Рекурсивна складова інформаційної технології

Сортування й обмеження множини термінів за дисперсією (Блок 2.5) визначає остаточний результат дисперсійного оцінювання елементів множини оригінальних термінів M_{T1}. Спершу проводиться сортування елементів множини оригінальних термінів M_{T1} за зменшенням їх дисперсійної оцінки, після чого проводиться видалення всіх елементів, дисперсійна оцінка яких рівна 0. Таке значення дисперсії вказує на те, що даний термін присутній у тексті лише один раз, й не може бути інтерпретований як ключовий. Результуюча множина термінів M_{T2} є результатом пошуку ключових термінів на поточній ітерації й формує локальні вихідні дані для фіксації результатів ітерації (Блок 2.6).

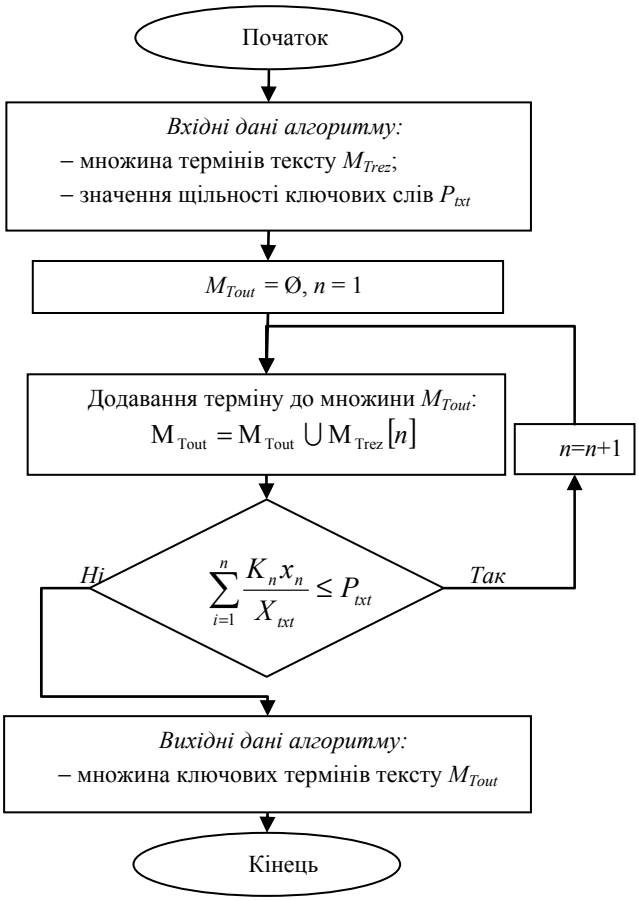

Рис. 3. Алгоритм формування множини ключових термінів тексту M_{Tout}

Перевірка умови рекурсії (Блок 2.7) здійснюються шляхом проведення перевірки, чи терміни всіх розмірностей були визначені протягом проведених ітерацій (рис. 2). Якщо умова $x = n$ справджується, то поточна ітерація визначається фінальною; якщо ж умова хибна ($x<n$), то виконується ще одна ітерація зі збільшенням розмірності термінів, що аналізуються ($x = x+1$).

Завершальна ітерація (Блок 3) акумулює результати всіх ітерацій пошуку термінів та формує вихідні дані інформаційної технології. Так, *об'єднання множин термінів різної розмірності* (Блок 3.1) полягає в додаванні результуючих множин термінів M_{T2} кожної з ітерацій до загальної множини ключових термінів M_{Trez}. Одержана множина термінів M_{Trez} містить ключові терміни всіх розмірностей ($x \le n$), у яких дисперсійна оцінка більша нуля.

На етапі *сортування результуючої множини термінів* (Блок 3.2) Елементи множини ключових термінів M_{Trez} сортуються за зменшенням їх дисперсійної оцінки, після чого їх кількість обмежується.

Обмеження результуючої множини термінів (Блок 3.3) є заключним етапом формування множини ключових термінів тексту. Кількість елементів в одержаній вихідній множині ключових термінів M_{Tout} визначається відповідно до показника граничної щільності ключових слів [8]. Щільність ключових слів P_{txt} є відношенням кількості слів ключових термінів в тексті до загальної кількості слів у тексті й для навчальних матеріалів становить 11–15% (рис. 3). Відповідно, до порожньої результуючої множини ключових термінів M_{Tout} додаються терміни з множини M_{Trez} з найбільшими значеннями оцінки важливості доти, доки справджується рівність:

$$\sum_{i=1}^{n} \frac{K_n x_n}{X_{txt}} \le P_{txt}, \quad (1)$$

де K_n – кількість появ терміну n в множині M_{T1}; x_n – кількість слів у терміні n; X_{txt} – загальна кількість слів у тексті; n – поточна кількість термінів у множині M_{Tout}.

Вихідними даними інформаційної технології інформаційної технології рекурсивного пошуку

ключових термінів у цифрових текстах є множина ключових термінів тексту M_{Tout}, що в відповідній програмній системі може бути виведена користувачеві на екран або збережена у базі даних для подальшого використання.

Прикладна реалізація інформаційної технології

Розроблена інформаційна технологія рекурсивного пошуку ключових термінів у цифрових текстах була реалізована в тестовому програмному продукті. Вхідними даними для системи є електронний документ із цифровим текстом (рис. 4), а вихідними даними є множина ключових термінів, відповідна досліджуваному фрагменту текстового контенту електронного документу (рис. 5). Для написання програмного продукту на платформі .NET було використано мову програмування C# та розширення Spire.Doc.dll для аналізу рівнів структури документу Heading та доступу до елементів контенту TextRange, який є найнижчим рівнем структури документу та визначає фрагменти тексту однакового стилю [10].

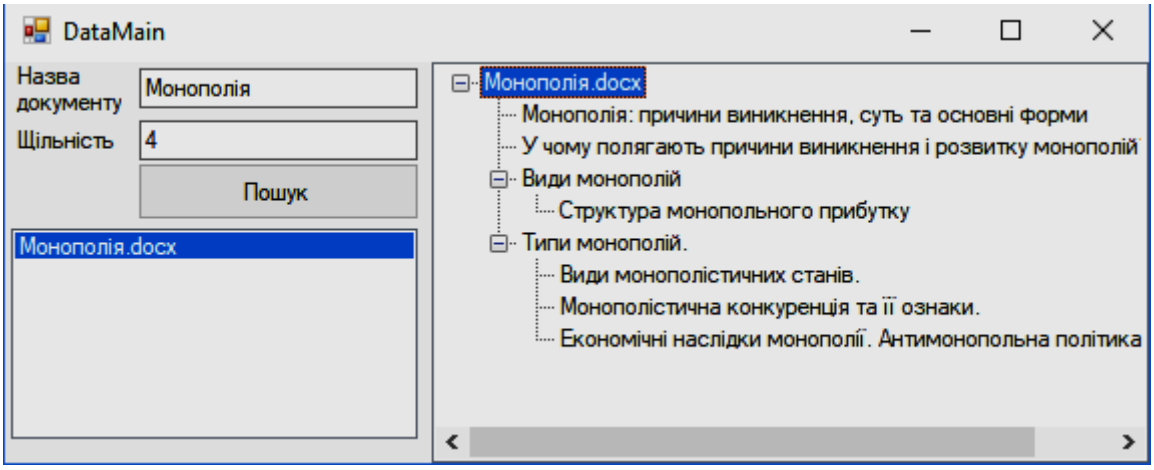

Рис. 4. Форма вибору тексту для аналізу в тестовому програмному продукті

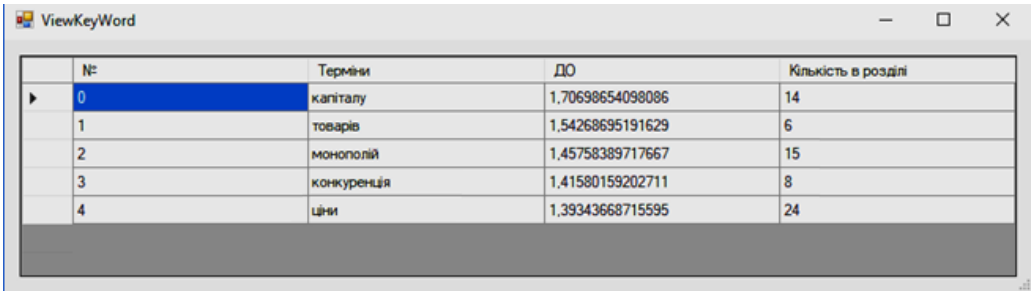

Рис. 5. Форма відображення вихідних даних у тестовому програмному продукті

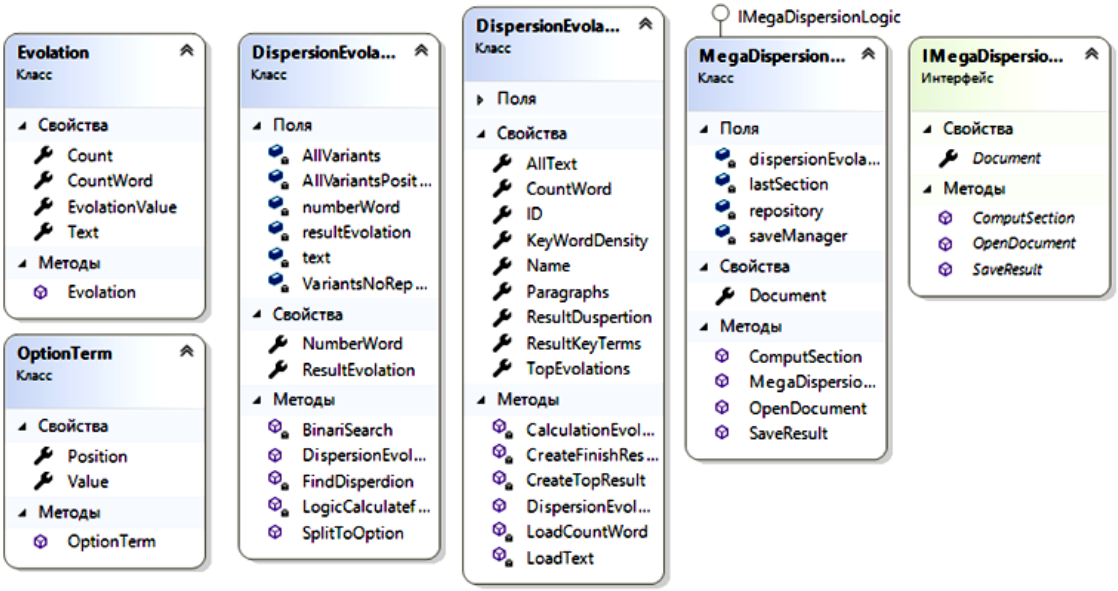

Рис. 6. Діаграма класів модулю розрахунків тестового програмного продукту

Структуру програмного продукту, що відповідає безпосередньо за рекурсивний пошук ключових

термінів у цифрових текстах, у вигляді діаграми класів модулю розрахунків відображено на рисунку 6. Відповідно до діаграми, клас OptionTerm використовується в процесі розрахунку для збереження проміжних даних про терміни-кандидати, а саме значення терміну та його позиції в тексті, що аналізується. Клас Evolation призначений для збереження даних про ключові терміни, зокрема кількість появ терміну в тексті, кількість слів в терміні, вагу терміну і символьне значення терміну. Клас DispersionEvolation реалізує пошук та оцінку ключових термінів, які складаються *x* слів. Головним методом, який проводить оцінку термінів, є «FindDisperdion», який повертає перелік із ключових термінів з оціночними даними класу типу Evolation.

Клас «DispersionEvolationOfSection» є основним в розрахунку ключових термінів. Його робота базується на результаті попереднього класу, оскільки попередній лише проводить оцінку термінів. Даний клас групує його результати, чим формує перелік з найвищими результатами незалежно від кількості слів в терміні, що реалізовано в методі «CreateTopResult». Також в класі у методі «CreateFinishResult» реалізовано обрахунок кінцевого результату з урахуванням щільності ключових термінів, яку задає користувач. Решта методів є другорядними щодо мети дослідження й відповідають за роботу з документами .docx, взаємодію користувача з програмою або візуалізацію результатів роботи системи. Кінцевим результатом роботи тестового програмного продукту є множина ключових термінів тексту. Таким чином, в тестовому програмному продукті реалізовано етапи обробки даних відповідно до інформаційної технології рекурсивного пошуку ключових термінів у цифрових текстах.

Дослідження ефективності інформаційної технології

Ефективність практичного застосування запропонованої інформаційної технології було визначено шляхом використання наведеного тестового програмного продукту за показниками точності (Precision) та повноти (Recall) [11]. Точність пошуку *P* – це відношення кількості релевантних ключових термінів, знайдених автоматично, до загальної кількості знайдених ключових термінів у досліджуваному тексті. Повнота пошуку *R* – це відношення кількості релевантних ключових термінів, знайдених автоматично, до загальної кількості релевантних ключових термінів у досліджуваному тексті.

Точність пошуку *P* і повнота пошуку *R* обчислюються наступним чином:

$$P = \frac{\left|M_{TK}^{E} \bigcap M_{TK}\right|}{\left|M_{TK}\right|}, R = \frac{\left|M_{TK}^{E} \bigcap M_{TK}\right|}{\left|M_{TK}^{E}\right|}, \quad (2)$$

де M_{TK}^{E} – множина релевантних ключових термінів, сформована експертом; M_{TK} – множина знайдених автоматично ключових термінів.

З метою визначення ефективності практичного застосування розробленої інформаційної технології, тестовим програмним продуктом було оброблено тестову вибірку з 40 файлів, що містили досліджуваний текстовий контент.

Отримані множини ключових термінів порівнювалися з множинами ключових термінів, сформованих авторами текстів, шляхом обрахунку показників точності пошуку *P* та повноти пошуку *R*. За результатами обрахунку було одержано відповідно дві множини по 40 показників кожна, з використанням яких було обчислено показники середньої точності пошуку $\overline{P}$ та середньої повноти пошуку $\overline{R}$ за наступними формулами:

$$\overline{P} = \frac{\sum_{i=1}^{k} P_k}{k}, \overline{R} = \frac{\sum_{i=1}^{k} R_k}{k}, \quad (3)$$

де *k* – кількість текстів у тестовій вибірці.

Середня точність пошуку склала 0,705, а повнота пошуку склала 0,512. Мінімальна точність пошуку склала 0,385, мінімальна повнота пошуку – 0,458; максимальна точність пошуку – 0,896, максимальна повнота пошуку – 0,933 (рис. 7).

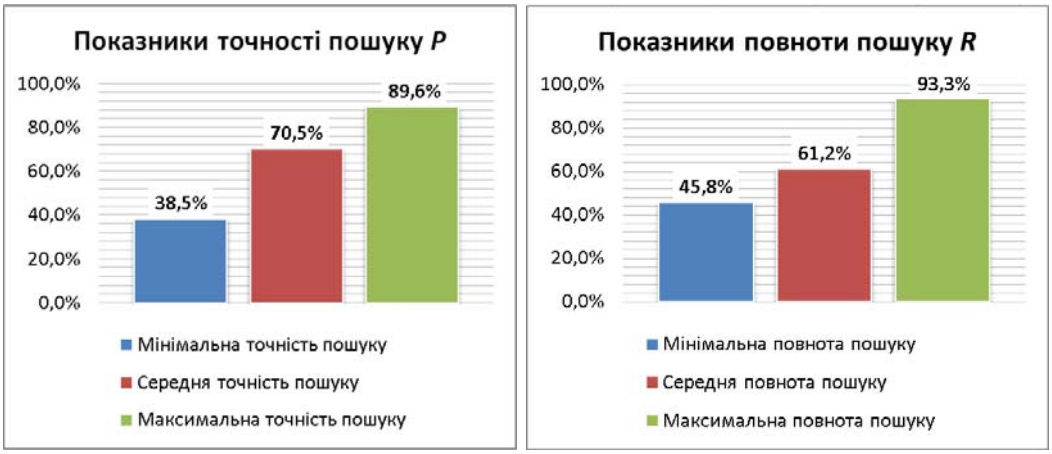

Рис. 7. Результати дослідження ефективності інформаційної технології (точність та повнота пошуку)

Оскільки в задачах розробки даної інформаційної технології було відзначено відсутність необхідності використання лексичних баз даних корпусів слів для ідентифікації слів у текстах, було порівняно одержані результати дослідження ефективності з відповідними результатами, одержаними при використанні подібної інформаційної технології [8], яка використовує лексичну базу даних корпусу слів української мови для ідентифікації слів. Для аналізу використовувались тотожні вибірки файлів із текстовим контентом. Результати порівняння надано на рис. 8.

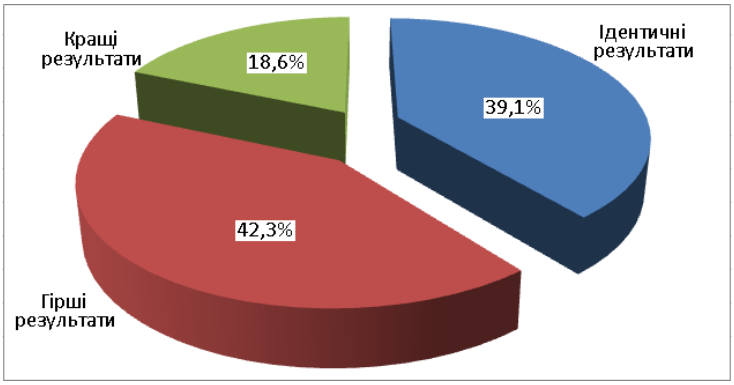

Рис. 8. Результати порівняння ефективності інформаційної технології

Таким чином, результати порівняння ефективності інформаційної технології рекурсивного пошуку ключових термінів у цифрових текстах визначено неоднозначні, уникнення використання лексичних баз даних корпусів слів для ідентифікації слів у текстах у 42,3% випадках негативно вплинуло на якість результату пошуку, проте в 18,6% випадків такий підхід виявив навіть кращий результат.

Дискусія

Перевагами розробленої інформаційної технології рекурсивного пошуку ключових термінів у цифрових текстах, яка проводить аналіз текстового контенту із використанням методу дисперсійної оцінки, є:

1) відсутність необхідності використання лексичних баз даних корпусів слів;
2) суттєве прискорення швидкодії (як наслідок п.1);
3) можливість використання для текстів на різних мовах;
4) можливість використання для текстів із кількома мовами;
5) кращі результати при обробці вузькоспеціалізованого контенту.

У зв'язку з предметною областю, що містить неоднозначні та важко формалізовані сутності, відзначено наступні фактори, що ускладнюють процес оцінки ефективності інформаційної технології:

1) деякі семантично важливі терміни тексту автори суб'єктивно ігнорують;
2) деякі терміни автори включають до переліку ключових, хоча вони розглядаються мінімально;
3) на деяких термінах автори акцентують надмірну увагу попри їх семантичну другорядність в тексті.

Тому відсутність програмно визначених термінів у множині автора не завжди характеризує недолік розглядуваної технології, як і результати порівняння з іншими методами пошуку ключових термінів.

Зважаючи на переваги, розроблена інформаційної технології рекурсивного пошуку ключових термінів у цифрових текстах найбільш ефективно може бути використана для аналізу текстів із невідомими властивостями тематики й мови.

Висновки

В статті розглянуто інформаційну технологію рекурсивного пошуку ключових термінів у цифрових текстах, яка проводить аналіз текстового контенту із використанням методу дисперсійної оцінки та без використання лексичних баз даних корпусів слів. Дана інформаційна технологія може бути ефективно використана для аналізу текстів із невідомими властивостями тематики та мови.

Розроблене відповідно до інформаційної технології рекурсивного пошуку ключових термінів програмне забезпечення в результаті обробки вхідних даних у вигляді цифрового документу формату .docx із текстовим контентом дозволяє одержувати вихідні дані у вигляді відповідної тексту множини ключових термінів.

Проведені за допомогою розробленого відповідно до інформаційної технології тестового програмного забезпечення дослідження підтвердили можливість ефективно автоматизовано формувати множини ключових семантичних термінів текстів із показниками точності пошуку до 89,6% й повноти пошуку до 93,3%.

Подальші дослідження спрямовані на більш детальне дослідження ефективності інформаційної технології й розгорнутий аналіз результатів із метою визначення причин зниження ефективності, а також пошуку областей застосування інформаційної технології для випадків, що характеризуються високими показниками ефективності її застосування.

Література

1. Мазурець О. В. Онтологічний підхід до побудови семантичної моделі навчальних матеріалів / О. В. Мазурець // Вісник Хмельницького національного університету. Серія: Технічні науки. – 2017. – № 6. – С. 223–229.

2. Сергієва О. О. 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.

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

4. SEO-аналізатор «Адвего» [Електронний ресурс]. – Режим доступу : http://advego.ru/text/seo/.

5. SEO-платформа «Serpstat» [Електронний ресурс]. – Режим доступу : https://serpstat.com/

6. Семантичний онлайн-аналізатор тексту «Seozor» [Електронний ресурс]. – Режим доступу : http://seozor.ru/tools/analyzer.php.

7. Ventura J. New Techniques for Relevant Word Ranking and Extraction / J. Ventura, J. Silva // Proceedings of the artificial intelligence 13th Portuguese conference on Progress in artificial intelligence, EPIA'07. – Berlin: Springer-Verlag, Berlin, Heidelberg, 2007. – P. 691–702.

8. Krak Y. The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials / Y. Krak, O. Barmak, O. Mazurets // CEUR Workshop Proceedings, 2139. – 2018. – P. 245–254.

9. Ландэ Д. В. Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения» / КПИ. – Киев, 2013. – С. 158–164.

10. Мазурець О. В. Використання спеціалізованих програмних розширень для автоматизації роботи з цифровими документами навчальних матеріалів / О. В. Мазурець, О. В. Ковальчук, В. О. Слободзян // Вісник Хмельницького національного університету. Серія: Технічні науки. – 2018. – № 1. – С. 61–69.

11. Manning C. Introduction to Information Retrieval / C. Manning, P. Raghavan, H. Schutze – Cambridge University Press, 2008. – 482 p.

References

1. MAZURETS, O. V. (2017) Ontological Approach to Building a Semantic Model of Educational Materials. Herald of Khmelnytskyi national university. Technical Sciences, Issue 6, 2017 (255). p. 223-229.

2. SERHIIEVA, O. O. & MAZURETS, A. V. (2017) Intelligent System of Automated Texts Compression // Collection of scientific works on the materials of the VIth international scientific and practical conference "ICST-ODESSA-2017". p. 223-229.

3. BARMAK, O. V. & MAZURETS, O. V. (2015) Methods of Automation of Definition of Semantic Terms in Educational Materials // Herald of Khmelnytskyi national university. Technical Sciences, Issue 2, 2015 (223). p. 209-213.

4. ADVEGO (2019) SEO-analazer "Advego". [Online] Available from: http://advego.ru/text/seo/ [Accessed: 25 February 2019]

5. SERPSTAT (2019) SEO-analazer "Serpstat". [Online] Available from: http://seozor.ru/tools/analyzer.php [Accessed: 25 February 2019]

6. SEOZOR (2019) Semantic online-analazer of texts "Seozor". [Online] Available from: http://seozor.ru/tools/analyzer.php [Accessed: 25 February 2019]

7. VENTURA, J. & SILVA, J. (2007). New Techniques for Relevant Word Ranking and Extraction. In Proceedings of 13th Portuguese Conference on Artificial Intelligence, Springer-Verlag, p. 691-702.

8. KRAK, Y., BARMAK, O. & MAZURETS, O. (2018) The Practice Implementation of the Information Technology for Automated Definition of Semantic Terms Sets in Content of Educational Materials. CEUR Workshop Proceedings, 2139. p. 245-254.

9. LANDE, D. V. & SNARSKIY, A. A. (2013) Kompaktificirovanniy Gorizontalniy Graf Vidimosti dlya Seti Slov / D.V. Lande, A. A. Snarskiy // Trudi Mejdunarodnoy Nauchnoy Konferencii «Intellektualniy Analiz Informacii IAI-2013. Znania I Rassujdenia». p 158-164.

10. MAZURETS, O. V., KOVALCHYK, O. V. & SLOBODZIAN, V. O. (2018) Using specialized software packeges for automation of work with digital documents of educational materials // Herald of Khmelnytskyi national university. Technical Sciences, Issue 1, 2018 (257). p. 61-69.

11. MANNING, C., RAGHAVAN, P. & SCHUTZE, H. (2008) Introduction to Information Retrieval. Cambridge University Press.

Рецензія/Peer review : 23.3.2019 р. Надрукована/Printed : 2.6.2019 р.
Рецензент: д.т.н., проф. Сорокатий Р. В.