УДК 004.42

В. Б. Мокін, д. т. н., проф.;

В. В. Войтко, к. т. н., доц.;

С. В. Бевз, к. т. н., доц.;

О. В. Гавенко, студ.;

І. А. Білоус, студ.

АВТОМАТИЗОВАНА СИСТЕМА ПЕРЕВІРКИ ТЕКСТІВ НА ПЛАГІАТ

Запропоновано моделі автоматизованої системи перевірки текстів на плагіат та пошукові методи символьної ідентифікації. Розроблена автоматизована система дозволяє визначати коефіцієнти наявності плагіату в досліджуваних документах.

Вступ

Сучасний розвиток інформаційних технологій, а саме розвиток мережі Інтернет, значно спростив процес обміну інформацією. Доступність інформаційних ресурсів безперечно сприяє розвитку особистості та суспільства в цілому. У зв'язку з цим, актуальною постає задача ідентифікації даних у процесі аналізу інформаційних потоків, що є головним завданням сучасних пошукових систем, спрямованих на підбір унікального контенту Web-ресурсів. Як відомо, популярні пошукові системи, такі, наприклад, як Google, при виведенні списку сайтів за запитом користувача віддають перевагу сайтам з унікальним контентом [1]. Крім того, важливого значення набуває проблема визначення унікальності даних в інформаційному просторі з метою виключення явища плагіату у процесі ідентифікаційного аналізу наявних ресурсів баз даних. Такі питання є актуальними, перш за все, для вищих навчальних закладів та науково-дослідних установ.

Специфіка призначення автоматизованих систем пошуку плагіату (АСПП) зумовлює потребу зміни та модифікації існуючих методів і засобів встановлення унікальності даних. Так, наприклад, популярні Web-ресурси, такі як copyscape, Web-content, findcopy.ru [2], що спеціалізуються на визначенні унікальності текстової інформації, є практично непридатними для вирішення задачі пошуку плагіату, оскільки не мають засобів статистичної оцінки ідентифікованої кількості повторюваних семантичних масивів та не враховують можливості зміни структури наявного інформаційного ресурсу. Найвідомішим в Україні спеціалізованим програмним забезпеченням такого типу є програми, які використовують експерти ВАК України для перевірки дисертацій на плагіат [3]. Але розробка нових та удосконалення відомих методів і засобів пошуку плагіату та їх реалізація в автоматизованій пошуковій системі залишається актуальною.

Метою роботи є підвищення швидкодії та забезпечення універсальних характеристик методів пошуку плагіату. Під об'єктом дослідження розуміємо пошук плагіату шляхом аналізу текстових масивів даних. Предметом дослідження є методи і засоби пошуку плагіату.

Основними задачами є розробка принципів, методів і алгоритмів пошуку плагіату в інформаційних ресурсах (текстових файлах та базах даних) та їх програмна реалізація; розробка моделі автоматизованої пошукової системи, яка реалізуватиме запропоновані методи та дозволить користувачеві задавати режими роботи за пріоритетністю швидкодії чи універсальних характеристик пошукових процесів.

Розробка методів та засобів пошуку плагіату

Єдиного загальноприйнятого визначення терміну «плагіат» не існує [4]. Сутність плагіату розкривають його базові принципи:

1) копіювання чужої праці (як без, так і з відома автора) та оприлюднення її під своїм іменем [5];

2) презентація суміші власних та запозичених аргументів без належного цитування першодже-

рел використаної інформації [6—8];

3) подання чужих ідей як своїх власних [2];

4) перефразування чужої роботи без належного оформлення посилань на автора [9—11];

5) фальсифікація результатів роботи [10];

6) використання власних, раніше опублікованих ідей у нових наукових роботах [8, 11, 12] та ін.

З огляду на вказані принципи плагіату, що презентують широкий спектр його характеристик, пошукові методи потребують впровадження різних алгоритмів ідентифікації з метою забезпечення надійності роботи пошукових систем. Так, перший і другий принципи плагіату зумовлюють необхідність розробки лінгвістичних алгоритмів ідентифікації текстових даних, орієнтованих відповідно на пошук еталонних семантичних масивів заданої довжини.

Третій принцип потребує використання більш складних алгоритмів пошуку даних за набором ключових слів та подальшого інтелектуального аналізу змісту ідентифікованих ідей.

Четвертий принцип плагіату полягає у зумисному перефразуванні результатів чужої роботи, тому пошукові алгоритми повинні проводити циклічну перевірку лінійних семантичних масивів на предмет виявлення фактів перестановки слів у реченні чи заміни деяких літер їх аналогами з іншомовних алфавітів. Крім того, алгоритми цього класу потребують додаткових засобів ідентифікації модифікованих слів з урахуванням зміни структури слова (зміни закінчення чи префіксально-суфіксальної модифікації) та здійснення пошукових процесів у спеціалізованих базах синонімічних архівів з метою виявлення фактів можливої заміни обраних слів їх синонімами. Тому метод пошуку плагіату за четвертим принципом ідентифікації забезпечує універсальні характеристики пошукових процесів, проте, у свою чергу, потребує значних часових затрат на виконання ресурсоємних операцій.

П’ятий принцип плагіату, за визначенням, ускладнює умову пошукової задачі, тому передбачає використання експертних систем у процесі аналізу текстових масивів даних.

Шостий принцип зумовлює додаткове включення в оперативну ідентифікаційну базу даних текстів з авторством особи, праці якої перевіряють.

З метою забезпечення швидкодійних режимів роботи пошукової системи розглядається можливість вибору обмежень ідентифікаційних вимог та реалізація буферного принципу збереження оперативних інформаційних ресурсів, що сприяє оптимізації часових затрат на проведення пошукових операцій.

Розробка моделі структури автоматизованої системи перевірки текстів на плагіат

Узагальнена модель автоматизованої системи перевірки текстів на плагіат (рис. 1) презентує структурну взаємодію базових блоків, що формують програмне середовище АСПП.

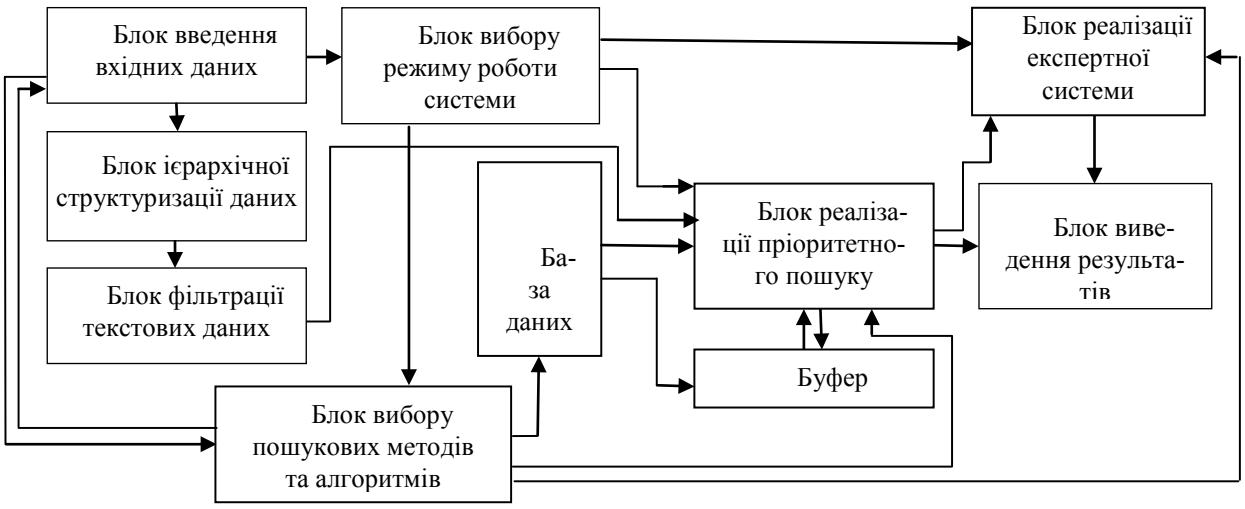

Рис. 1. Узагальнена модель структури АСПП

Блок введення вхідних даних (рис. 2) реалізує інтерфейсну людино-машинну взаємодію в середовищі АСПП, забезпечуючи початковий та діалоговий режими введення інформації.

Блок ієрархічної структуризації даних (див. рис. 1) формує набір елементарних об’єктів дослідження (рис. 3). Такий підхід дозволяє підвищити швидкодію пошукових процесів та забезпе-

чити функціональну незалежність алгоритмів роботи від вхідних даних.

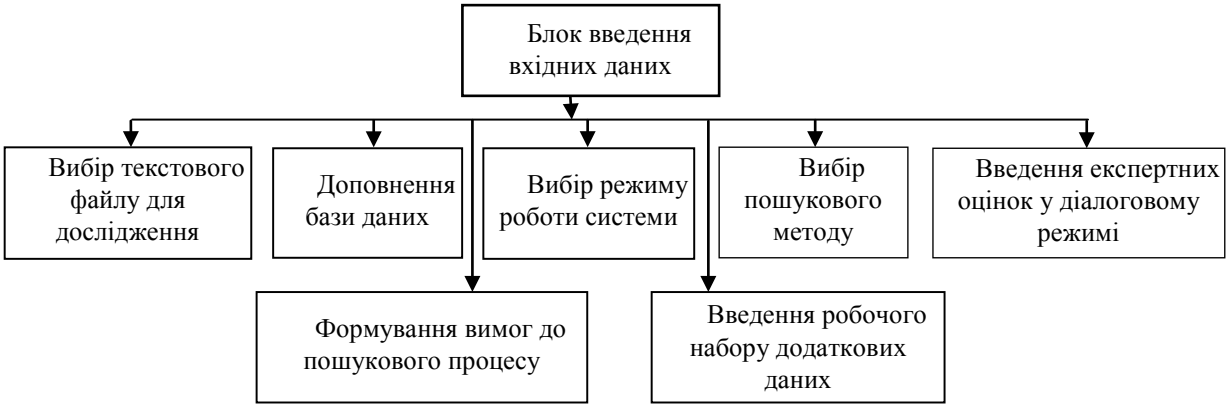

Рис. 2. Структура блоку введення вхідних даних

Ієрархічна структуризація вхідної інформації зумовлює підтримку єдиного підходу до формування інформаційного забезпечення досліджуваного ресурсу, що передбачає можливість оперування даними нетекстового формату без суттєвих змін базових алгоритмів.

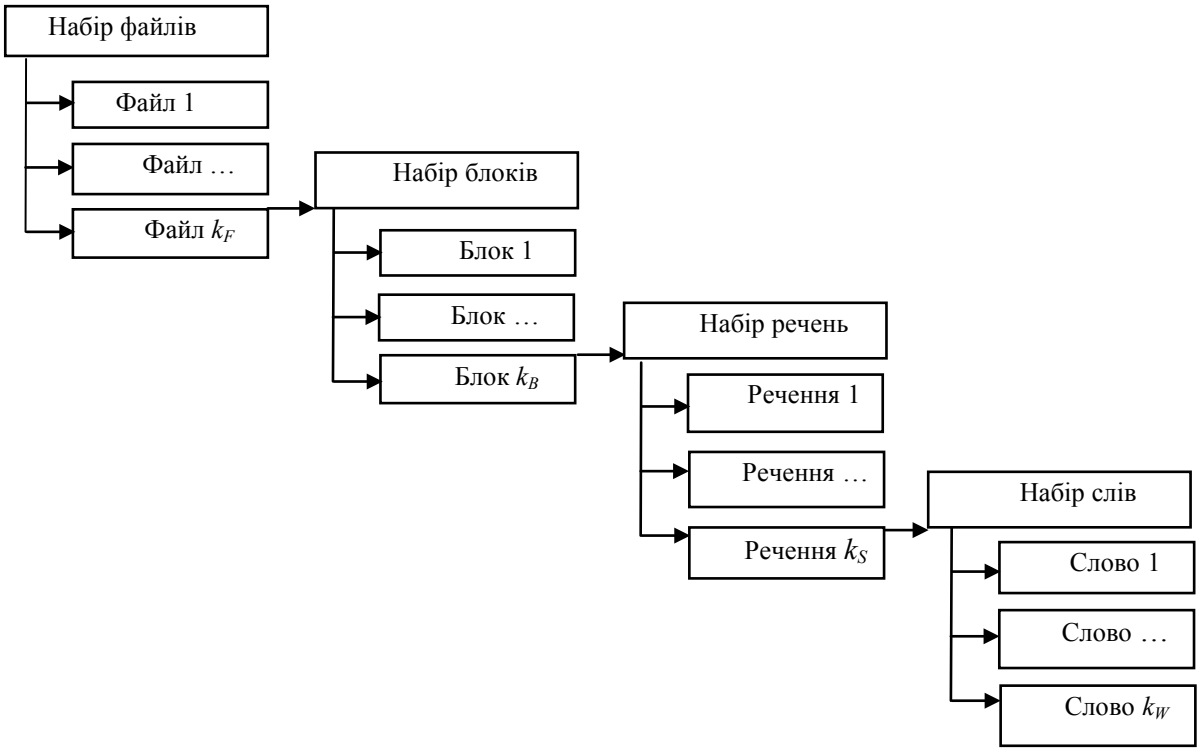

Рис. 3. Структура ієрархічного подання даних

Блок фільтрації (див. рис. 1) дозволяє оптимізувати набір вхідних даних текстового файлу, поданого для дослідження, шляхом мінімізації інформативних ресурсів засобами ієрархічної фільтрації з метою збільшення швидкодії АСПП. Процес фільтрації здійснюється у два етапи. На першому етапі проводиться аналіз вхідного файлу на рівні речень. Відфільтровуються речення, які містять менше, ніж 30 % символів робочого алфавіту, та речення, довжина яких не перевищує 30 символів. На другому етапі виконується процес фільтрації слів, довжина яких обмежується 4 символами. Такий підхід дозволяє суттєво зменшити семантичний набір досліджуваного ресурсу шляхом вилучення неінформативних даних та не впливає на достовірність кінцевого результату пошукового процесу.

Блок вибору режиму роботи АСПП (рис. 4) призначений для встановлення пріоритетності пошукових умов, які впливають на кінцевий вибір методів і алгоритмів перевірки текстів на плагіат у середовищі автоматизованої системи.

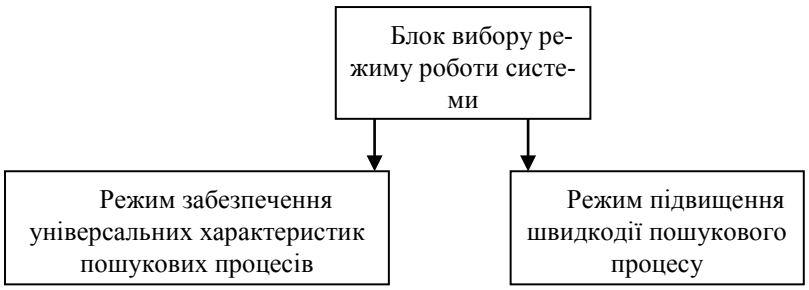

Рис. 4. Режими роботи блока вибору режиму роботи АСПП

Режим підвищеної швидкодії пошукового процесу передбачає мінімізацію набору вхідних вимог, що дозволяє обмежити кількість обраних алгоритмів текстової ідентифікації у вказаному пошуковому методі, та використання буфера пам'яті, куди записуються вибрані з бази даних (БД) робочі інформаційні ресурси з метою їх подальшого дослідження в оперативному режимі.

Блок вибору пошукових методів та алгоритмів (рис. 5) забезпечує вибір набору робочих алгоритмів з урахуванням вимог користувача до умов проведення пошукових процесів. У середовищі АСПП реалізовані базові методи перевірки текстів на плагіат, в основу яких покладені принципи визначення сутності плагіату.

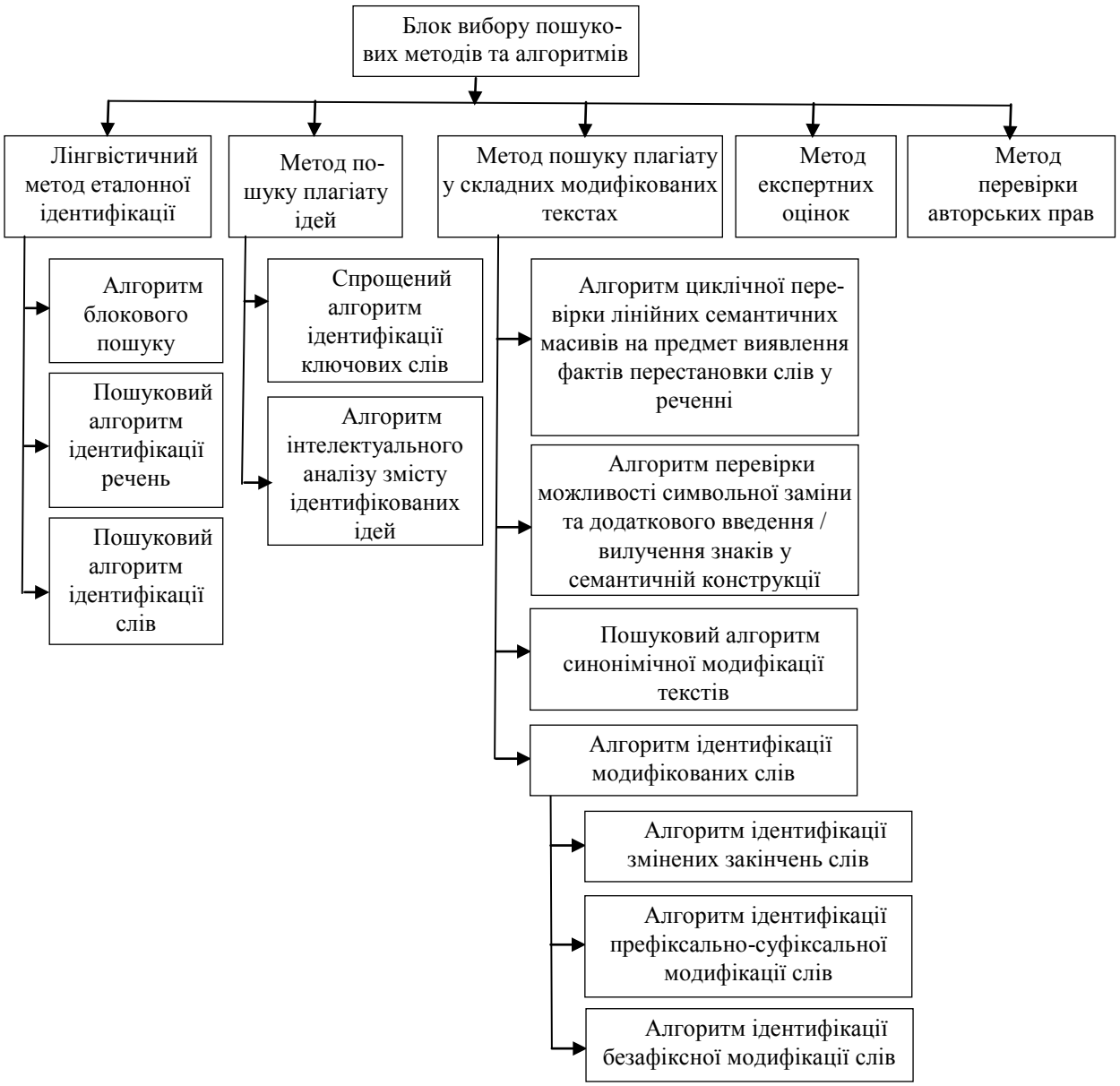

Рис. 5. Структура формування блоку вибору пошукових методів та алгоритмів

Лінгвістичний метод еталонної ідентифікації базується на першому і другому принципах плагіату. Набір алгоритмів методу орієнтований на опрацювання обраних конструкцій структурованих вхідних даних.

Метод пошуку плагіату ідей використовує третій принцип визначення сутності плагіату і реалізує спрощений алгоритм ідентифікації ключових слів за умови підтримки режиму підвищеної швидкодії пошукового процесу та алгоритм інтелектуального аналізу змісту ідентифікованих ідей, який дозволяє проведення ґрунтовніших досліджень.

Метод пошуку плагіату в складних перефразованих текстах розкриває четвертий принцип плагіату. Набір робочих алгоритмів методу дозволяє забезпечення універсального підходу до перевірки текстових документів за умови їх модифікації.

Метод експертних оцінок орієнтований на використання п’ятого принципу плагіату та передбачає залучення професійного експерта для встановлення ступеня фальсифікації результатів роботи. В основу методу покладено математичний апарат нечіткої логіки для автоматизованого аналізу результатів експертних оцінок.

Метод перевірки авторських прав базується на шостому принципі плагіату і передбачає дослідження робіт автора, праці якого перевіряються, використовуючи описані вище алгоритми, вибір яких зумовлюється умовами ведення пошукових процесів.

Блок реалізації пріоритетного пошуку (див. рис. 1) забезпечує підтримку обраних режимів та реалізує виконання вказаних методів з ідентифікованим набором робочих алгоритмів.

Блок реалізації експертної системи (див. рис. 1) містить інтелектуальні методи та алгоритми, зручний інтерфейс, який дозволяє експертам систематизувати та переглядати результати обробки даних за різними принципами, різними методами на різних ітераціях та з різною деталізацією. Система автоматично проводить усі розрахунки, але тільки експерти можуть об’єктивно прийняти остаточне рішення щодо наявності плагіату в тих чи інших частинах інформаційних ресурсів чи ідентифікувати випадковий збіг чи вживання відомих тез. Адже ж дві статті на одну досить вузьку тему можуть мати до 70 % ключових однакових слів та понять, але бути викладенням різних ідей та досягнень. В основу реалізації такої експертної системи слід покладати математичний апарат нечіткої логіки для мінімізації суб’єктивізму експертів-користувачів системи.

Блок виведення результатів (див. рис. 1) презентує результати перевірки у зручному для користувача вигляді. АСПП визначає коефіцієнт плагіату в досліджуваному документі та забезпечує можливість виведення ідентифікованого збігу семантичних конструкцій в ілюстративному режимі. Крім того, програмно забезпечена можливість виведення результатів на друк та їх збереження в окремому файлі. На рис. 6 подано приклад таких результатів, які виводяться авторською програмою, що реалізовує запропоновані у статті методи, структури та алгоритми.

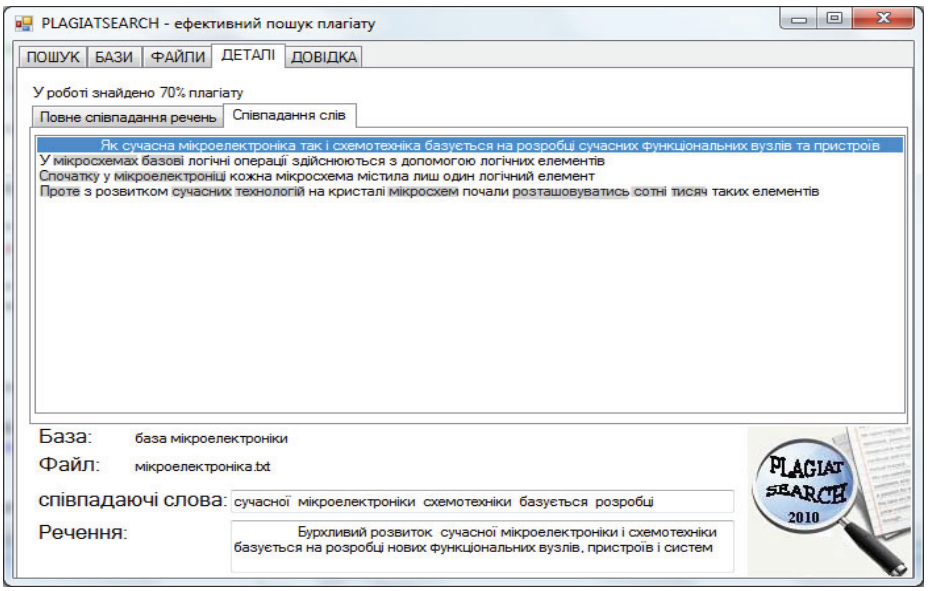

Рис. 6. Приклад результатів пошуку плагіату за лінгвістичним методом еталонної ідентифікації на рівні слів в авторській програмі

Коефіцієнт плагіату визначається відношенням:

$P = 100\frac{k_p}{k_s}$, де k_s – загальна кількість структурних елементів у відфільтрованих вхідних даних; k_p – кількість структурних елементів, у яких було виявлено плагіат.

Висновки

Запропоновано модель та методи роботи автоматизованої системи перевірки текстів на плагіат, які реалізують основні принципи визначення сутності плагіату. Система комплексно забезпечує універсальні характеристики пошукових процесів та підтримує режим підвищеної швидкодії за рахунок вибору обмежень ідентифікаційних вимог, використання засобів ієрархічної фільтрації вхідних даних та реалізації буферного принципу збереження оперативних інформаційних ресурсів, що сприяє оптимізації часових затрат на проведення пошукових операцій. Результатом роботи системи є визначення коефіцієнту наявності плагіату в досліджуваному документі та ілюстративне виведення ідентифікованого збігу семантичних конструкцій. Розроблено та успішно апробовано програму, яка реалізовує запропоновані методи, модель та підходи.

СПИСОК ЛІТЕРАТУРИ

1. Печников В. Н. Создание Web-сайтов / В. Н. Печников. — М. : Технический бестселлер, 2006. — 464 с. — ISBN 5-89392-137-2.

2. Петренко І. Питання виявлення плагіату літературного твору / І. Петренко // Теорія і практика інтелектуальної власності. — 2009. — № 4.

3. Литвиненко О. Є. Система порівняльного аналізу тексту / О. Є. Литвиненко // Інформаційні технології в системі управління освітою України : міжн. наук.-техн. конф, 29—30 верес. 2005 р. : тези доп. — Х., 2005.

4. Бусел В. Т. Великий тлумачний словник сучасної української мови (з додат. і допов.) / Уклад. і голов. ред. В. Т. Бусел. — К. : ВТФ «Перун», 2005. — 1728 с.

5. Святоцький О. Д. Інтелектуальна власність : словник-довід. у 2-х т. / О. Д. Святоцький: Т. 1 Авторське право і суміжні права ; за ред. О. Д. Святоцького, В. С. Дроб'язка. — Уклад. В. С. Дроб'язко, Р. В. Дроб'язко. — К. : Видавничий Дім «Ін Юре», 2000. — 356 с.

6. Липцик Д. Авторское право и смежные права / Д. Липцик ; пер. с фр.; предисловие М. Федотова. — М. : Радомир; Издательство ЮНЕСКО, 2002. — 788 с.

7. Вишневецкий Л. М. Формула приоритета: Возникновение и развитие авторского и патентного права / Л. М. Вишневецкий, Б. И. Иванов, Л. Г. Левин. — Ленинград, 1990. — 205 с.

8. Закон України Про авторське право і суміжні права // Відомості Верховної Ради України (ВВР). — 1994. — № 13. — С. 64.

9. Ионас В. Я. Произведения творчества в гражданском праве / В. Я. Ионас. — М. : Юридическая литература, 1972. — 168 с.

10. Попович О. С. Науково-технологічна та інноваційна політика: основні механізми формування та реалізації / О. С. Попович. — К.: Фенікс, 2005. — 226 с.

11. Закон України «Про наукову і науково-технічну діяльність» // Відомості Верховної Ради України (ВВР). — № 12. — С. 165.

12. Закон України «Про пріоритетні напрями інноваційної діяльності в Україні» // Відомості Верховної Ради (ВВР). — № 13.

Рекомендована кафедрою моделювання та моніторингу складних систем

Надійшла до редакції 4.12.07
Рекомендована до друку 12.02.08

Мокін Віталій Борисович — завідувач кафедри моделювання та моніторингу складних систем, ***Войтко Вікторія Володимирівна*** — доцент кафедри програмного забезпечення, ***Бевз Світлана Володимирівна*** — доцент кафедри електричних станцій та систем, ***Гавенко Олег Віталійович*** — студент Інституту магістратури, аспірантури та докторантури, ***Білоус Ігор Анатолійович*** — студент Інституту інформаційних технологій та комп'ютерної інженерії.

Вінницький національний технічний університет