**УДК 004.56**
**DOI https://doi.org/10.32851/tnv-tech.2021.6.8**

# ВИЗНАЧЕННЯ КОЕФІЦІЄНТУ УНІКАЛЬНОСТІ ТЕКСТОВОГО ДОКУМЕНТУ З ВИКОРИСТАННЯМ КОЕФІЦІЄНТУ ЖАККАРДА

***Савчук Т.О.*** *– PhD,*
*професор кафедри комп'ютерних наук*
*Вінницького національного технічного університету*
*ORCID ID: 0000-0002-0061-6206*
***Кучевський Ю.А.*** *– студент кафедри комп'ютерних наук*
*Вінницького національного технічного університету*
*ORCID ID: 0000-0002-3053-5271*

*Стрімкий розвиток мережі Інтернет, поряд зі зростаючою комп'ютерною грамотністю, сприяє проникненню плагіату в різні сфери людської діяльності: плагіат є гострою проблемою в освіті, промисловості та науковому співтоваристві. Відповідно до [1] під плагіатом розуміють незаконне використання або розпорядження охоронюваними результатами чужого творчої праці, яке супроводжується доведенням до інших осіб неправдивих відомостей про себе як про дійсного автора. Плагіат може бути порушенням авторсько-правового законодавства і патентного законодавства і в якості таких може спричинити за собою юридичну відповідальність. З іншого боку, плагіат можливий і в областях, на які не поширюється дія будь-яких видів інтелектуальної власності, наприклад, в математиці та інших фундаментальних наукових дисциплінах. Плагіат з появою Інтернету перетворився в серйозну проблему. Потрапивши в Інтернет, знання стає надбанням всіх, дотримуватися авторське право стає все важче, а іноді навіть і неможливо. Тому перевірка унікальності серед документів є актуальною задачею. У статті досліджено проаналізовано сучасні методи та засоби перевірки текстової інформації на унікальність. Для кожного з них наведено приклад роботи, переваги та недоліки. Зазначено, що актуальною задачею є підвищення точності при перевірці текстів на унікальність. Ідентифіковано метод шинглів як найбільш поширений та ефективний метод перевірки текстової інформації на плагіат. На базі методу шинглів запропоновано удосконалений алгоритм перевірки текстів на унікальність з використанням коефіцієнту Жаккарда. Було пораховано складність запропонованого алгоритму відносно використання пам'яті та процесорної потужності. Наголошено, що з введенням додаткових покращень швидкодія алгоритму не погіршилась.*

***Ключові слова:*** *плагіат, антиплагіат системи, алгоритм шинглів, коефіцієнт Жаккарда.*

***Savchuk T.O., Kuchevskyi Y.A. Calculation of a coefficient of uniqueness for text document using Jaccard similarity***

*The rapid development of the Internet, along with the growing computer literacy, is contributing to the penetration of plagiarism in various areas of human activity: plagiarism is an acute problem in education, industry and the scientific community. According to [1], plagiarism is understood as the illegal use or disposal of the protected results of another's creative work, which is accompanied by bringing to others wrong information about himself as a real author. Plagiarism can be a violation of copyright and patent law and as such can lead to legal liability. On the other hand, plagiarism is possible in areas that are not covered by any type of intellectual property, such as mathematics and other basic scientific disciplines. Plagiarism with the advent of the Internet has become a serious problem. Once on the Internet, knowledge becomes the property of all, it becomes increasingly difficult and sometimes impossible to enforce copyright. Therefore, checking the uniqueness of documents is the important task. The article analyzes modern methods and means of checking textual information for uniqueness. For each of them there are example of work, advantages and disadvantages provided. It is noted that the important task is to increase the accuracy when verifying texts for uniqueness. The shingles method has been identified as the most common and effective method of plagiarizing textual information. Based on the shingles method, an improved algorithm for checking texts for uniqueness using the Jaccard*

*coefficient is proposed. The complexity of the proposed algorithm in terms of memory and processing power was considered. It is noted that with the introduction of additional improvements, the performance of the algorithm has not deteriorated.*
***Key words:*** *plagiarism, antiplagiarism systems, shingles, Jaccard similarity.*

**Постановка задачі дослідження.** Нехай дано базу даних у якій зберігається множина оброблених документів. Модифікація бази даних та додавання нових документів виконується користувачем системи.

Нехай задано вхідний вектор X(x1, x2), де:

x1 – база документів;

x2 – документ для оцінення на ступінь унікальності.

Тоді, задачу аналізу тексту на унікальність можна подати у вигляді:

$$F(X) = Y,$$

де Y(y1, y2) – вихідний вектор;

y1 – чисельний показник унікальності наданого документу;

y2 – множина документів, які були джерелами плагіату з відповідними чисельними показниками.

Аналіз сучасних антиплагіат методів і засобів. Сучасні підходи перевірки документів на унікальність характеризуються по типу оцінки подібності. Глобальна оцінка використовує великі частини тексту або документа для знаходження подібності в цілому, в той час, як локальні методи на вході перевіряють обмежений сегмент тексту.

Натепер одним із найбільш поширених підходів є дактилоскопія, сутність якої полягає в наступному: з ряду документів вибирається набір з декількох підстрічок, які і є «відбитками». Розглянутий документ буде порівнюватися з «відбитками» для всіх документів колекції. Знайдені відповідності з іншими документами вказують на загальні сегменти тексту [3]. Перевірка документа дослівним перекриттям тексту представляє собою класичне порівняння рядків. Перевірка підозрілих документів в цій ситуації вимагає розрахунку і зберігання ефективно порівнянні подання всіх документів в довідковій колекції, які порівнюються попарно [4]. Як правило, використовують моделі, такі як суфіксне дерево або суфіксний масив, які були адаптовані для виконання цього завдання в контексті комп'ютерного виявлення плагіату. Перевагою цього алгоритму є дуже проста модель реалізації [5]. Однак зіставлення підстрічки є нежиттєздатним рішенням для перевірки великих колекцій документів (алгоритм відпрацьовує в середньому 2h порівнянь, де h – довжина рядка, в якій ведеться пошук), що є недоліком [6].

Аналіз «кошик слів» є спрощеним уявленням, що використовується в обробці природної мови і пошуку інформації [7]. У цій моделі текст представлений як невпорядкований набір слів. Документи представлені у вигляді одного або декількох векторів, які використовуються для попарного обчислення подібності [8]. Приклад реалізації: наступні моделі створюють текстовий документ за допомогою кошику слів. Ось два простих текстових документа:

(1) Джон любить дивитися фільми. Марія теж любить фільми.

(2) Джон також любить дивитися футбольні матчі.

На основі цих двох текстових документів, для кожного документа будується список таким чином:

(1) «Джон», «любить», «дивитися», «фільми», «Марія», «любить», «фільми», «теж».

(2) «Джон», «також», «любить», «дивитися», «футбольні», «матчі».

Кожен список надалі стає об'єктом, де ключем виступає слово з кошику, а значенням – кількість входжень цього ключа в рядок. Порядок елементів при цьому може бути вільний. На практиці модель «кошик слів» використовується в основному як інструмент формування ознак, що є корисним при порівнянні двох фрагментів тексту на унікальність [9; 10]. Перетворивши текст на «кошик слів», можна утворювати різні міри, що характеризують текст. Найбільш поширеним типом характеристик, або ознак, розрахованим за моделлю «кошик слів», є частота термів, а саме, кількість разів, скільки терм з'являється в тексті [11].

Модель «кошик слів» — це не впорядковане представлення документу де важлива лише кількість слів. Наприклад, у наведеному вище прикладі «Джон любить дивитися фільми. Мері теж любить фільми», представлення кошику слів не вкаже не те, що дієслово «любить» завжди йде за ім'ям людини в цьому тексті. Відсутність врахування зв'язку між словами є недоліком наведеного алгоритму. Як альтернатива, n-грамована модель може зберігати цю просторову інформацію. Застосовуючи цей самий приклад, модель bigram буде аналізувати текст на наступні одиниці і зберігатиме термін частоти кожної одиниці, як раніше.

Цитування – комп'ютерний метод виявлення плагіату, призначений для використання в наукових документах, що дозволяє використовувати цитати і довідковий матеріал. Визначає загальні цитати двох наукових робіт.

Стильометрія або вивчення мовних стилів – це статистичний метод для виявлення авторства анонімних документів і для комп'ютерної перевірки на плагіат. Будуються стилометричні моделі для різних фрагментів тексту, уривків, які стилістично відрізняються від інших. І шляхом порівняння моделей можна виявити плагіат. Перевагою методу є знаходження плагіату навіть якщо текст був повністю змінений. Недоліком є велика складність реалізації та необхідність наявності якомога більшої кількості робіт конкретних авторів в навчальній базі задля досягнення допустимої точності в знаходженні оригінального автору за стилем [7].

Алгоритм шинглів призначений для нечіткого пошуку дублікатів тексту. Слово «нечіткий» означає, що входження дублів шукається не точно, а розмито. Наприклад, можливий дублікат не тільки рядків, а й окремих словосполучень. В основному модифікація алгоритму шинглів використовується системами антиплагіату, пошуковими системами для боротьби з пошуковим спамом, копіпастом, а також для визначення унікальності рерайта.

Шингли – виділені для порівняння з тіла тексту окремі фрагменти (підрядка), з певною кількістю слів в його послідовності для перевірки на унікальність. Шингли можуть бути на будь-яку кількість слів, чим шингл коротше, тим точніше буде результат перевірки, але водночас, тим більше обчислювальних ресурсів затрачатиметься на процес перевірки.

Існують різні методи розбиття тексту на шингли:

– один за одним, шингли не перетинаються (рис. 1);

– з перетином, коли підстрічки включають в себе частину попереднього підрядка (рис. 2).

Спосіб формування шинглів і кількість слів або символів в шинглі, а також зрушення шингли (на скільки слів або знаків зсувається наступна підстрічка) сильно впливає на точність результату. При визначенні розмірності підстрічки вибір залежить від обчислювальної потужності, обсягів пам'яті і необхідної точності результатів.

Після поділу підрядка на шингли також існують різні підходи до обчислення контрольних сум і подальшого їх попарного порівняння для оцінки подібності

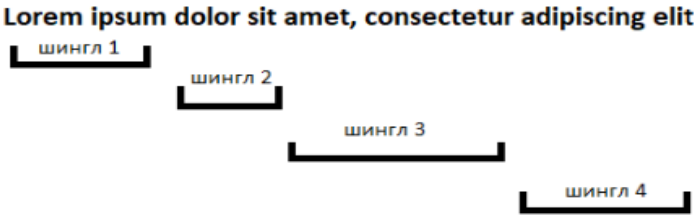


*Рис. 1. Візуалізація розбиття тексту на шингли методом один за одним*

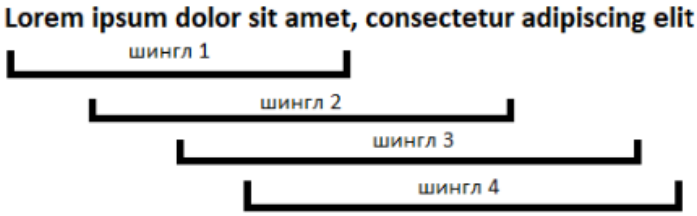


*Рис. 2. Візуалізація розбиття тексту на шингли методом з перетином*

тексту. Контрольні суми можна отримати за допомогою хешування за різними алгоритмами таких як SHA1, SHA3, CRC32, MD5. Далі потрібно оцінити збіг отриманих контрольних сум для двох порівнюваних текстів.

Для ефективного порівняння потрібно задати правильні параметри алгоритму. Чим менше шингл, тим точніше будуть виявлені збігаються слова. Також і із зсувом-менше ймовірності «перестрибнути» повторювані словесні звороти. Однак чим більше текст, тим простіше знайти в ньому збіги якщо вони є, і немає необхідності вибирати мінімальне значення шингли. Важливо зазначити, що більш точна обробка на великому тексті може бути більш повільною.

Таким чином в результаті порівняння наведених алгоритмів перевірки текстів на унікальність було виявлено їх характеристики, що наведені у таблиці 1:

Таблиця 1

**Переваги та недоліки розглянутих методів перевірки текстів на унікальність**

| Характеристика | Дактило-скопія | Кошик слів | Шаблон цитат | Стильо-метрія | Шингли |
|---|---|---|---|---|---|
| Простота реалізації | - | + | - | - | + |
| Висока точність | - | - | + | - | - |
| Великі обсяги документів | - | - | - | + | + |
| Виявлення в зміненому тексті | - | - | - | + | - |
| Робота на невеликій базі | + | + | - | - | + |
| Зв’язок між словами | + | - | + | + | - |

Розглянемо наявні засоби перевірки текстів на унікальність. Система «Антиплагіат» розроблена компанією «Форексис» [9]. Система здійснює онлайн пошук

по великій кількості документів, що зберігаються воснові системи, по базам даних партнерів, в тому числі: наукова електронна бібліотека ELibrary.ru, компанія Lexpro, а також по базі даних користувача. «Антиплагіат» здійснює пошук по мережі Інтернет власними засобами і тому має меншу оперативністю ніж системи, що використовують Яндекс.XML. У безкоштовної версії системи доступна тільки скорочена форма звіту. Перевагою системи є висока точність порівняння та велика база документів та партнерів, та можливість пошуку в Інтернеті. Недоліком є обмеженість в додаванні власної бази документів.

Сервіс Unplag Unplag [10] може здійснювати перевірку на плагіат як в режимі реального часу онлайн, так і порівнювати документ зі збереженою базою документів в бібліотеці користувача. Підтримує роботу з різними типами документів. Перевагою даного засобу є можливість порівняння файлів в Інтернеті та варіація типів підтримуваних документів. Водночас недоліком є менша база документів і, відповідно, менша точність.

Система Plagiat inform перевіряє документи на наявність запозичень як в локальній базі, так і в мережі Інтернет [11]. Перевагою наведеного підходу є те, що система вміє знаходити плагіат у вигляді документів, скомпонованих з «перемішаних» шматків тексту декількох джерел. Недоліками є відсутність перетворення букв та відсутня можливість вільного використання або тестування системи.

Таблиця 2

**Переваги та недоліки засобів перевірки текстів на унікальність**

| Характеристика | Антиплагіат | Unplag | Plagiatinform |
|---|---|---|---|
| Висока точність | + | - | - |
| Додавання власної бази документів | - | - | - |
| Велика база документів та партнерів | + | - | + |
| Варіація типів документів | + | + | + |
| Вільне використання або тестування | + | - | - |
| Перетворення букв | + | + | - |
| Різні режими аналізу | - | + | + |
| Перевірка з багатьох джерел | - | - | + |
| Можливість пошуку в Інтернеті | + | + | + |

Таким чином, актуальною задачею є підвищення точності при перевірці текстів на унікальність.

**Запропоноване рішення.** Нерідко пишуть, що алгоритм шинглів не здатний визначити ідентичність таких фраз, як «Викладач дає студенту матеріал / Викладачі дають студентам матеріали». І дійсно, багато сервісів перевірки унікальності, засновані на алгоритмі шинглів, покажуть, що фрази унікальні, хоча для пошукових систем вони ідентичні. Справа тут не в недоліках алгоритму шинглів, а в методах канонізації тексту, тобто його очищення. Якщо в канонізації використовується морфологія, тобто все слова наводяться до своєї нормальної формі, то алгоритм легко розпізнає фрази як однакові, не залежно від їх закінчень.

Таким чином, доцільним є використання невеликої довжини шинглу для підвищення точності алгоритму. Стандартне використання алгоритму шинглів приводить до збільшення кількості даних, які потрібно обробити, що обмежує його використання [1; 2; 3]. Стандартний вигляд даного алгоритму передбачає складність O(n×m^2), а витрати пам'яті на зберігання текстів, при цьому, будуть складати

O((n×m)/k)], що усувається за рахунок введення етапу попередньої обробки текстової інформації, яка додається в базу даних текстових робіт [4].

Перед обробкою вхідних даних потрібно провести їх канонізацію, за допомогою морфологічних словників, коли кожне слово приводиться до нормальної форми. Для усунення впливу слів, які не мають впливу на процес перевірки на унікальність, потрібно прибрати сполучники та пунктуаційні знаки, що покращить точність перевірки тексту на унікальність.

Для досягнення найбільшої точності алгоритму візьмемо довжину шингла за 1.

Надалі кожен хешований фрагмент неоднорідності (шингл) додається в базу даних типу «ключ-значення». Ключем виступає сам фрагмент, а значенням – множина ідентифікаторів документів, які містять даний фрагмент. Враховуючи те, що бази типу "ключ-значення" гарантують доступ до кортежу за O(1), то складність алгоритму становитиме O(m) порівняно з O(n×m^2).

Для обчислення наочного коефіцієнту унікальності текстового документу використаємо коефіцієнт Жаккарда для кожної пари вхідного документу і документу, що має найбільшку кількість однакових шинглів шляхом виконання наступних обчислень:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Удосконалений алгоритм можна подати наступним чином:

1. Розбиття вхідних даних на шингли довжиною 1.
2. Канонізація даний за допомогою морфологічних словників.
3. Видалення допоміжних слів таких як сполучники, прийменники тощо.
4. Для кожного шинглу виконання запиту для отримання множини ідентифікаторів документів що містять такий самий шингл. Водночас для кожного отриманого ідентифікатора оновлюємо довідник в пам'яті, інкрементуючи значення змінної під ключем ідентифікатора
5. Сортування пар ключ-значення в порядку спадання за значенням.
6. Для перших k пар дістаємо з бази даних текстові документи з відповідними ідентифікаторами.
7. Для кожного документу з бази даних обраховуємо коефіцієнт Жаккарда використовуючи шингли вхідного документу та того, що був отриманий з бази даних.
8. Для відображення наочного коефіцієнту використовуємо коефіцієнт Жаккарда як показник рівня плагіату відносно певного існуючого документу і 1-jaccard як коефіцієнт унікальності.

Запропонований удосконалений алгоритм перевірки текстової інформації на унікальність базується на використанні коефіцієнту Жаккарда, що підвищує точність алгоритму. Також, зберігання фрагментів неоднорідності в базі даних типу «ключ – значення» забезпечує швидку обробку великих обсягів текстової інформації. Так як кількість остаточних документів для обчислення фінального коефіцієнту є малою, то є доцільним обрахунок цього коефіцієнту на різних довжинах шинглів, що також підвищує точність.

Доцільність використання удосконаленого алгоритму перевірки текстів на унікальність було підтверджено детальним обчисленням складності алгоритму та подальшим порівнянням зі складністю стандартного алгоритму. Результати проведених досліджень щодо залежності трудомісткості операцій по перевірці текстів на унікальність від обсягу пам'яті яка необхідна для зберігання фрагментів неоднорідності наведені в таблиці 3.

Таблиця 3

**Порівняльна характеристика стандартного та удосконаленого алгоритму**

| К-ть шинглів | К-ть текстів | К-ть операцій (станд.) | К-ть операцій (удоск.) | Затрачена пам'ять (станд.) | Затрачена пам'ять (удоск.) |
|---|---|---|---|---|---|
| 10 | 10 | 1 000 | 10 | 100 | 100 |
| 10 | 50 | 25 000 | 10 | 500 | 500 |
| 50 | 10 | 5 000 | 50 | 500 | 500 |
| 50 | 50 | 125 000 | 50 | 2 500 | 2 500 |
| 100 | 100 | 1 000 000 | 100 | 10 000 | 10 000 |
| 100 | 200 | 4 000 000 | 100 | 20 000 | 20 000 |
| 200 | 200 | 8 000 000 | 200 | 40 000 | 40 000 |

Таким чином, використання коефіцієнту Жаккарда дає більш точну оцінку під час обчислення унікальності текстового документу. Водночас потреби в пам'яті та процесорній потужності залишились незмінними.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Савчук Т.О., Кучевський Ю.А. Удосконалений алгоритм перевірки текстів на унікальність. *INTERNET-EDUCATION-SCIENCE* : Proceedings of the XII International scientific-practical conference, м. Вінниця, 26–29 травня 2020 р. Вінниця : Вінницький національний технічний університет, 2020. С. 237–239. URL: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/30970/WORK-IES-2020-269-271.pdf?sequence=1&isAllowed=y.

2. Савчук Т.О., Кучевський Ю.А. Підхід до аналізу на унікальність курсових розробок. *Матеріали XLIX науково-технічної конференції підрозділів Вінницького національного технічного університету*, м. Вінниця, 27–28 квітня 2020 р. URL: https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2020/paper/view/8929/7739.

3. Monostori K., Zaslavsky A., Schmidt H. Document Overlap Detection System for Distributed Digital Libraries. Proceedings of the fifth ACM conference on Digital libraries. 2000. P. 226–227.

4. Leong A., Lau H., Rynson W. H. Check: A Document Plagiarisment Detection System. *Proceedings of ACM Symposium for Applied Computing.* 1997. P. 70–77.

5. Dreher H. Automatic Conceptual Analysis for Plagiarism Detection. *The Journal of Issues in Informing Science and Information Technology.* 2007. Vol. 4. P. 601–614.

6. Meyer zu Eissen S., Stein B. Intrinsic Plagiarism Detection. *European Conference on Information Retrieval*. Springer, 2006. P. 565–569.

7. Седов А.В., Рогов А.А. Анализ неоднородностей в тексте на основе последовательностей частей речи. *Современные проблемы науки и образования.* 2013. № 1. URL: https://science-education.ru/ru/article/view?id=8339.

8. Антиплагиат: обнаружение заиствований. Веб-сайт. URL: https://www.antiplagiat.ru/corporate/education.

9. Unichek. Сервіс перевірки на плагіат для найкращих результатів. Вебсайт. URL: https://unicheck.com/uk-ua.

10. Ширяев М.А., Мустакимов В. Plagiatinform избавит от плагиата в научных работах. *Educational Technology & Society.* 2008. № 11 (1). С. 367–374. URL: https://cyberleninka.ru/article/n/plagiatinform-izbavit-ot-plagiata-v-nauchnyh-rabotah/viewer.

11. Brin S., Davis J., Garcia-Molina H. Copy Detection Mechanisms for Digital Documents. *CM International Conference on Management of Data* (SIGMOD 1995), San Jose, California, May 22–25, 1995. P. 398–409.

**REFERENCES:**

1. Savchuk, T.O., Kuchevskyi, Y.A. (2020) An improved algorithm for checking texts for uniqueness. URL: http://ies.vntu.edu.ua/reports/program/WORK-IES-2020.pdf.

2. Savchuk, T.O., Kuchevskyi, Y.A. (2020) An approach to analysis of course works for uniqueness. URL: https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2020/paper/view/8929/7739.

3. Monostori, K., Zaslavsky, A., Schmidt, H. (2000) Document Overlap Detection System for Distributed Digital Libraries.

4. Leong, A., Lau, H., Rynson, W.H. (1997) Check: A Document Plagiarisment Detection System.

5. Dreher, H. (2007) Automatic Conceptual Analysis for Plagiarism Detection. Information and Beyond: The Journal of Issues in Informing Science and Information Technology.

6. Meyer zu Eissen, S., Stein, B. (2006) Intrinsic Plagiarism Detection. Springer.

7. Sedov, A.V., Rogov, A.A. (2013) Analysis of indifferences in text basing on words sequence. Modern problems of science and education. Vol. 1.

8. Antiplagiarism for rising the level of education. URL: https://www.antiplagiat.ru/corporate/education.

9. Plagiarism Detection in Schools & Universitie. URL: https://unicheck.com/plagiarism-check-for-k-12-and-higher-education.

10. Plagiatinform as antiplagiarism system in science papers. URL: https://cyberleninka.ru/article/n/plagiatinform-izbavit-ot-plagiata-v-nauchnyh-rabotah/viewer.

11. Brin, S., Davis, J., Garcia-Molina, H (2001). Copy Detection Mechanisms for DigitalDocuments.