УДК 681.31

Магістрант Павлюченко В.О., ст. викладач Дробязко І.П., к.т.н., доцент Тесленко О.К.

Національний технічний університет України «Київський політехнічний інститут»

ЗАСОБИ ШВИДКОГО ПОШУКУ ЗАПОЗИЧЕНЬ

Abstract

Valentyn O. Pavliuchenko, student; Iryna P. Drobyazko, lecturer; Olexandr K. Teslenko, assoc. prof., PhD

Methods of fast plagiarism search

This paper concerns the task of fast plagiarism search. Efficiency criteria is studied and discussed. The method of splitting the search into two steps for fast and reliable plagiarism detection is proposed. The ways of further research are proposed as well.

Вступ

Проблема ефективного пошуку запозичень завжди була актуальною, особливо останнім часом у зв’язку з активним використанням електронного документообігу та глибоким проникненням інформаційних технологій у процес навчання. Розвиток мережі Інтернет суттєво спростив отримання інформації і цим значно поширив явище запозичення чужих робіт, у том числі й найпримітивнішого — коли знайдена інформація подається у практично незмінному вигляді як власна робота. Задача пошуку запозичень піддається автоматизації і її вирішення може суттєво допомогти при визначенні рівня самостійної роботи студентів. Для цього можна і потрібно широко використовувати комп’ютерні системи.

Існуючі системи ([1-4], та ін) є прикладами практичного застосування пошуку запозичень, водночас вони показують недоліки в плані ефективності їх застосування в освітній сфері. В освітній сфері визначення наявності запозичень має за мету визначення рівня самостійності виконання роботи студентом і є лише однією з складових інтегральної оцінки навчальних робіт, яку формує викладач або відповідна комісія. Тому, в загальному випадку, ефективність систем визначення запозичень пов’язана з розміром реальної допомоги викладачам, яку попри наявності суб’єктивного фактору, можна оцінити за розміром зменшення навантаження на викладачів і рівнем достовірності результатів системи.

Однією з важливих проблем тут є швидкість пошуку. Крім того, при вирішенні проблеми ефективності виявлення запозичень звичайно недооцінюють один з його аспектів: надто потужний пошук запозичень не дозволяє розділяти дійсно оригінальні роботи студентів та добре перетворенні запозичення, що зменшує ефективність системи. В [1] ці випадки часто кваліфікуються як запозичення через занадто глибокий пошук, який здатен знаходити багато схожого у оригінальних матеріалах (тобто таких, які не мають запозичень).

Постановка задачі

Задача полягає у розробці засобів визначення запозичень з *достатньою* (для викладачів) ефективністю.

Критерії ефективності

Ефективність засобів визначення запозичень можна визначити за допомогою наступних критеріїв:

1. Швидкість роботи системи – достатня для користувачів.
2. Стійкість системи - витрати студента (наприклад, його особистого часу) на подолання спроможності системи виявити запозичення повинні значно перебільшувати витрати на самостійне вирішення завдання.
3. Достовірність – кількість не визначених запозичень та дійсно самостійних робіт, помилково визначених як запозичення, повинна бути мінімальною.

Специфіка і вимоги до системи пошуку

Вище зазначено, що велика глибина пошуку запозичень не тільки а) дозволяє знаходити сильно трансформовані запозичення, але і б) відносить оригінальні роботи до категорії запозичених. Слід відмітити, що роботи, які є сильно трансформованими запозиченнями, по-перше, є здебільшого виключенням, а по-друге, складність перетворення таких робіт більша за складність виконання самої роботи. Звідси випливає їх недоцільність.

До безпосередніх вимог до системи пошуку запозичень можна віднести:

1. Пошук запозичень із достатньою ефективністю.
2. Розподілене зберігання бази даних існуючих робіт для одночасного розподіленого пошуку запозичень у конкретному документі на багатьох ЕОМ і спрощення додавання нових матеріалів.

3. Можливість здійснення пошуку різного типу робіт (тексту, зображення тощо).

4. Достатньо висока швидкість пошуку.

Побудова алгоритму прискореного пошуку

Максимальна ефективність пошуку (коли помилкова кваліфікація оригінального матеріалу як запозиченого практично неможлива) є окремою, досить складною задачею, для вирішення якої потрібні суттєві практичні дослідження. У даній роботі ставиться завдання отримання хоча й не максимальної ефективності, але такої, коли частка знайдених запозичень порівнянна з часткою, яку дає алгоритм з максимальною ефективністю. Звичайно, останнє слово залишається за викладачем, який самостійно приймає рішення, в тому числі у випадку сумнівних робіт.

Швидкість пошуку важлива для практичного використання системи. Мала швидкість існуючих систем [1] також вимагає приділити особливу увагу цьому аспекту. Очевидно, що можливі два шляхи вирішення цієї проблеми: а) прискорення алгоритму пошуку, б) нарощування обчислювальних потужностей.

Нарощування потужностей найкраще вирішується за рахунок розпаралелювання обчислень, що може бути досягнуто шляхом створення розподіленої системи. Розподілене зберігання бази даних існуючих робіт дозволяє спростити додавання нових робіт (відсутнє централізоване додавання) і нових апаратних потужностей до системи. Це особливо ефективно при використанні однієї системи для багатьох ВНЗ, а також у межах одного ВНЗ – для різних кафедр і факультетів.

Зупинимось на задачі прискорення алгоритму пошуку детальніше. Самий процес порівняння робіт можна поділити на онлайн та офлайн частини. Офлайн частина виконується до безпосереднього порівняння і ставить за мету підготовку окремо кожної із двох порівнюваних робіт (тієї, яку перевіряють, та однієї з існуючих), тобто виділення необхідної інформації для порівняння. Онлайн частина здійснює безпосередньо порівняння, і потребує інформацію про обидва документи від відповідних офлайн частин.

Поділ процесу порівняння на ці дві складові дозволяє побачити шляхи вирішення задачі його прискорення. Особливість полягає у можливості виконання офлайн частин для існуючих робіт (з бази даних) одноразово, замість їх виконання при кожному порівнянні. Для роботи, що перевіряється, офлайн частину також потрібно буде виконувати лише один раз перед усіма онлайн частинами порівняння. Таким чином, задача

прискорення порівняння зводиться до максимального перенесення алгоритму з онлайн до офлайн частин.

Максимальне перенесення алгоритму до офлайн частин здатне звести онлайн частину до простого посимвольного порівняння та пошуку спільних частин. Це забезпечить максимальну швидкість пошуку. Крім того, при посимвольному порівнянні фактично не аналізується семантична складова даних, що дає наступні переваги: а) онлайн частина стає універсальною, тобто може бути застосована для порівняння текстів як українською, так і практично будь-якою іншою мовою, а також для як початкових текстів програм, так і виконуваних програм, аудіо файлів і т.п.; б) помилки у виявленні запозичень зменшуються, оскільки, головним чином, вони зумовлені трактуванням семантичної складової.

Таким чином, при створенні алгоритму швидкого пошуку слід починати з онлайн частини, яка здійснює посимвольний пошук спільних блоків даних. Оптимізація цього алгоритму пошуку була виконана у [5]. Дещо більш детальний опис можна знайти у [6]. Подальша реалізація пошуку є реалізацією офлайн частини. Потрібно вирішити наступні задачі:

1. Зменшення об’єму даних, який передається до офлайн частини.
2. Оптимізація формату подання даних до онлайн частини.
3. Забезпечення стійкості алгоритму до модифікації даних.
4. Забезпечення стійкості алгоритму до методів ускладнення пошуку, що базуються на використанні специфічних для формату можливостей.

Перша задача ставить за мету, головним чином, прискорення пошуку шляхом зменшення об’єму даних для порівняння. Ущільнення інформації можливе на основі інформації про формат її подання (наприклад, замінити слова їх номерами із певного словника). Особливістю ущільнення є отримання пошуку, нечутливого до деяких змін у вхідних даних (наприклад, до закінчень слів). Крім того, ущільнення може не виконуватись, якщо неможлива його реалізація без втрати важливої для порівняння частини даних.

Задача оптимізації формату подання даних до онлайн частини ставить за мету перетворення даних для порівняння щоб прискорити порівняння в онлайн частині. Зберігання результатів офлайн частини у базі даних існуючих робіт є доцільним у форматі, який спеціально розроблено для такого прискорення. Формат включає попередньо відсортовані та оптимізовані для більш швидкого пошуку структури даних. Розроблений алгоритм онлайн частини містить опис цих структур із оптимізованим форматом подання даних. Робота алгоритму була описана у [5].

Забезпечення стійкості алгоритму до модифікації даних є задачею, яку необхідно вирішити для результативного пошуку. Прикладом модифікації для векторних форматів зображень є зміна координат

елементів без зміни відносного розміщення елементів. Але для пошуку запозичень у ВНЗ достатньо невисокої стійкості, яка в свою чергу залежить від формату даних. Слід зазначити, що частина модифікацій усувається при вирішенні задач ущільнення інформації та забезпечення стійкості до ускладнення пошуку.

Задача забезпечення стійкості алгоритму до методів ускладнення пошуку, які базуються на використанні специфічних для формату можливостей, є суто інженерною.

Висновки

На основі дослідження проблеми та існуючих систем пошуку запозичень сформульовано вимоги до подібних систем у ВНЗ. Однією з основних вимог є забезпечення достатньої швидкості пошуку. Визначено недоцільність створення надто потужного пошуку. Сформульовано критерії ефективності пошуку.

Запропоновано алгоритм пошуку запозичень з поділом власне пошуку на дві частини, одна з яких виконується заздалегідь та забезпечує стійкість та достовірність системи, а друга - швидкість. Таке рішення дозволяє спростити алгоритм порівняння і тим самим значно підвищити ефективність роботи системи в цілому.

У подальшому можлива деталізація алгоритмів попередньої обробки робіт з використанням зазначених принципів та алгоритмів для створення комплексної системи пошуку запозичень.

Література

1. http://www.antiplagiat.ru
2. http://searchinform.com
3. http://www.plagiarism-detector.com
4. http://turtin.com
5. *Павлюченко В.О., Тесленко О.К.* Алгоритм визначення кореляції символьних послідовностей з використанням сигнатур // Свідоцтво про реєстрацію авторського права на твір №23869 від 03.2008 р.
6. *Drobiazko I.P , Pavliuchenko V.O., Teslenko O.K.* GRID-oriented Techniques for Effective Plagiarism Search. // Матеріали 4-ої міжнародної науково-технічної конференції “Сучасні комп’ютерні системи та мережі: розробка та використання” (ACSN’2009), Львів 2009. с. 87-89.