УДК 510.635:004.891 (045)

А. Вавіленкова
Національний авіаційний університет (м. Київ)

СТРУКТУРА СИСТЕМИ ПОРІВНЯЛЬНОГО АНАЛІЗУ ЕЛЕКТРОННИХ ТЕКСТОВИХ ДОКУМЕНТІВ ЗА ЗМІСТОМ

Проаналізовано схему функціонування сервісів для визначення унікальності електронних текстових документів, розглянуто їхні основні характеристики під час перевірки на оригінальність наукової статті. Наведено структуру системи порівняльного аналізу електронних текстових документів за змістом, описано принцип функціонування кожної з її основних компонент.

Ключові слова: природна мова, порівняльний аналіз, пошук, логіко-лінгвістична модель, база знань.

The study analyzes the scheme of services for determining the uniqueness of electronic text documents, considered their main characteristics while checking the originality of the article. An author presents structure of the system for comparative analysis of electronic text documents by content, she outlines the operating principle each of its major components.

Key words: natural language, comparative analysis, search, logic and linguistic model, knowledge base.

Постановка проблеми

Необхідність створення унікального контенту, боротьба за авторські права, поява величезної кількості компаній та сервісів із здійснення рерайтингу, а також безупинний плагіат наукових робіт спричиняють необхідність створення нових якісних механізмів для виявлення копій електронних текстових документів на парадигматичному рівні [1]. Сьогодні в мережі Інтернет існує багато різноманітних сервісів для визначення відсотка унікальності текстових документів. Загальну схему їх функціонування ілюструє рис. 1 [2].

Фактично кожна система визначення відсотка унікальності електронних текстових документів містить два контури: перший відповідає за інструменти порівняння заданого документа з іншими, що знаходяться у базі, а другий шукає документи в мережі Інтернет для занесення до бази та подальшого порівняння. Якщо другий контур у всіх існуючих сервісах функціонує за однаковим принципом і використовує методологію роботи інформаційно-пошукових систем (у цьому разі швидкість роботи системи залежить лише від технічних умов), то перший контур, тобто методика здійснення порівняльного аналізу, і відрізняє сервіси один від одного.

Процес порівняння – це зіставлення об’єктів з метою виявлення спільних рис або різниці між ними. Прийом порівняння використовують у процесі узагальнення, коли необхідно виявити тотожності, збіги та протиріччя в об’єктах дослідження. Тут тотожність – це повний збіг усіх ознак, а протиріччя – коли ознаки одних об’єктів відсутні в інших. Для порівняння необхідні ознаки, що визначають можливі відношення між об’єктами.

Основною проблемою функціонування існуючих сервісів визначення відсотка унікальності електронних текстових документів є відсутність лінгвістичної компоненти порівняння, тобто механізмів, які б зіставляли тексти не лише за кількісними показниками, але й за змістом.

Аналіз останніх досліджень та публікацій

Значних успіхів досягнуто в області інформаційного пошуку, що опосередковано дає можливість підвищити якість визначення відсотка унікальності електронних текстових документів.

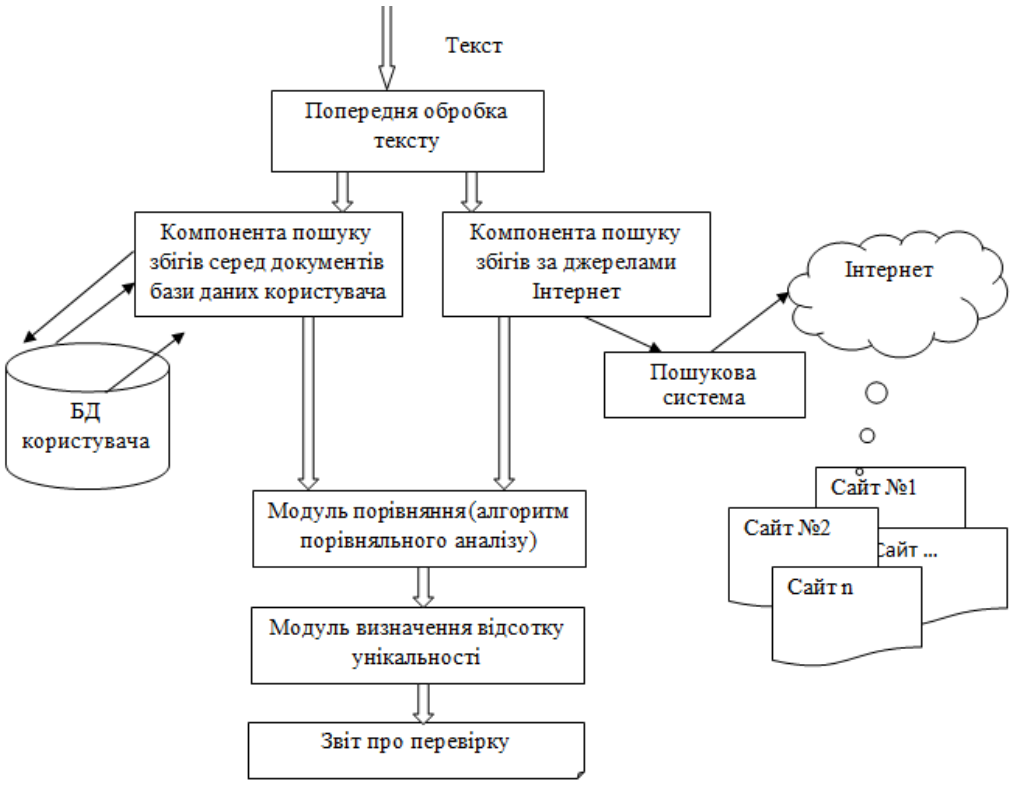


Рис. 1. Схема функціонування системи визначення унікальності електронних текстових документів

Так, у роботі [3] автори пропонують використовувати для моделювання інформаційних потоків фрактальний аналіз, що ґрунтується на властивостях збереження внутрішньої структури масивів документів у разі зміни їх розмірів. У навчальному посібнику "Системи штучного інтелекту" [4] автори піднімають питання розуміння тексту, інтерпретуючи процес сприйняття природної мови через рівні, кожному з яких відповідає конкретний вид абстрактної моделі. Детально схему роботи пошукової системи описано в статті А. М. Глибовця та А. С. Шабінського "Один підхід до побудови інтелектуальної пошукової системи" [5]. Питання організації інформаційного пошуку широко популяризується [6], а технічні можливості його здійснення удосконалюються із року в рік.

Проте питання про якісний змістовний аналіз джерел інформації, що дасть можливість оцінювати унікальність текстів на семантичному рівні, поки що залишається відкритим. Про це свідчить порівняльна характеристика існуючих сервісів визначення унікальності електронних текстових документів (таблиця).

Характеристики функціонування сервісів визначення унікальності електронних текстових документів

Параметр /Назва сервісу	Unplag.com	Content-watch.ru	Strike Plagiarism	Advego Plagiatus	Text.ru
1	2	3	4	5	6
Тип тексту	наук. ст.	наук. ст.	наук. ст.	наук. ст.	наук. ст.
Обмеження довжини тексту	500 слів	10 000 символів	-	-	10 000 символів
Час обробки	1 хв.	30 с.	12 хв.	8 хв.	4 хв.
Відсоток унікальності	**4,63 %**	**100 %**	**-**	**70 %**	**65,77 %**
Процес обробки документа	Без встановлення спеціальних додатків	Безпосередньо на сайті без черги	Без встановлення спеціальних додатків	Зі встановленням додатка на ПК	Безпосередньо на сайті в порядку черги

Продовження таблиці

1	2	3	4	5	6
Методика перевірки	Нейронні мережі	Багатоступінчатий алгоритм власної розробки	Аналіз N-грам	Метод “шинглів”	Алгоритм власної розробки
Концепція отримання результатів порівняльного аналізу	Показує відсоток оригінальності тексту, джерела дублювання та відсоток збігу з ними	Показує відсоток унікальності тексту, джерела дублювання та відсоток збігу з ними	Показує загальний відсоток збігу, джерела дублювання та відсоток збігу з ними	Показує відсоток унікальності тексту, загальний відсоток збігу, джерела дублювання та відсоток збігу з ними	Показує відсоток унікальності тексту, джерела дублювання та відсоток збігу з ними
Можливість порівняння з документами бази даних користувача	+	-	+	-	-
Мова	укр., англ., рос., нім., іспан., турецька, франц., італ.	укр., англ., рос.	укр., англ., рос.	укр., англ., рос.	укр., англ., рос.
Формат файлів	.pdf,.docx,.doc,.odt,.txt,.zip або.html	.txt,.doc,.rtf	.doc,.odt,.txt,.pdf	txt, doc, rtf	txt, doc, rtf
Загальний відсоток збігу, %	**95,37 %**	**-**	**2,6 %**	**100 %**	**-**

Для перевірки на унікальність було взято текст наукової статті (урізаний до 10 000 символів, зі списком посилань та анотацією), що розміщена у науково-метричних базах даних на сайті видавництва та в електронних бібліотеках. Тобто, якщо сервіс обробляє ресурси хоча б з одного джерела, де розміщені сайти з науковими статтями, то відсоток унікальності цієї роботи має бути мінімальним. Із таблиці видно, що сервіси [7–11] оцінювали за такими критеріями, як: час обробки, методика перевірки, можливість порівняння з документами бази даних користувача, мова, формат файлів та відсоток унікальності. Незважаючи на те, що в керівництвах з експлуатації деяких сервісів вказано, що вони функціонують за алгоритмами власної розробки, отримані результати перевірки вказують на те, що в основу механізмів порівняльного аналізу поалкдено модифікований метод “шинглів” або кореляційні методи [12]. Поставлений експеримент дає змогу робити висновки щодо коректності функціонування технічного контуру пошуку інформаційних джерел за повного дублювання тексту електронного файла.

Метод “шинглів” продовжує залишатися основним для порівняльного аналізу двох текстових документів, що означає принципову неможливість якісного лінгвістичного аналізу текстової інформації. Це відбувається через те, що метод “шинглів” ґрунтується на порівнянні контрольних сум деякої послідовності слів, при цьому жодним чином не враховуючи синтаксичних та семантичних зв’язків між ними, а тим більше між реченнями у тексті. Порівняння двох різних текстів, розбитих на однакові за довжиною “шингли”, дасть низький відсоток унікальності лише у тому випадку, коли буде виявлено повний збіг контрольних сум, якщо коли “шингли” будуть ідентичні.

Формулювання цілі статті

Дослідження існуючих сервісів для перевірки електронних текстових документів на плагіат та визначення відсотка їх унікальності (без заглиблення у методики здійснення самого порівняння) показало, що технологія обчислення відсотка оригінальності текстів потребує удосконалення.

Тому метою статті є розкриття основних принципів функціонування систем визначення відсотка унікальності електронних текстових документів та демонстрація роботи системи порівняльного аналізу як підґрунтя для існування інформаційної технології порівняльного аналізу електронних текстових документів за змістом.

Виклад основного матеріалу

Систему порівняльного аналізу створено для реалізації проекту інформаційної технології порівняльного аналізу електронних текстових документів за змістом та верифікації нових методів порівняльного аналізу, запропонованих автором матеріалів [13].

Методологія інформаційної технології порівняльного аналізу текстових документів за змістом передбачає:

– декомпозицію процесу знаходження відсотка збігів між двома електронними текстовими документами, тобто його розбиття на окремі взаємопов'язані складові та етапи;

– програмну реалізацію послідовності виконання операцій з метою підвищення ефективності пошуку відсотка збігів, що відтворена у вигляді системи порівняльного аналізу електронних текстових документів за змістом;

– інструкції щодо виконання окремих дій з аналізу результатів роботи створеної системи.

Отже, система порівняльного аналізу електронних текстових документів за змістом являє собою додаток, створений на основі клієнт-серверної архітектури (рис. 2) і містить такі основні компоненти.

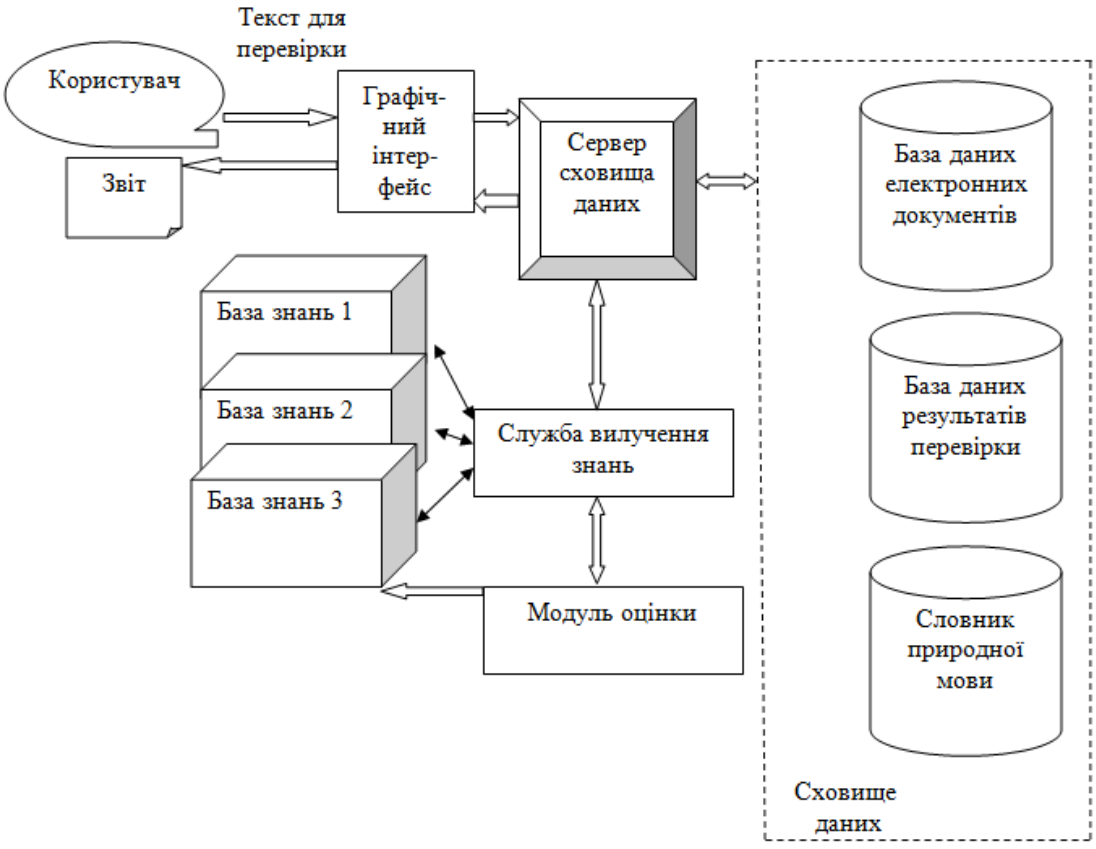


Рис. 2. Архітектура системи порівняльного аналізу електронних текстових документів

Перша компонента системи порівняльного аналізу електронних текстових документів за змістом – це **графічний інтерфейс користувача**, що відповідає за комунікацію між користувачем та системою аналізу даних, візуалізацію отриманих результатів у різноманітних формах [14].

Інтерфейс програми представляє собою вікно для введення тексту, який необхідно порівняти з текстами, що знаходяться в базі та відкриті для пошуку в Інтернет. У поле для введення інформації копіюється текст, що необхідно перевірити на унікальність (рис. 3).

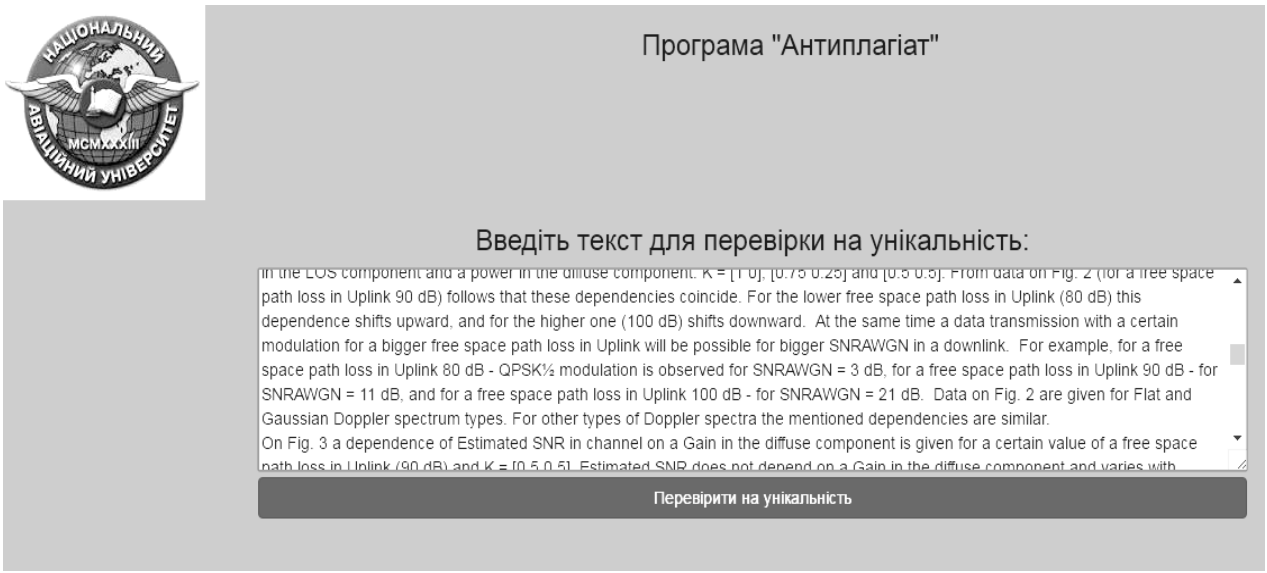


Рис. 3. Інтерфейс системи порівняльного аналізу електронних текстових документів за змістом

Після натискання кнопки “Перевірити на унікальність” система видає результат у вигляді відсотка збігу та посилань на джерела, збіг з якими знайдено (рис. 4).

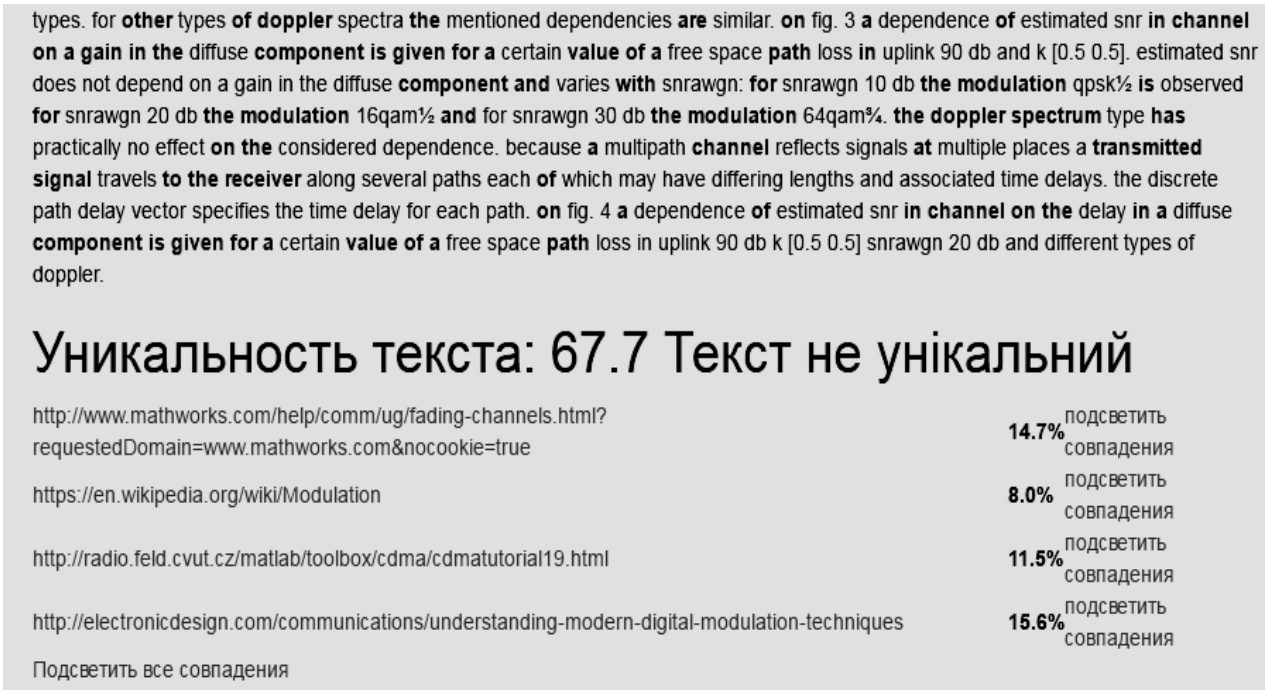


Рис. 4. Результат роботи системи порівняльного аналізу електронних текстових документів за змістом

Унікальність заданого тексту – 67,7 %. Це означає, що текст не унікальний. Найбільший відсоток збігу знайдено з четвертим джерелом, що становить 15,6 % збігів. Текстові збіги, знайдені в інших джерелах, зазначено у тексті жирним шрифтом.

Система порівняльного аналізу електронних текстових документів враховує у роботі такі критерії:

– відсотковий показник, що видається у вікні для перевірки документа, відповідає за унікальність тексту, тобто за те, скільки відсотків нової інформації від автора надано у тексті, що розглядається;

– другий відсотковий показник у цьому самому вікні вказує на загальний відсоток збігів (рерайт), що враховує відсоток збігів за всіма знайденими посиланнями;

– якщо отриманої відсоткові показники 95–100 % / 90–100 %, то система видає висновок про відмінну унікальність;

– якщо відсоткові показники 90–94 % / 90–100 %, то висновок про хорошу унікальність;

– якщо відсоткові показники 80–89 % / 90–100 %, то висновок про задовільну унікальність та можливий рерайт;

– якщо відсоткові показники <80 % / <90 %, то система видає повідомлення про не унікальний текст.

Отже, перший відсотковий показник відповідає за обсяг тексту, що був запозичений із джерела, знайденого за посиланням. Другий відсотковий показник відповідає за обсяг рерайта тексту, знайденого за посиланням (рис. 5).

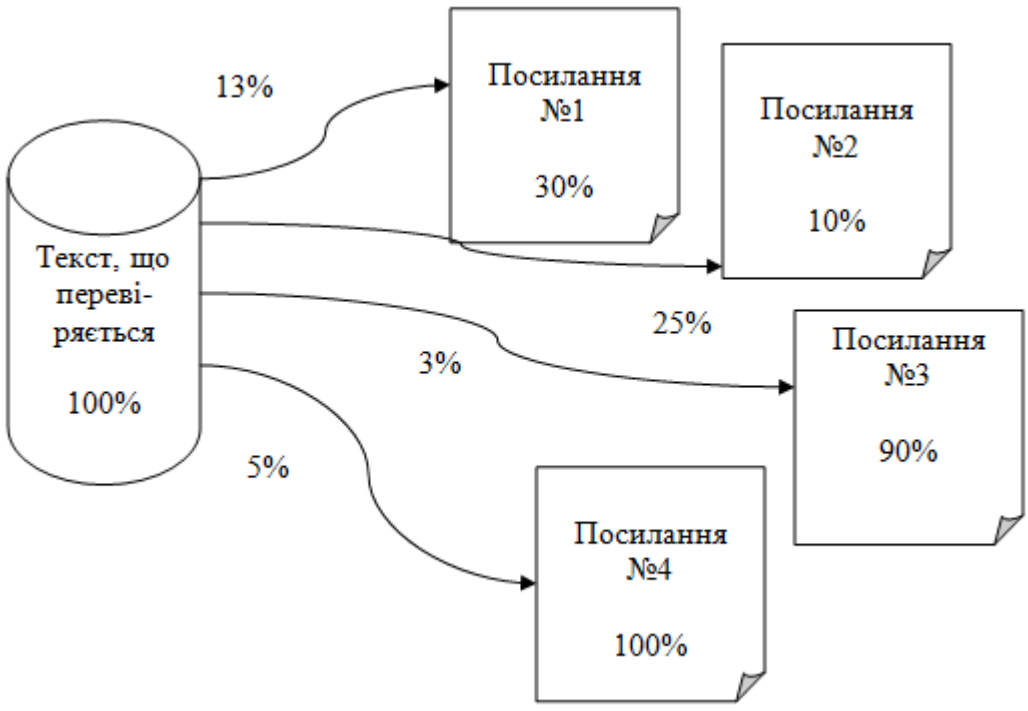


Рис. 5. Графічна інтерпретація визначення унікальності тексту з-поміж декількох запозичених

З рис. 5 видно, що весь текст, який перевіряється, приймається за 100 %, 13 % якого становить 30 % від першого знайденого посилання, 25–10 % від другого посилання, 3 – 90 % посилання № 3 та 5 % – це повністю скопійоване четверте джерело. Тобто 46 % тексту, що перевіряється на унікальність, запозичено з інших джерел, два з яких майже повністю скопійовано. На цьому етапі система перевіряє посилання зі списку літератури і відповідність між авторством. Наприклад, на унікальність може перевірятися монографія, а як повністю скопійовані джерела

можуть бути знайдені посилання на статті цього самого автора. Знаходження таких відповідностей прямо пропорційно впливає на розрахунок унікальності тексту. Також при визначенні унікальності враховують дублювання списків літератури.

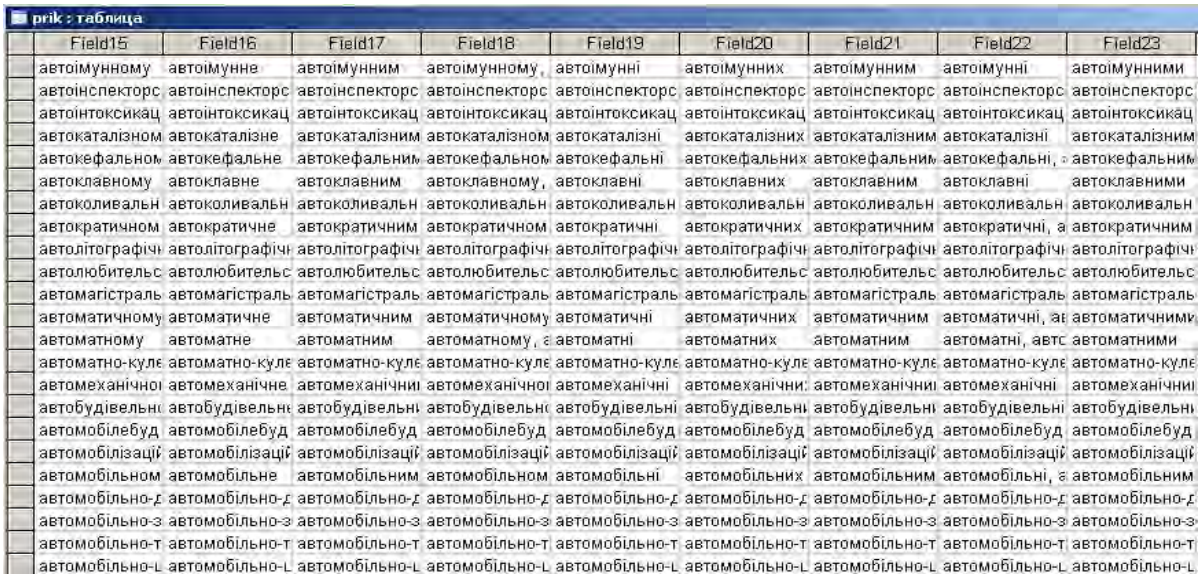
Система порівняльного аналізу електронних текстових документів за змістом шукає джерела не лише в мережі Інтернет, але й в базі даних відкритих дипломних проектів та дисертаційних робіт, забезпечуючи цим повноту отриманої інформації для об'єктивного оцінювання унікальності.

Друга компонента системи порівняльного аналізу електронних текстових документів за змістом – це **сховище даних**, що являє собою сукупність трьох баз даних, які необхідні для видачі сервером результатів на запит користувача.

База даних електронних документів – це сукупність текстових документів, збережених у форматі.txt. Це всі документи, які завантажив користувач системи для перевірки.

База даних результатів перевірки – це реляційні таблиці, у яких міститься інформація про автора, тип роботи, мову, якою написано текст, та результати перевірки (відсоток збігу, список посилань, дата перевірки, кількість слів та символів роботи).

Словник природної мови – це реляційна база даних, кожна із таблиць якої відповідає певній частині мови. Кожне поле таблиці – це слово, що залежно від номеру стовпчику, в якому знаходиться, володіє заздалегідь визначеними граматичними характеристиками, такими як: відмінок, рід, число, час, особа та ін. (рис. 6).


prik : таблица

Field15	Field16	Field17	Field18	Field19	Field20	Field21	Field22	Field23
автоімунному	автоімунне	автоімунним	автоімунному,	автоімунні	автоімунних	автоімунним	автоімунні	автоімунними
автоінспекторс	автоінспекторс	автоінспекторс	автоінспекторс	автоінспекторс	автоінспекторс	автоінспекторс	автоінспекторс	автоінспекторс
автоінтоксикац	автоінтоксикац	автоінтоксикац	автоінтоксикац	автоінтоксикац	автоінтоксикац	автоінтоксикац	автоінтоксикац	автоінтоксикац
автокаталізном	автокаталізне	автокаталізним	автокаталізном	автокаталізні	автокаталізних	автокаталізним	автокаталізні	автокаталізним
автокефальном	автокефальне	автокефальним	автокефальном	автокефальні	автокефальних	автокефальним	автокефальні,	автокефальним
автоклавному	автоклавне	автоклавним	автоклавному,	автоклавні	автоклавних	автоклавним	автоклавні	автоклавними
автоколивальн	автоколивальн	автоколивальн	автоколивальн	автоколивальн	автоколивальн	автоколивальн	автоколивальн	автоколивальн
автократичном	автократичне	автократичним	автократичном	автократичні	автократичних	автократичним	автократичні, а	автократичним
автолітографічн	автолітографічн	автолітографічн	автолітографічн	автолітографічн	автолітографічн	автолітографічн	автолітографічн	автолітографічн
автолюбительс	автолюбительс	автолюбительс	автолюбительс	автолюбительс	автолюбительс	автолюбительс	автолюбительс	автолюбительс
автомагістраль	автомагістраль	автомагістраль	автомагістраль	автомагістраль	автомагістраль	автомагістраль	автомагістраль	автомагістраль
автоматичному	автоматичне	автоматичним	автоматичному	автоматичні	автоматичних	автоматичним	автоматичні, ав	автоматичними
автоматному	автоматне	автоматним	автоматному, а	автоматні	автоматних	автоматним	автоматні, автс	автоматними
автоматно-куле	автоматно-куле	автоматно-куле	автоматно-куле	автоматно-куле	автоматно-куле	автоматно-куле	автоматно-куле	автоматно-куле
автомеханічної	автомеханічне	автомеханічни	автомеханічної	автомеханічні	автомеханічни	автомеханічни	автомеханічні	автомеханічни
автобудівельн	автобудівельн	автобудівельн	автобудівельн	автобудівельні	автобудівельн	автобудівельн	автобудівельні	автобудівельн
автомобілебуд	автомобілебуд	автомобілебуд	автомобілебуд	автомобілебуд	автомобілебуд	автомобілебуд	автомобілебуд	автомобілебуд
автомобілізаці	автомобілізаці	автомобілізаці	автомобілізаці	автомобілізаці	автомобілізаці	автомобілізаці	автомобілізаці	автомобілізаці
автомобільном	автомобільне	автомобільним	автомобільном	автомобільні	автомобільних	автомобільним	автомобільні, а	автомобільним
автомобільно-д	автомобільно-д	автомобільно-д	автомобільно-д	автомобільно-д	автомобільно-д	автомобільно-д	автомобільно-д	автомобільно-д
автомобільно-з	автомобільно-з	автомобільно-з	автомобільно-з	автомобільно-з	автомобільно-з	автомобільно-з	автомобільно-з	автомобільно-з
автомобільно-т	автомобільно-т	автомобільно-т	автомобільно-т	автомобільно-т	автомобільно-т	автомобільно-т	автомобільно-т	автомобільно-т
автомобільно-L	автомобільно-L	автомобільно-L	автомобільно-L	автомобільно-L	автомобільно-L	автомобільно-L	автомобільно-L	автомобільно-L

Рис. 6. Внутрішня організація таблиці прикметників реляційної бази даних системи порівняльного аналізу електронних текстових документів за змістом

Після звернення до словника кожному слову тексту, що подається на вхід системи порівняльного аналізу електронних текстових документів за змістом, ставиться у відповідність масив характеристик. Це значно спрощує подальшу роботу з порівняння та оцінювання унікальності текстів.

Служба вилучення знань системи порівняльного аналізу електронних текстових документів за змістом відповідає за безпосереднє перетворення тексту, що перевіряється на унікальність, на логіко-лінгвістичну (формальну) модель. Для цього використовують алгоритм побудови логіко-лінгвістичної моделі речення природної мови, метод автоматизованого формування логіко-лінгвістичних моделей текстової інформації та алгоритм побудови логіко-лінгвістичної моделі текстового документа [13]. Для побудови логіко-лінгвістичної моделі тексту, що перевіряється на унікальність, використовують три бази знань.

Перша база знань містить правила формування словосполучень природної мови, на основі яких будується логіко-лінгвістична модель речень природної мови. Приклад програмної реалізації правила формування словосполучення природної мови показано на рис. 7.

```
protected void rule01phrase(int val) {
    if(this.sentence[val].getPrioritet() == 0){
      for (int i = val + 1; i < this.sentence.length; i++) {
        //=========================================
        if (((this.sentence[val].selectedProperties("прикметник", 0, 0, 0, 0, 0, 0))
                && (this.sentence[i].selectedProperties("іменник", 0, 0, 0, 0, 0, 0))
                )
                && ((this.sentence[val].equalls(this.sentence[i], this.sentence[val].getC
                    this.sentence[val].getCurrent().getN(),0, 0, 0, 0)))

                ) {
            setText("\nП-1. Словосочетание   -", this.sentence[val], this.sentence[i]);

if(i+1 < this.sentence.length){
    if((this.sentence[i+1].selectedProperties("дієслово",0,0,0,0,0,0))){
    return;
}}
     if(!testSpoluchnikAndImennik(val,i)){
    return;
    }
```

Рис. 7. Правило формування словосполучень із першої бази знань

Фактично кожне правило – це окремий клас.

Друга база знань містить правила, за якими синтезують логіко-лінгвістичні моделі, тобто об’єднують та замінюють їхні структурні компоненти на основі виявлення способів логічного зв’язку між реченнями природної мови.

Третя база знань призначена для формування семантико-синтаксичної складової логіко-лінгвістичної моделі тексту та містить правила визначення типів абзаців, тематичних прогресій, порядку викладення думки тощо, що наділяє текст змістом і створює його текстуальність.

Модуль оцінки системи порівняльного аналізу електронних текстових документів за змістом використовує метод порівняльного аналізу простих речень природної мови, метод порівняння речень природної мови довільної складності, методи порівняльного аналізу логіко-лінгвістичних моделей текстових документів [13]. Цей модуль безпосередньо обчислює відсоток збігів та відсотку унікальності заданого тексту.

Висновки

Створення системи порівняльного аналізу електронних текстових документів передбачає можливість автоматично здійснювати такі функції:

– отримувати з речення природної мови довільної складності набір словосполучень;

– отримувати граматичні характеристики кожного слова тексту;

– надавати вичерпну інформацію про граматичні параметри слів, навіть якщо для формування словосполучення підходить лише один набір таких параметрів;

– автоматично обирати коректні граматичні характеристики слова для формування словосполучення за наявності омонімії;

– надавати інформацію про те, за якими лінгвістичними правилами сформовано словосполучення;

– автоматично формувати логічні зв’язки між концептами речення природної мови з наданням інформації про їхні характеристики;

– автоматично порівнювати речення природної мови;
– обчислювати відсоток збігу з виявленням частин речення, що дублюються за змістом;
– надавати інформацію щодо умов тотожності речень природної мови;
– автоматично порівнювати електронні природномовні тексти;
– обчислювати відсоток збігу з виявленням частин текстів, що дублюються за змістом;
– надавати інформацію щодо умов тотожності природномовних текстів;
– надавати розширену оцінку тотожності текстів.

Теоретичні основи, що покладено в основу роботи системи порівняльного аналізу електронних текстових документів, можна використовувати для удосконалення інформаційного пошуку, SEO-оптимізації, синтезу текстів та їх відновлення.

1. Марчук Ю. Н. Компьютерная лингвистика: учеб. пособие / Ю. Н. Марчук. – М.: АСТ: Восток-Запад, 2007. – 317 с. 2. Шарапов Р. В. Система проверки текстов на заимствования из других источников [Электронный ресурс] / Р. В. Шарапов, Е. В. Шарапова // 13-я Всероссийская научная конференция "Электронные библиотеки: перспективы и методы в технологии, электронные коллекции (RCDL' 2011)": сб. трудов. – Воронеж, 2011. – С. 121–126. – Режим доступа: http://ceur-ws.org/Vol-803/paper16.pdf. 3. Ландэ Д. В. Интернетика: навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ , А. А. Снарский , И. В. Безсуднов. – М.: Либроком, 2009. – 264 с. 4. Глибовець М. М. Системи штучного інтелекту: навч. посібник / М. М. Глибовець, О. В. Олецький. – К.: Вид-во "Км "Академія", 2002. – 366 с. 5. Глибовець А. М. Один підхід до побудови інтелектуальної пошукової системи / А. М. Глибовець, А. С. Шабінський // Наукові записки НаУКМА. – (Серія "Комп'ютерні науки"). – 2010. – Т. 112. – С. 26–29. 6. Коцюба И. Ю. Основы проектирования информационных систем: учеб. пособ./ И. Ю. Коцюба, А. В. Чунаев, А. Н. Шиков. – СПб.: Университет ИТМО, 2015. – 206 с. 7. Plagiarism Detection Engine [Електронний ресурс]. – Режим доступу: https://ua.unplag.com /check-paper-for-plagiarism/. 8. Content watch [Електронний ресурс]. – Режим доступу: https://content-watch.ru. 9. StrikePlagiarism [Електронний ресурс]. – Режим доступу: http://strikeplagiarism.com. 10. Проверка уникальности текста Advego Plagiatus [Электронный ресурс]. – Режим доступа: http://advego.ru/. 11. Text.ru [Електронний ресурс]. – Режим доступу: https://text.ru/about. 12. Позняк Д. Уникальность текста по text.ru и content-watch.ru – пароход идет! [Электронный ресурс] / Д. Позняк, 2013. – Режим доступа: http://www.toboom.name/2013/04/text-ru-content-watch-ru.html. 13. Вавіленкова А. І. Структура інформаційної технології порівняльного аналізу текстових документів / А. І. Вавіленкова // Технічні науки та технології. – 2016. – № 1 (3). – С. 103–109. 14. Вавіленкова А. І. Структура системи порівняльного аналізу електронних текстових документів / А. І. Вавіленкова // Міжнар. наук.-техн. конф. "Інтелектуальні технології лінгвістичного аналізу": тези доповідей. – К.: НАУ, 2016. – С. 18.