УДК 520.6.05

Магдич Б.В.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

ЗАДАЧА СТВОРЕННЯ АВТОМАТИЗОВАНОЇ СИСТЕМИ ДЛЯ ПЕРЕВІРКИ СТУДЕНТСЬКИХ РОБІТ НА ПЛАГІАТ ДЛЯ ЗАДАНОЇ ПРЕДМЕТНОЇ ОБЛАСТІ

У роботі наведений огляд базових алгоритмів визначення відсотка унікальності заданого тексту щодо іншого. Також запропоновані модифікації в предметній області, що дозволяють детальніше визначити унікальність тексту.

Ключові слова: *плагіат, алгоритм шинглів, дублікат, копія, авторські права.*

Постановка проблеми. Стрімкий розвиток мережі Інтернет поряд зі зростанням комп'ютерної грамотності широкої аудиторії користувачів стали причиною поширення плагіату в різних сферах людської діяльності, зокрема в галузі освіти та науки. В освіті проблема плагіату досить гостра й активно обговорюється. На жаль, студенти сьогодення не проводять багато часу в бібліотеках і архівах через те, що більшість інформації можна «скачати» із джерел мережі Інтернет або запозичити роботи студентів минулих років навчання і під час здачі матеріалу викладачеві банально замінити прізвище на своє, тобто порушується авторське право на інтелектуальну власність. Тому для викладачів досить актуальною є проблема виявлення плагіату в роботах студентів.

На жаль, викладач не може перевірити виконання завдання до лабораторної роботи чи реферату на унікальність за короткий проміжок часу. Щоб оцінити роботу студентів, спочатку необхідно порівняти роботи за однаковим завданням – знайти в них подібності, і тільки після цього перевіряти саму роботу. Для збільшення продуктивності та швидкості порівняння робіт і пошуку плагіату розробляють спеціальні системи, що могли б максимально автоматизувати процес. Бази даних таких систем обов'язково мають поповнюватися вже виконаними роботами кожного семестру чи року, щоб студенти не мали змоги користуватися запозиченнями з робіт студентів за попередні періоди.

Призначення такої системи полягає у визначенні авторства роботи студента, забезпеченні пошуку плагіату серед наявних робіт студентів і накопиченні робіт у базі даних системи. Пошук засновується на алгоритмах – детекторах тексту.

Цілями створення комп'ютеризованої системи є:

– отримання оціночної інформації про роботу й узагальненого коефіцієнта плагіату в роботі;

– підвищення ефективності оцінювання роботи студентів.

Аналіз алгоритмів для виявлення плагіату в текстових документах. Якщо говорити про методи виявлення плагіату в довільних текстах, то часто замість слова «плагіат» вживають термін «нечіткий дублікат». Ці методи можна розділити на два великі класи. Алгоритми, які використовують певні знання про всю розглянуту колекцію документів, називають глобальними, в іншому разі – локальними [4].

Спочатку розглянемо локальні алгоритми. Основна ідея таких методів зводиться до синтаксичного аналізу документа. На основі цього аналізу визначається відповідність документа певній кількості сигнатур.

LongSent

Найпростішим прикладом може бути алгоритм, який обчислює хеш-функцію (MD5, SHA-2, CRC32) від конкатенації двох найдовших пропозицій у документі. Це і буде його сигнатурою. Точність такого алгоритму досить велика, але він має істотні вади щодо безпеки. Такий алгоритм легко обдурити. Досить відкорегувати лише два найдовших речення.

Методи на основі міри TF

Більш ефективним способом знаходження нечітких дублікатів може стати метод, заснований на TF (term frequency – частота слова). TF – це відношення числа вживання конкретного слова до загальної кількості слів у документі. Так оцінюється важливість слова в межах окремого доку-

мента. Для кожного слова в документі обчислюється його вага, що дорівнює відношенню числа входження цього слова до загальної кількості слів документа. Далі зчіплюються n упорядкованих слів із найбільшим значенням ваги і обчислюється хеш-функція. Такий підхід дозволяє поліпшити ситуацію, але для вирішення реальних завдань цей спосіб не підходить.

Методи, які використовують семантичні мережі

Також цікавим підходом є використання семантичної мережі. Завдання визначення факту запозичення зводиться до порівняння моделей, що відображають смислове навантаження текстів. Аналіз ведеться з використанням алгоритмів на графах, модифікованих і оптимізованих для застосування в межах даного завдання. Використання схем аналізу даних у цьому методі дозволяє виявляти факт запозичення, навіть якщо оригінал був певним чином модифікований (виконаний переклад, слова були замінені на синоніми, текст був викладений із використанням іншої лексики тощо).

Методи, які використовують поняття шинглів

Один із перших методів, який був застосований на практиці (компанією AltaVista), ґрунтувався на понятті шинглів. Даний підхід був запропонований A. Broder [1].

Сьогодні він є найпопулярнішим алгоритмом для пошуку плагіату в довільних текстах. Він розроблений для пошуку копій (дублікатів) розглянутого тексту в документі та є потужним інструментом, що може боротися із проявами плагіату.

Метод заснований на представленні текстів у вигляді множини послідовностей фіксованої довжини, що складаються із сусідніх слів. За значного перетину таких множин документи будуть схожі один на одного.

Розберемо, через які етапи проходить текст [2], що буде порівнюватися:

1. Канонізація тексту і видалення «стоп-символів» і «стоп-слів».
2. Розбиття на шингли.
3. Обчислення хешів шинглів.

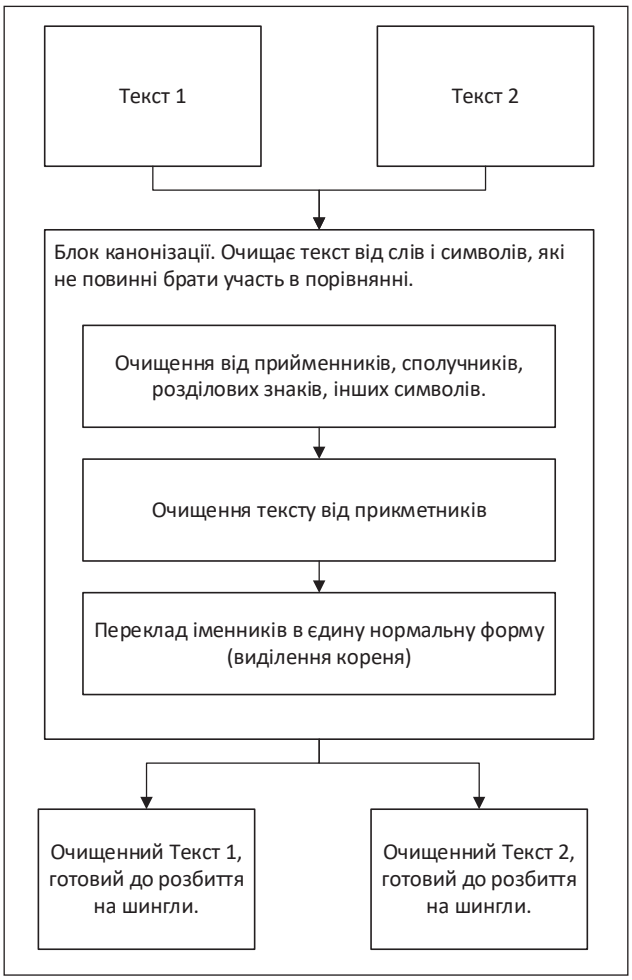

Рис. 1. Послідовність робіт етапу канонізації в алгоритмі шинглів

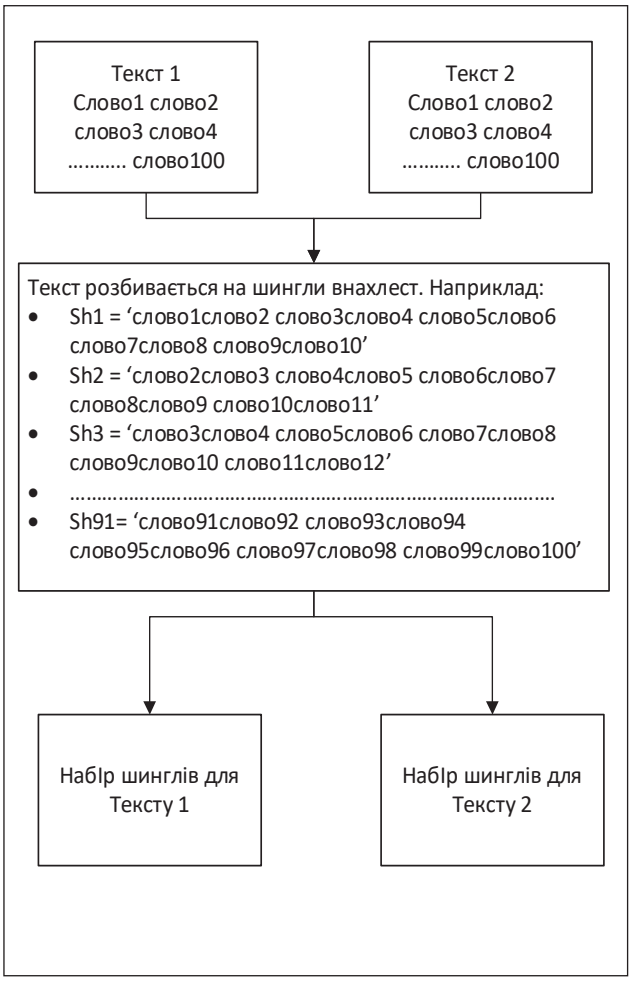

Рис. 2. Послідовність робіт етапу розбиття на шингли

4. Порівняння та визначення результату.

Розглянемо даний алгоритм детальніше:

Канонізація тексту

Канонізація тексту приводить оригінальний текст до єдиної нормальної форми. Текст очищається від «стоп-символів» і «стоп-слів» (прийменників, сполучників, знаків пунктуації тощо), які не повиненні брати участь у порівнянні. Іноді також пропонується видаляти з тексту прикметники, бо вони не мають смислового навантаження. Також на етапі канонізації тексту можна приводити іменники до називного відмінку, єдиного числа або залишати від них тільки корінь. На виході цього етапу ми маємо текст, очищений від «сміття» і готовий до порівняння (рис. 1).

Розбиття на шингли

Шингли – виділені підпослідовності слів. Необхідно з порівнюваних текстів виділити підпослідовності слів, що йдуть один за одним по N штук (довжина шингла). Розбиваючи у такий спосіб текст на підпослідовності, ми отримаємо набір шинглів. Дії за кожним із пунктів виконуються для кожного з порівнюваних текстів (рис. 2).

Обчислення хешів шинглів

Принцип алгоритму шинглів полягає в порівнянні випадкової вибірки контрольних сум шинглів (підпослідовностей) двох текстів між собою. Тепер у кожного з текстів є свої набори шинглів. Треба розрахувати контрольну суму кожного із шинглів. Для розрахунку можна використовувати відомий алгоритм CRC3 2 (рис. 3).

Порівняння та визначення результату

На цьому етапі порівнюємо між собою всі елементи першого масиву з відповідними елементами другого масиву, зчитуємо відношення однакових значень і отримуємо кінцевий результат (рис. 4).

Специфіка предметної області, що дозволяє модифікувати наявні алгоритми для покращення показників виявлення плагіату

Предметна область, в якій застосовується один із вищезазначених алгоритмів, має свою специфіку, що відкриває великий простір для модифікацій [3]. Під специфікою мається на увазі, що кожна область буде мати досить багато сталих виразів, словосполучень, слів. Наприклад, якщо взяти за область застосування тексти з історії, то це будуть різні дати, назви подій, прізвища йімена популярних в історії людей тощо. Якщо взяти за область математику, то нам необхідно звернути увагу на формули, назви геометричних фігур (квадрат, трикутник тощо), назви теорем, різні правила математики. Тому необхідно вилучати такі слова і словосполучення з тексту, щоб він мав вигляд, за якого порівняння з іншим текстом буде максимально якісне.

Під час застосування алгоритмів необхідно враховувати задані специфікації предметної області, оскільки вони впливають на відсоток унікальності. А нам необхідно максимально чітко визначити унікальність роботи студента і дізнатися, запозичив він текст чи ні.

Модифікації, що сприяють виявленню плагіату в заданій предметній області

Для більш ефективного виявлення плагіату в заданій предметній області необхідно враховувати її специфіку. Тобто треба з тесту вилучати

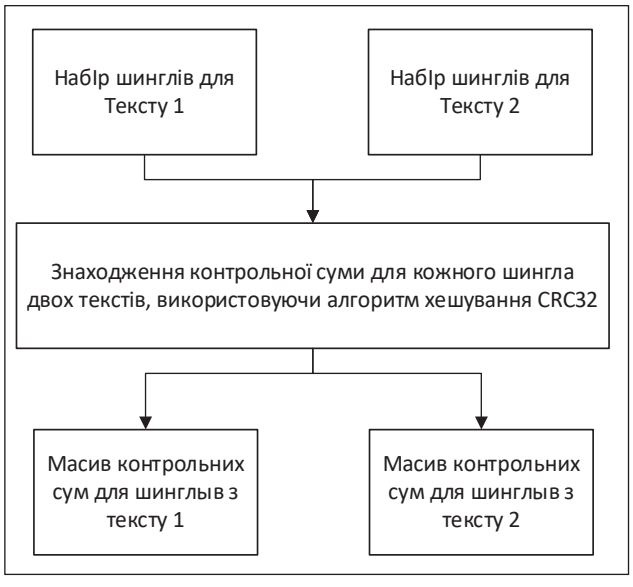

Рис. 3. Послідовність робіт під час знаходження контрольних сум шинглів

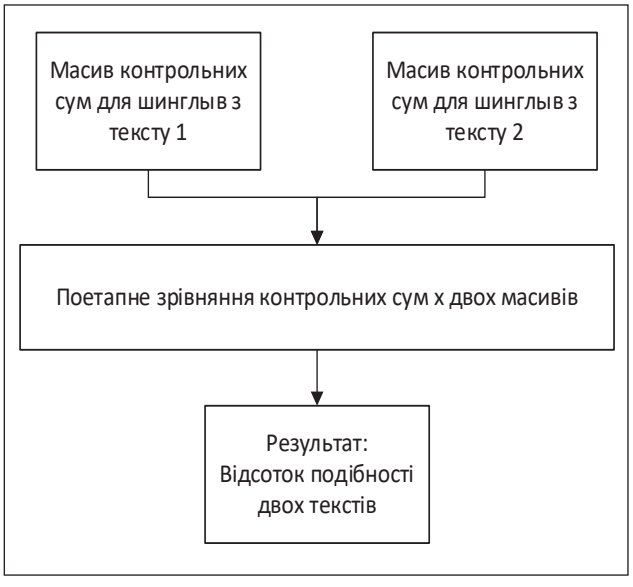

Рис. 4. Послідовність під час порівняння елементів масиву

всі сталі слова та словосполучення. Також необхідно враховувати те, що слова і словосполучення мають синоніми, що теж впливає на роботу алгоритму.

Алгоритми для виявлення плагіату не розрізняють слова на часто вживані та нечасто вживані, вони будуть включати всі слова та словосполучення в оброблення, і в разі повтору у двох текстах однакових слів сприймають це за плагіат. Наприклад, візьмемо словосполучення «Друга світова війна», яке досить часто вживається в рефератах зі всесвітньої історії, це словосполучення має також синонім «Велика вітчизняна війна», якщо один студент використав перше словосполучення, а другий студент − інше словосполучення, то під час оброблення обох рефератів алгоритм буде зчитувати та видавати подібність у слові «війна», хоча насправді його не потрібно взагалі враховувати, оскільки воно є часто вживаним у різних історичних документах.

Як можна побачити, такі модифікації дадуть нам змогу максимально почистити текст від усього, що не потрібно для аналізу. На виході ми будемо мати очищений текст саме в нашій області застосування, який вже можна буде, за допомогою того чи іншого алгоритму, обробляти та порівнювати з іншим текстом.

Модифікована автоматизована система виявлення плагіату

Отже, ми отримаємо модифікований алгоритм, що дозволить нам досить коректно виявляти плагіат у заданій галузі. Проте необхідно створити саме автоматизовану систему на базі цього модифікованого алгоритму, така система буде мати декілька цінних переваг, а саме: перевірка роботи на унікальність, накопичення робіт і можливість накопичувати сталі словосполучення та слова. Необхідно створити базу знань, в якій можна зберігати унікальні роботи (зазвичай відсоток унікальності – понад 80%), сталі слова, словосполучення та їх синоніми. Цінність такої бази знань буде збільшуватися з кожним циклом роботи, тому що вона буде поповнюватися. Більшість популярних платформ, що перевіряють роботи на плагіат, працюють саме за таким принципом.

Така система не дасть можливості студентам використовувати наробітки попередніх поколінь. А викладач буде знати чіткий процент того, наскільки робота була запозичена чи вона оригінальна, а в разі виявлення оригінальності роботи зможе поповнити базу знань. Надалі викладач матиме змогу ділитися своєю базою знань з іншими викладачами, так можна передавати свої надбання з того чи іншого предмета новим викладачам, що дозволить їм перевірити роботу на унікальність і правильно оцінити знання студентів.

Висновки. Впровадження такої системи на рівні університету дозволить викладачам швидко й якісно виявляти запозичення в працях студентів. Що, у свою чергу, сприятиме самостійному виконанню студентами своїх робіт. Якість знань у студентів буде збільшуватися, тому що будуть збільшуватися вимоги до робіт у викладачів.

Також дана система має великі перспективи для подальших модифікацій. Наприклад, аналіз роботи на плагіат із використанням засобів пошукових платформ (пошук в Інтернеті), тому що досить часто студенти саме звідти беруть запозичену інформацію, з порушенням авторських прав чи без. Надалі можливе розроблення і впровадження алгоритмів аналізу картинок, схем, кодів програм тощо. Також можна створити бази знань для різних предметних областей, що дозволить застосовувати автоматизовану систему не тільки в межах однієї навчальної дисципліни, але й у всьому навчальному просторі навчального закладу.

Список літератури:

1. Broder M. Charikar et al. Min-wise independent permutations, Proceedings of the thirtieth annual ACM symposium on Theory of computing. 1998.

2. W-Shinging. URL: https://en.wikipedia.org/wiki/W-shingling.

3. Moss. A. System for Detecting Software Plagiarism. URL: https://theory.stanford.edu/~aiken/moss/.

4. Miller C. Detecting duplicates: a searcher's dream come true. URL: https://www.questia.com/magazine/1G1-9065495/detecting-duplicates-a-searcher-s-dream-come-true.

ЗАДАЧА СОЗДАНИЯ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ДЛЯ ПРОВЕРКИ СТУДЕНЧЕСКИХ РАБОТ НА ПЛАГИАТ ДЛЯ ЗАДАННОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

В работе приведен обзор базовых алгоритмов для определения процента уникальности заданного текста относительно другого. Также определены модификации в предметной области, позволяющие подробнее определить уникальность текста.

Ключевые слова: *плагиат, алгоритм шинглов, дубликат, копия, авторские права.*

TASK OF THE AUTOMATION SYSTEM CREATION FOR VERIFICATION STUDENT WORKS FOR PLAGIARISM FOR A GIVEN SUBJECT AREA

This paper is the review of basic algorithms by which is possible to determine the percentage of a given text unique compared to another. The modifications in a subject area is describe, enabling more acceptable text to determine uniqueness.

Key words: *plagiarism, algorithm Shingle, duplicate, copy, copyright protection.*