ПРОБЛЕМИ ПРИЙНЯТТЯ РІШЕНЬ
І УПРАВЛІННЯ В ЕКОНОМІЧНИХ, ТЕХНІЧНИХ,
ЕКОЛОГІЧНИХ ТА СОЦІАЛЬНИХ СИСТЕМАХ

УДК 004.931
DOI: 10.20535/SRIT.2308-8893.2017.4.03

ПІДВИЩЕННЯ ДОСТОВІРНОСТІ ПЕРЕВІРКИ УНІКАЛЬНОСТІ ТЕКСТІВ З ВИКОРИСТАННЯМ КОМБІНОВАНИХ СИСТЕМ РОЗПІЗНАВАННЯ ОБРАЗІВ

О.С. МЕНЯЙЛЕНКО, О.І. ЗАХОЖАЙ, П.І. БІДЮК

Анотація. Крім очевидних переваг пошуку та використання інформації з мережі Інтернет, останнім часом почала проявлятися негативна тенденція — намагання користувачів мережі розпоряджатись нагромадженою інформацією та подавати її як продукт власної творчої праці. Для підвищення достовірності визначення унікальності тексту в роботі запропоновано побудову бази даних текстових фрагментів на основі комбінованої системи розпізнавання образів. Використання запропонованих технічних рішень дало змогу підвищити достовірність визначення частки унікальності тексту в середньому на 54%.

Ключові слова: інформаційні технології, мережа Інтернет, плагіат, достовірність розпізнавання тексту, система для розпізнавання образів.

ВСТУП

Інтенсивний розвиток інформаційних технологій, інструментів і засобів автоматизованого оброблення та зберігання надвеликих масивів інформації створює сприятливі умови для формування глобальних, централізованих і розподілених сховищ даних для численних користувачів. В останні десятиріччя глобальна мережа Інтернет об'єднала ці сховища і разом із сучасними пошуковими сервісами дозволила створити майже необмежений для подальшого розширення та удосконалення ресурс доступу до різноманітної інформації з можливостями розгорнутого пошуку та упорядкування даних. Окрім очевидних переваг такого прогресу, проявляється також і негативна тенденція — намагання використовувати та розпоряджатися нагромадженою інформацією та подавати її як продукт власної творчої праці. На жаль, термін «плагіат» стає дедалі поширенішим і «нормою» для будь-яких сфер людської діяльності. Така ситуація спостерігається, зокрема, у галузі освіти, де під час підготовки студентських робіт, доповідей, рефератів, курсових і дипломних робіт, а іноді й дисертацій суб'єктами освітнього процесу реалізуються спроби запозичення результатів чужої інтелектуальної праці.

Натепер у сучасній науці й освіті стає дедалі популярнішим підхід до написання робіт, який полягає у звичайному копіюванні інформації з одного

або декількох джерел, подальшому редакторському обробленні та поданні отриманих матеріалів як результату особистої інтелектуальної праці [1].

Одним зі способів протидії негативним аспектам є використання спеціальних засобів перевірки текстів на унікальність, які дозволяють за результатами аналізу поданих продуктів інтелектуальної праці виявляти запозичені матеріали, визначати їх обсяг, а також ідентифікувати джерела запозичення [2–9]. Такі засоби отримали назву систем антиплагіату [3, 5, 7].

Принцип функціонування систем перевірки текстів на унікальність полягає у порівнянні поданих текстів з наявними у базі даних. У процесі роботи системи збирається і фільтрується інформація про наявні продукти інтелектуальної праці [5]. Після попереднього оброблення тексти індексуються і вносяться до бази даних. Надалі цю базу даних можна використовувати для порівняння з поданими для аналізу текстами. Кожний документ, завантажений для перевірки, ставиться в чергу для оброблення. Після перевірки необхідного документа система формує звіт, у якому детально подаються усі результати пошуку та рівень унікальності перевіреного матеріалу. Зазвичай ті фрагменти, що не є унікальними, для наочності виділяються в тексті [8].

Існує велике розмаїття програмних продуктів, які дозволяють перевірити текст на унікальність та ідентифікувати ознаки плагіату. Кожний з них може мати специфічні особливості, а також переваги та недоліки [2–6, 9]. Аналіз таких систем дав змогу виявити основні функції, властиві більшості з них, а саме:

- визначення частки унікальності поданого тексту;
- пошук фрагментів тексту, що були запозичені, а також ідентифікація джерел інформації, що були використані у плагіаті;
- маркування запозичених фрагментів тексту кольором для наочності, а також використання різних кольорів для фрагментів, запозичених з різних джерел;
- формування детальних звітів після перевірки тексту на унікальність.

Програмне забезпечення, як правило, має можливість гнучкого налагодження алгоритмів та інструментарію пошуку подібності (наприклад, визначення кількості слів у шинглі, кількість вибірок з тексту і т. ін.).

Додаткові функції програмного забезпечення перевірки текстів на унікальність:

- перевірка контенту різноманітних web-ресурсів на унікальність з формуванням докладного звіту, що містить зафіксовані джерела плагіату;
- робота зі списком проксі;
- пакетна перевірка поданих текстових файлів;
- можливість повторної перевірки текстів після видалення неунікальних фрагментів;
- збереження історії перевірок та ведення журналу подій під час функціонування алгоритму.

Незважаючи на існуючу достатньо велику розмаїтість систем перевірки унікальності текстів, їх поєднує один негативний аспект — неоднозначність отриманих результатів аналізу, рівень достовірності яких істотно залежить від умов перевірки і більшою мірою від організації бази даних текстів, за якими виконується перевірка. Можливість отримання достовірного резуль-

тату аналізу тексту ускладняється застосуванням методів трансформування тексту, які штучно підвищують унікальність тексту, що не дозволяє системі подавати адекватний результат перевірки.

Отже, можна зробити висновок, що удосконалення систем перевірки текстів на унікальність, особливо в частині логічної організації бази даних текстів та алгоритмів перевірки, є актуальним науково-технічним завданням, розв'язання якого надасть можливість ефективно протидіяти незаконному використанню запозичених продуктів інтелектуальної праці.

ПОСТАНОВКА ЗАВДАННЯ

Для визначення основних аспектів завдання, що значною мірою впливають на достовірність результатів перевірки унікальності текстів, було заплановано та проведено експеримент, у якому різним програмам перевірки на плагіат подавались на вхід повністю запозичені текстові фрагменти, їх частка унікальності становила 0%.

В експерименті використано чотири найбільш відомі та поширені системи перевірки на антиплагіат, назви яких з метою недопущення реклами або антиреклами, не будуть розкриватися. Назвемо досліджувані системи: «система 1», «система 2», «система 3», «система 4». Для перевірки достовірності експериментального дослідження на вхід кожної з чотирьох систем подавалися для аналізу повністю запозичених 20 фрагментів тексту (відповідно до планування експерименту це 20 груп дослідів).

Дисперсію середнього значення розрахункової частки унікальності тексту визначали так [10, 11]:

$$s_i^2 = \frac{\sum_{j=1}^{m}(y_{ji} - \overline{y_i})^2}{m-1}, \qquad (1)$$

де m — кількість рівнобіжних дослідів; y_{ji} — відгук j-го рівнобіжного досліду; $\overline{y_i}$ — середній відгук у досліді.

Відтворюваність середнього значення частки унікальності тексту оцінювалась за критерієм Кохрена [10, 11]:

$$G = \frac{s_{i\text{макс}}^2}{\sum_{i=1}^{n} s_i^2},$$

де n — кількість груп дослідів; s_i^2 — дисперсія i-го досліду, яка визначається за формулою (1); $s_{i\,\text{макс}}^2$ — максимальне значення дисперсії в групі дослідів.

Результати виконаного експерименту з аналізу текстових фрагментів, поданих мовою оригіналу, зведено у табл. 1.

Проаналізувавши результати дослідів, поданих в табл. 1, можна зробити висновок, що всі системи майже однаково визначили поданий плагіат. Значення коефіцієнта Кохрена не перевищувало табличного значення відпо-

відно до методики [10]. Такі результати можна назвати очікуваними, оскільки тексти бралися із загально доступних джерел і в базах даних досліджуваних систем містився однаковий обсяг відомостей про наявний плагіат.

Таблиця 1. Результати аналізу текстових фрагментів, поданих мовою оригіналу з очікуваною часткою унікальності 0%

Параметр	Система 1	Система 2	Система 3	Система 4
Середнє значення частки унікальності тексту $\overline{y_i}$	4,65	5,60	10,35	8,65
Середня дисперсія значення частки унікальності тексту $\overline{s_i^2}$	2,33	4,36	2,72	1,74
Значення коефіцієнта Кохрена (за довірчого інтервалу 0,95)	0,37	0,39	0,32	0,31

Оскільки системи перевірки можуть аналізувати тексти за умови їх можливого перекладання іншими мовами, на наступному етапі досліджень виконувалося автоматизоване перекладання текстів загальновідомими системами перекладу з української мови оригіналу на англійську та з української мови оригіналу на російську. Для збереження чистоти експерименту ручне редагування перекладених текстів не проводилося.

Результати аналізу на унікальність текстів, перекладених з української мови на англійську та з української мови на російську, наведено в табл. 2 і 3 відповідно.

Таблиця 2. Результати аналізу текстових фрагментів, перекладених з української мови на англійську з очікуваною часткою унікальності 0%

Параметр	Система 1	Система 2	Система 3	Система 4
Середнє значення частки унікальності тексту $\overline{y_i}$	65,95	62,85	56,18	67,34
Середня дисперсія значення частки унікальності тексту $\overline{s_i^2}$	6,38	8,03	7,64	5,12
Значення коефіцієнта Кохрена (за довірчого інтервалу 0,95)	0,35	0,36	0,34	0,30

Таблиця 3. Результати аналізу текстових фрагментів, перекладених з української мови на російську з очікуваною часткою унікальності 0%

Параметр	Система 1	Система 2	Система 3	Система 4
Середнє значення частки унікальності тексту $\overline{y_i}$	40,30	38,74	44,57	41,05
Середня дисперсія значення частки унікальності тексту $\overline{s_i^2}$	7,32	8,94	7,42	5,04
Значення коефіцієнта Кохрена (за довірчого інтервалу 0,95)	0,36	0,38	0,36	0,32

Аналіз результатів дослідів, поданий в табл. 2 і 3, показав достатньо високу відтворюваність усіма чотирма системами, і переклад тексту з часткою плагіату 100% дає можливість значно підвищити розрахункову частку унікальності. При цьому переклад англійською мовою дає вищий результат унікальності. Отримані результати пояснюються тим, що всі системи перевірки текстів на унікальність фактично здійснюють пошук послідовності символів, що збігаються з фрагментами, наявними в базі даних, повністю на кшталт пошуковим системам. Таким чином, синтаксично текст російською мовою буде мати більшу кількість збігів з текстом українською мовою завдяки більшій кількості однакових символів. Цей аспект може бути використаний плагіаторами для підвищення унікальності тексту через автоматичну підміну на латинські символи, наприклад, в українському тексті букв, що мають однакове написання (тобто підміна «а», «Н», «К», «р» та інших символів кирилиці на відповідні символи латиниці). Візуально це не помітно, але під час аналізу буде отримано вищу частку унікальності.

Очевидно, що найбільш ефективним і універсальним способом підвищення достовірності визначення унікальності тексту може бути повна відмова від синтаксичного аналізу текстів на користь логічного (смислового) аналізу. Утім велике розмаїття текстів, їх тематики та спрямованості, а також достатньо велика розбіжність специфічних понять з різних галузей знань не дозволить у сучасних умовах створити систему штучного інтелекту з такою високорозвинутою семантикою та забезпечити її ефективне функціонування.

На сучасному етапі розвитку інформаційних технологій має науково-практичну значущість розв’язання задачі з удосконалення існуючих систем синтаксичного аналізу текстів на унікальність з метою підвищення достовірності результатів, що отримуються особливо за умови достовірності подання перекладеного або трансформованого тексту.

РОЗВ’ЯЗАННЯ ЗАДАЧІ

На основі виконаного аналізу встановлено, що достовірність визначення частки унікальності текстових фрагментів істотно залежить від форми подання матеріалу на аналіз, а також форми та вмісту бази даних текстових фрагментів, з якими ведеться порівняння. Якщо для порівняння текстових фрагментів система має його трансформувати (наприклад, під час порівняння фрагментів різними мовами), достовірність результату значно зменшується. Отже, можна зробити висновок, що підвищення достовірності аналізу можливе у випадку усунення необхідності попередньої трансформації текстових фрагментів та забезпечення прямого порівняння наданих фрагментів з уміс-том бази даних системи перевірки на плагіат. Такий варіант стає можливим у випадку спеціальної організації бази даних систем перевірки текстів на унікальність. Для підвищення достовірності визначення частки унікальності текстів пропонується використовувати апарат комбінованих систем розпізнавання образів [10–14].

У комбінованих системах розпізнавання об’єкт подається сукупністю образів, ознаки яких однозначно характеризують об’єкт розпізнавання [10, 11]. За кожним з цих образів (або за довільною їх сукупністю) мож-

на віднести об'єкт до одного з наперед визначених класів. У цьому випадку стає можливим отримання коректного рішення щодо класифікації за різних умов спостереження об'єкта розпізнавання, оскільки окремі образи за різних умов спостереження мають різний ступінь інформативності. Відповідно до цього комбінована система для визначених умов спостереження об'єкта розпізнавання здійснює селекцію найбільш інформативних образів для отримання якомога достовірнішого результату класифікації.

Таким чином, відповідно до теорії розпізнавання образів об'єктом розпізнавання є текстовий фрагмент, що надається для аналізу, а фрагменти з якими він порівнюється, є апріорно визначеними класами, ступінь подібності до яких визначатиме частку унікальності тексту, що перевіряється. Кожен з цих класів $C_1 - C_n$ характеризується відповідними образами $P_1 - P_n$. Узагальнену схему порівняння тексту з наявними в базі даних фрагментами з погляду теорії розпізнавання образів показано на рис. 1.

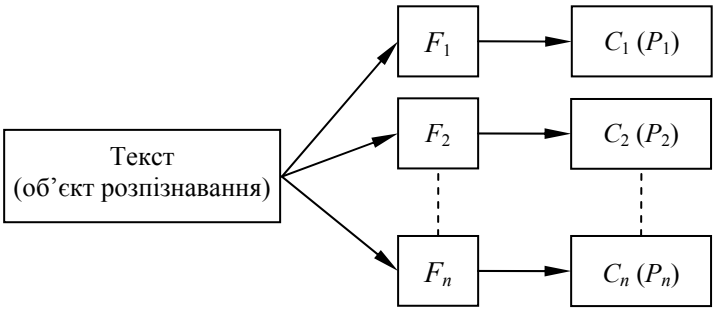

Рис. 1. Узагальнена схема порівняння тексту з наявними в базі даних фрагментами з погляду теорії розпізнавання образів

Схема, показана на рис. 1, повністю відповідає традиційному підходу до побудови систем перевірки на унікальність текстової інформації. Відповідно до неї для врахування можливих трансформацій тексту, що подається для аналізу, в систему вводяться функції перетворення $F_1 - F_n$, які дають змогу порівняти різноманітні форми подання інформації з наявними у базі (з класами $C_1 - C_n$, що характеризуються образами $P_1 - P_n$ відповідно). Наприклад, таке трансформування використовується для порівняння тексту з урахуванням можливих варіантів його перекладу іншими мовами.

Для уникнення трансформування тексту узагальнена схема, показана на рис. 1, перетворюється до варіанта використання комбінованої системи розпізнавання образів (рис. 2).

У комбінованій системі розпізнавання класи $C_1 - C_n$ характеризуються множиною репрезентативних образів $\{P_R\}$, причому за кожним образом з цієї множини можна визначити для наданого для аналізу тексту відповідний клас. Відмінність між традиційною та запропонованою схемами організації бази даних систем перевірки тексту на унікальність проілюстровано на рис. 2.

Відмінності схеми організації порівняння тексту з використанням теорії комбінованих систем розпізнавання образів полягають у тому, що в базі даних кожний клас подається сукупністю репрезентативних образів, які відповідають можливим, апріорно заданим формам подання текстової інформації.

У процесі аналізу фрагмент тексту, що аналізується, порівнюється з кожним репрезентативним образом. Частка унікальності тексту визначається відповідно до міри відстані поданого для аналізу образу від репрезентативних образів. Міра відстані до *i*-го репрезентативного образу визначається як

$$D_i = \left\| P_i - P^R \right\| = \sqrt{(P_i - P^R)'(P_i - P^R)} ,$$

де D_i — характеристика відстані до *i*-го репрезентативного образу; P_i — образ, що надається для аналізу; P^R — репрезентативний образ.

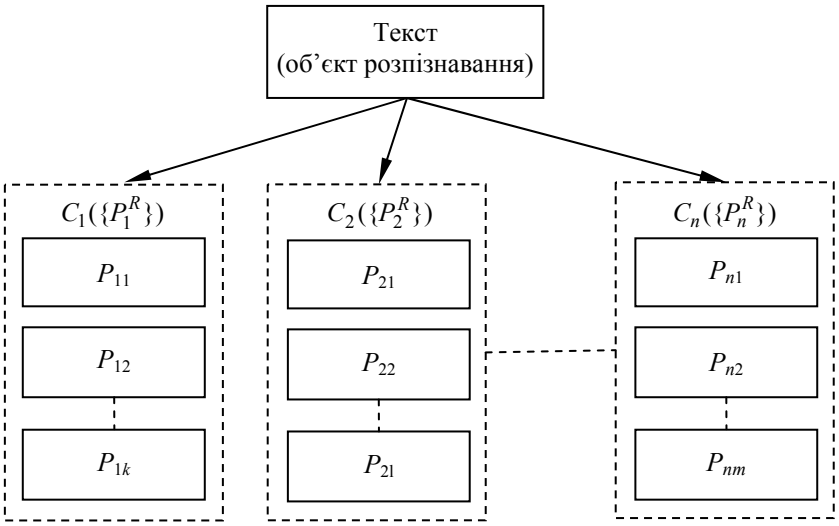

Рис. 2. Узагальнена схема порівняння тексту за наявними в базі даних фрагментами з погляду теорії комбінованих систем розпізнавання образів

Частка унікальності текстового фрагмента визначається пропорційно значенню відстані, тобто

$$D_i \to 0 \Rightarrow y_i \to 0 ,$$

де y_i — частка унікальності текстового фрагмента, %.

Перевірка запропонованого технічного рішення щодо підвищення достовірності визначення частки унікальності текстової інформації проводилася на прикладі двох з чотирьох раніше досліджених систем. На двох системах розробниками не передбачена можливість довільного формування бази даних за запропонованою схемою. Тому довелося повторно виконати досліди з аналізу тексту з очікуваною часткою унікальності 0% за умови його перекладу англійською і російською мовами. Результати дослідів зведено у табл. 4.

Аналіз даних, поданих у табл. 4, вказав на істотне збільшення достовірності визначення частки унікальності тексту. Порівняно з результатами дослідів, поданих у табл. 2 і 3, достовірність збільшилась у середньому на 54%. Це вказує на більш коректне порівняння та наявність позитивного ефекту від усунення попередньої трансформації текстових фрагментів, а також їх аналізу не мовою оригіналу.

Таблиця 4. Результати аналізу текстових фрагментів, перекладених з української мови на англійську і російську та з очікуваною часткою унікальності 0%

Параметр	Система 1	Система 2	Система 1	Система 2
	Англійська мова		Російська мова	
Середнє значення частки унікальності тексту $\overline{y_i}$	6,84	7,12	5,21	6,84
Середня дисперсія значення частки унікальності тексту $\overline{s_i^2}$	5,16	3,07	4,18	2,72
Значення коефіцієнта Кохрена (за довірчого інтервалу 0,95)	0,35	0,31	0,33	0,30

ВИСНОВКИ

У результаті виконаних досліджень отримано такі науково-технічні результати.

1. Установлено, що в системах пошуку плагіату на достовірність визначення частки унікальності тексту істотно впливає форма подання матеріалу.

2. Для підвищення достовірності визначення частки унікальності тексту запропоновано побудову бази даних текстових фрагментів на основі комбінованої системи розпізнавання образів.

3. Використання запропонованих технічних рішень дало змогу підвищити достовірність визначення частки унікальності тексту в середньому на 54%.

4. Незважаючи на запропоновану можливість підвищення достовірності визначення частки унікальності текстів, існує безліч способів штучного збільшення цього показника, а спроби врахувати можливі варіанти трансформування текстів значно ускладнюють базу даних текстових фрагментів. Тому за напрям подальших досліджень можна вибрати створення систем не синтаксичної, а смислової перевірки та експертної оцінки наявності запозиченої інформації. Але це потребує створення високорозвиненої системи штучного інтелекту.

ЛІТЕРАТУРА

1. *Сичивиця О.* Авторство і псевдоавторство в науці. Стаття друга. Плагіат і плагіатори / О. Сичивиця, В.К. Жданкин // Соціогуманітарні проблеми людини. — 2008. — № 3. — С. 39–47.
2. *Ліннік І.* Програмне забезпечення для виявлення плагіату: практичний аспект / І. Ліннік // Науковий блог НаУ «Острозька Академія». — 2013. — Режим доступу: http://naub.oa.edu.ua/2013/prohramne-zabezpechennya-dlya-vyyavlennya-plahiatu-praktychnyj-aspekt
3. *Михайловський Ю.Б.* Система Anti-Plagiarism як інструмент запобігання плагіату в навчальній та науковій діяльності / Ю.Б. Михайловський,

Н.А. Длугунович // Вісн. Хмельн. нац. ун-ту. Технічні науки. — 2013. — № 3. — С. 162–168.

4. *Шарапова Е.В.* Универсальная система проверки текстов на плагиат «Автор.net» / Е.В. Шарапова, Р.В. Шарапов // Информатика и её применения — 2012. — № 3 (6). — С. 52–58.
5. *Hariharan Sh.* Automatic Plagiarism Detection Using Similarity Analysis / Sh. Hariharan // The International Arab Journal of Information Technology. — 2012. — N 4 (9). — P. 322–326. — Available at: http://www.ccis2k.org/iajit/PDF/vol.9,no.4/2796-4.pdf.
6. *Kharat R.* Semantically Detecting Plagiarism for Research Papers / R. Kharat, P.M. Chavan, V. Jadhav, K. Rakibe // International Journal of Engineering Research and Applications (IJERA). — 2013. — N 3 (3). — P. 077–080. — Available at: http://www.ijera.com/papers/Vol3_issue3/P33077080.pdf
7. *Shenoy M.* Automatic Plagiarism Detection Using Similarity Analysis [online] / M. Shenoy, K.C. Shet, U.D. Acharya // Advanced Computing: An International Journal (ACIJ). — 2012. — N 3 (3). — P. 59–62. — Available at: http://airccse.org/journal/acij/papers /0512acij06.pdf
8. *Singh R.* Duplicity Detection System for Digital Documents / R. Singh, C. Dutta// International Journal of Soft Computing and Engineering (IJSCE). — 2012. — N 5 (2). — P. 24–28. — Available at: http://www.iaeng.org/publication/IMECS2011/IMECS2011_pp272-277.pdf
9. *Tschuggnall M.* Detecting Plagiarism in Text Documents through Grammar-Analysis of Authors / M. Tschuggnall, G. Specht // 15th GI-Symposium Database Systems for Business, Technology and Web, 11th March-15th March, 2013. — P. 241–259. — Available at: http://www.btw-2013.de/proceedings/Detecting%20Plagiarism%20in%20Text%20Documents%20through%20GrammarAnalysis%20of% 20Authors.pdf
10. *Румшиский Л.З.* Математическая обработка результатов эксперимента: справочное руководство. — М.: Изд-во «Наука». — 1971. — 192 с.
11. *Довгий С.О.* Системи підтримки прийняття рішень на основі статистичноймовірнісних методів / П.І. Бідюк, О.М. Трофимчук. — К.: Логос, 2014. — 419 с.
12. *Рябенький В.М.* Комбіновані системи розпізнавання образів / В.М. Рябенький, О.І. Захожай // Проблеми інформаційних технологій. — Херсон: ХНТУ. — 2011. — № 01 (009). — С. 156–160.
13. *Захожай О.І.* Екстенсіонально-інтенсіональний підхід до синтезу інформаційних технологій автоматизованої обробки інформації і управління на базі багатопараметричних комбінованих систем розпізнавання образів / О.І. Захожай // Проблеми інформаційних технологій. — Херсон: ХНТУ, 2015. — № 02 (018). — С.106–111.
14. *Меняйленко О.С.* Комбіновані системи розпізнавання образів при аналізі просторового розподілу температури коксового пирога / О.С. Меняйленко, О.І. Захожай // Електротехнічні та комп'ютерні системи. — 2013. — № 12(88). — С. 147–154.

Надійшла 18.07.2017