УДК 004.02

В. Е. Климнюк

Харківський національний економічний університет імені Семена Кузнеця, Харків

ПРОБЛЕМИ СТВОРЕННЯ КАФЕДРАЛЬНОЇ АНТИПЛАГІАТНОЇ СИСТЕМИ

У роботі розглядається можливість побудови кафедральної системи перевірки на плагіат студентських робіт . Запропоновано новий підхід перевірки документів на основі нечітких дублікатів за допомогою алгоритму шинглів і з урахуванням шаблонів. Проведено дослідження з визначення оптимальної кількості шинглів для перевірки унікальності коротких документів , що характерно для студентських робіт. Показано, що з ростом числа шинглів для коротких документів зростання помилок стає істотним. Запропоновано технологію створення кафедральної системи перевірки студентських робіт у вигляді сайту

Ключові слова: *шингл, нечіткі дублікати, плагіат.*

Постановка проблеми

Для того щоб стати справжнім фахівцем, необхідно не тільки діставати інформацію, але і вміти самостійно її застосовувати для виконання лабораторних, курсових, дипломних та наукових робіт. Але на жаль, у зв'язку з доступністю інформації через Інтернет або іншим способом, плагіат став буденною справою. У передових ВНЗ в обов'язковому порядку проходять перевірку на плагіат всі наукові та навчально-методичні матеріали, а також дипломні роботи за допомогою спеціальних зовнішніх програм. Однак не всі студентські роботи, особливо поточні, потрапляють в Інтернет і проходять таку перевірку. Вони часто залишаються всередині університету, де переходять з рук в руки, з року в рік новим студентам. Оскільки більшість програм, які призначені для перевірки робіт на унікальність, користуються базою Інтернету, то виникає необхідність створення кафедральної системи перевірки унікальності студентських робіт.

Мета проведення досліджень

Щоб створити таку систему необхідно ураховувати велику кількість поточних студентських робіт та їх особливості – наявність деяких шаблонів, за якими вони виконуються і порівняно невеликий за обсягом їх об’єм – звіти з лабораторних та практичних робіт, есе, тези та наукові статті, курсові та дипломні роботи.

Крім того необхідно вирішити проблему виявлення нечітких дублікатів [1], як одну з найбільш важливих і важких завдань аналізу даних і пошуку інформації в Інтернеті. Актуальність цієї проблеми визначається різноманітністю додатків, у яких необхідно враховувати «схожість», наприклад, текстових документів. Основною перешкодою для успішного вирішення даної задачі є гігантський обсяг даних, що зберігаються в базах сучасних пошукових машин. Такий обсяг робить практично неможливим її «пряме» рішення шляхом попарного порівняння текстів документів. Тому останнім часом велика увага приділяється розробці методів зниження обчислювальної складності створення евристичних алгоритмів (хешування певного фіксованого набору «значущих» слів або речень документу, семплювання набору підрядків тексту, використання дактилограм та ін.) І, нарешті, ще одним ключовим вимогою, що пред'являються до якості алгоритмів детектування нечітких дублікатів, є їх стійкість до «невеликих» змін вихідних документів і можливість впевнено обробляти короткі документи.

Аналіз методів перевірки текстів

Головною проблемою, при створенні антиплагіатної системи модулю, є необхідність вибору певного методу перевірки тексту на унікальність. Існує багато методів, більшість з яких постійно розвивається зі своїми перевагами і недоліками. Серед них можна виділити «синтаксичні» і «лексичні» методи.

До «синтаксичних» методів можна віднести методи шинглів, методи Long Sent, Heavy Sent та багато інших.

У 1997 році А. Бродер запропонував «синтаксичний» метод оцінки подібності між документами, заснований на представленні документа у вигляді безлічі всіляких послідовностей фіксованої довжини, які складаються з сусідніх слів [3]. Такі послідовності були названі «шингли». Два документи вважалися схожими, якщо їх безлічі шинглів істотно перетиналися. Оскільки число шинглів приблизно дорівнює довжині документа в словах, тобто є достатньо великим, автором були запропоновані два методи семплювання для отримання репрезентативних підмножин. Перший метод залишав тільки ті шингли, чиї «дактилограми», які обчислюють за

алгоритмом Карпа-Рабіна, ділилися без залишку на деяке число m. Основний недолік – залежність вибірки від довжини документа і тому документи невеликого розміру представлялися або дуже короткими вибірками, або взагалі не мали таких. Другий метод відбирав тільки фіксоване число s шинглів з найменшими значеннями дактилограм або залишав всі шингли, якщо їх загальна кількість не перевищувала s.

Переваги методу шинглів:

можливість зміни кількості шинглів

більш точна перевірка текстів різного об'єму

швидкість роботи

Long Sent. Документ розбивається на речення, які упорядковуються за зменшенням довжини, вираженої кількістю слів, а при рівності довжин - у алфавітному порядку. Потім вибираються і зчіплюються в рядок в алфавітному порядку 2 самі довгі речення. В якості сигнатури документа обчислюється контрольна сума отриманого рядка.

Недоліки метода – низька швидкість перевірки та низький процент плагіату при перевірці великого тексту.

Інший сигнатурний підхід, заснований вже не на синтаксичних, а на лексичних принципах, був запропонований А. Коудури в 2002 р. і вдосконалений в 2004 р. [5]. Основна ідея такого підходу полягає в обчисленні дактилограми I-Match для подання змісту документів на основі спеціально побудованого словника. Два документи вважаються схожими, якщо у них збігаються I-Match сигнатури. Алгоритм має більш високу обчислювальну ефективність, ніж алгоритм А. Бродера. Іншою перевагою алгоритму є його висока ефективність при порівнянні невеликих за розміром документів. Основний недолік – нестійкість до невеликих змін змісту документа. Цей недолік було зменшено за рахунок введення можливості багаторазового випадкового перемішування основного словника і додавання додаткових словників, одержуваних шляхом випадкового видалення з початкового словника деякого невеликої фіксованої частини слів. Алгоритм показав високі результати при використанні в різних додатках веб-пошуку і фільтрації спаму.

Найбільш відомим представником «лексичних» методів є метод TF * RIDF – основна ідея якого полягає в побуддові словник, що ставить кожному слову у відповідність число документів, в яких воно зустрічається. Потім будується частотний словник документа і для кожного слова обчислюється його «вага».

Недоліки метода – в сигнатуру не включаються рідкісні слова та низький процент плагіату при перевірці великого тексту.

З «синтаксичних» (що використовують послідовності слів) методів найбільший інтерес представляє метод Log_Shingle, який забезпечує найвищу точність і швидкість роботи.

Основні етапи, які проходить текст, що піддається порівнянню:

1. Канонізація тексту.
2. Розбиття на шингли.
3. Обчислення хешей шинглів за допомогою 84х статичних функцій.
4. Випадкова вибірка 84 значень контрольних сум.
5. Порівняння, визначення результату.

Канонізація тексту призводить оригінальний текст до єдиної нормальної форми. Текст очищається від прийменників, прикметників, сполучників, розділових знаків, HTML тегів, який не повинен брати участь в порівнянні. Далі іменників переводяться в єдину нормальну форму –називний відмінок, однина або залишати від них лише корінь.

Розбиття на шингли. Необхідно з порівнюваних текстів виділити під послідовності слів, що йдуть один за одним по 10 штук (довжина шинглу). Вибірка відбувається внахлест, а не встик. Таким чином, розбиваючи текст на підпослідовності, можна отримати набір шинглів в кількості рівній кількості слів мінус довжина шинглу плюс один.

При обчисленні хешей шинглів розраховуються контрольні суми шинглів текстів.

Для кожного ланцюжка обчислюються 84 дактілограмми за алгоритмом Карпа Рабіна за допомогою взаємно-однозначних і незалежних функцій, що використовують випадкові набори простих поліномів. У результаті кожен документ представлявся 84 шинглами, мінімізуючими значення відповідної функції. Потім 84 шингли разбіваются на 6 груп по 14 незалежних шинглів у кожній. Ці групи називаються «супершинглами». Таким чином, кожен документ представляється всілякими попарними поєднаннями з 6 супершинглів. Принцип алгоритму шинглів полягає в порівнянні випадкової вибірки контрольних сум шинглів текстів між собою.

Якщо два документи мають схожість, наприклад, 95%, то 2 відповідних супершинглів в них збігаються з ймовірністю 0,49. Оскільки кожен документ представляється 6 супершинглів, то можна показати, що ймовірність того, що у двох документів співпадають не менше 2-х супершинглів, дорівнює0.90. Для порівняння: якщо два документи мають схожість тільки 80%, то ймовірність збігу не менше двох супершинглів складає всього 0,026 [4].

Проблема алгоритму полягає в кількості порівнянь, адже це безпосередньо позначається на продуктивності. Збільшення кількості шинглів для порівняння характеризується ростом операцій, що критично відіб'ється на продуктивності. Для збільшення продуктивності виконується випадкова вибірка контрольних сум.

Ключова перевага даного алгоритму полягає в тому, що, по-перше, будь-який документ (у тому числі і дуже маленький) завжди представляється вектором фіксованої довжини, і, по-друге, подібність визначається простим порівнянням координат вектора і не вимагає виконання теоретико-множинних операцій.

Існує безліч програм та ресурсів, котрі запобігають виникненню плагіату, і які б могли стати основою для розробки антиплагіатної системи кафедри.

Головним недоліком ресурсів (сайтів) є перевірка тексту певної довжини (від 500 до 3000 символів), що виключає можливість перевірки наукових робіт або звичайних робіт студентів. Іншим суттєвим недоліком є те, що за один день можна робити лише 10 перевірок.

Програма etxt антиплагіат є однією з найбільш популярних та якісних програм (серед програм у вільному доступі), призначених для перевірки унікальності тексту. Основною перевагою є точність перевірки, можливість завантаження роботи як за допомогою звичайного копіювання так і завдяки вибору файлу на комп'ютері. Серед недоліків є швидкість роботи програми та постійне втручання перевірки відвідувача за допомогою CAPTCHA, що знижує юзабіліті.

Основним і загальним недоліком вказаних ресурсів є те, що вони не використовують шаблонний метод перевірки і не можуть гарантувати точний результат аналізу тексту.

Це пов’язано з тим, що при оцінки студентських робіт необхідно враховувати правила оформлення робіт, можлива наявність спільної частини тексту для декількох студентських робіт, наприклад спільної титульної сторінки, назва, тема і мета лабораторних робіт тощо. Щоб підвищити вірогідність оцінки унікальності тексту необхідно створити шаблон, який зможе оцінити роботу студента без урахування основних елементів і результати будуть більш об'єктивні.

Окрім того слід враховувати того, що студенти можуть використовувати автозаміну літер кирилиці на схожі англійські літери.

Використовуючи алгоритм шинглів слід враховувати принцип перевірки тексту, котрий полягає в порівнянні випадкової вибірки контрольних сум шинглів текстів між собою. Збільшення кількості шинглів для порівняння характеризується ростом операцій, що критично відіб'ється на продуктивності. Для всіх методів перевірки текстів на унікальність, в тому числі і для метода шинглів, характерні помилки двох типів. Помилка першого типу полягає в тому, що система не «помітила» плагіат (пропуск плагіату). Помилка другого типу, навпаки, виникає при неправильному виявленні плагіату, коли насправді його нема (у допустимих межах). Оскільки вірогідність виникнення помилок обох типів залежить від кількості шинглів, необхідно проаналізувати результативність перевірки тексту на унікальність від кількості шинглів для документів різного об’єму.

Всі студентські роботи можна розділити на таки типи: тези, лабораторні роботи, есе, звіти, курсові та дипломні роботи.

Методика дослідження полягає втому, що було відібрано реальні студентські роботи різного типу, які були розміщені в базі даних антиплагіатної системи.

Для того, щоб щоб можна було точно відстежувати відсоток виправленого тексту, у кожній роботі було підраховано кількість слів, а потім замінено 20% тексту, тобто кожна робота мала 80% унікальності. При розбитті роботи на шингли, необхідно з порівнюваних текстів виділити під послідовності слів, що йдуть один за одним по 10 штук. Але при аналізі коротких документів, наприклад тезисів, об’єм яких займає лише одну сторінку аркушу, може статися неможливим розбиття тексту на 7 шинглів довжиною по 10 слів.

На рис. 1 показана залежність виникнення помилок визначення унікальності тексту від кількості шинглів ті об’єму текстів, що порівнюються.

Рис. 1. Залежність виникнення помилок від кількості шинглів ті об’єму текстів, що порівнюються

Аналіз результатів дослідження показує, що для менших за об’ємом робот треба вибирати довжину з меншою кількістю шинглів і навпаки. Якщо не дотримуватись цього правила, ймовірність помилки першого типу, тобто «пропуск» плагіату збільшується. Довжина у два та три шингли є найбільш придатною, тому що похибка при перевірці роботи будь якого розміру похибка варіюється від 0,56% до 3,47%.

Під час перевірки унікальності роботи слід враховувати те, що студенти використовують загальні правила оформлення робіт (шаблони). Особли-

во це є обов'язковим при перевірці не великих за об'ємом робіт, таких як тезиси чи лабораторні роботи. На рис. 2 показана залежність виникнення помилок визначення унікальності тексту від кількості шинглів з урахуванням основних правил оформлення тезисів та лабораторних робіт.

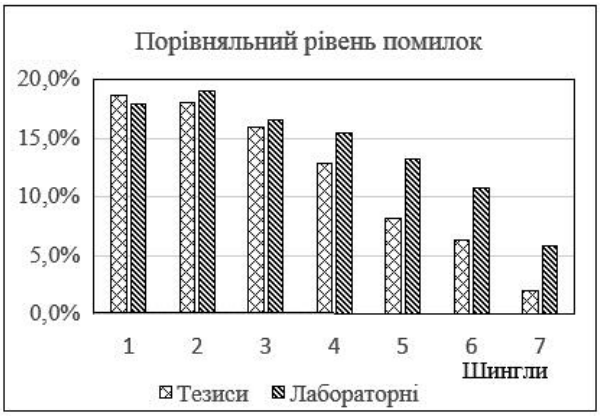

Рис. 2. Порівняльний рівень помилок з шаблонами і без шаблонами

Таким чином, для не великих за об'ємом робіт і при малих значеннях шинглів наявність стандартних елементів (шаблонів) дає суттєве зростання помилок в порівнянні з текстами, у яких стандартні елементи вилучені. З ростом числа шинглів для коротких документів зростання помилок стає дуже вагомим (рис.1), а різниця в рівні помилок між документами з шаблонами і без шаблонів стає незначною.

Кафедральну систему перевірки унікальності текстової інформації доцільно і досить просто розробити у вигляді інтерактивного веб-сайту через те, що у онлайн ресурсу є багато плюсів, наприклад:

доступ 24 години на добу;

немає необхідності встановлювати стороннє програмне забезпечення;

є можливість використовувати ресурс з будь-якого портативного пристрою і з будь-якої операційної системи.

Висновки

Результати проведених досліджень показують. Що найбільш оптимальна кількість шинглів для коротких текстів складає 3-4.

Для підвищення надійності виявлення плагіату необхідно враховувати наявність шаблонів в студентських роботах.

Реалізація системи у вигляді інтерактивного сайту дасть викладачам зручний і точний засіб перевірки в короткі терміни унікальність студентських робіт.

Список літератури

1. *Гасфилд Д. Строки, деревья и последовательности в алгоритмах / Д. Гасфилд. — СПБ. : «Невский диалект», БХВ-Петербург, 2003 – 654 с 2003.*

2. *Дербенев Н. Выявление нечетких дубликатов / Н. Дербенев. — СПБ. : Эксмо, 2010. — 768 с.*

3. *Broder, Andrei (2006). "Interview: "Search without a Box"". Yahoo! Search Blog. Retrieved 2006-03-04.*

4. *Manber U., Myers G. Suffix arrays, pages a new method for on-line search // SIAM J. Comput., 22, 1993. P. 935-948.*

5. *Kolcz A., Chowdhury A., Alspector J. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization // KDD 2004, 22-25 August, 2004, Seattle, Washington, USA*

Рецензент: д-р економічних наук, проф. О. І. Пушкар, Харківський національний економічний університет ім. Семена Кузнеця, Харків.

Автор*: КЛИМНЮК Віктор Євгенович*

Харківський національний економічний університет імені Семена Кузнеця, Харків, кандидат технічних наук, доцент, професор кафедри комп'ютерних систем і технологій.

Роб. тел. – 333-33-33, дом. тел. – 050-630-9462, E-mail – vek47@list.ru.

Проблемы создания кафедральной антиплагиатной системы

В. Е. Климнюк

В работе рассматривается возможность построения кафедральной системы проверки на плагиат студенческих работ. Предложен новый подход проверки документов на основе нечетких дубликатов с помощью алгоритма шинглов и с учетом шаблонов. Проведены исследования по определению оптимального количества шинглов для проверки уникальности коротких документов, что характерно для студенческих работ. Показано, что с ростом числа шинглов для коротких документов рост ошибок становится существенным. Предложена технология создания кафедральной системы проверки студенческих работ в виде сайта.

Ключевые слова: *шингл, нечеткие дубликаты, плагиат.*

Creating cathedral antiplagiarism system problems

V. E. Klymnjuk

The article considers the possibility of building a cathedral verification system for plagiarism student work. A new approach is proposed based on verification of documents fuzzy duplicate using an algorithm based on shingles and templates. Conducted studies to determine the optimal amount of shingles to check the uniqueness of short documents, which is typical for student work. It is shown that with increasing number of shingles for short documents growth becomes significant errors. The technology of creating a cathedral verification system in the form of student is proposed as site.

Keywords: *shingle, fuzzy duplicate, plagiarism.*