УДК 004.4
В.В. Зелений
(Тернопільський національний технічний університет імені Івана Пулюя)

АНАЛІЗ АЛГОРИТМІВ ПОШУКУ ПЛАГІАТУ ЛЕКСЕМ

UDC 004.4
V.V. Zelenyi

ANALYSIS OF PLAGIARISM SEARCH ALGORITHMS

Щоб виявити плагіат, важливо мати широкі знання про його можливі форми та типи, а також існування різних засобів та систем для його виявлення. Плагіат може мати місце у статті чи будь-якому текстовому виданні. З роками було запроваджено чимало інструментів та прийомів для виявлення плагіату. У цій доповіді буде висвітлено кілька перспективних методів виявлення плагіату та проаналізовано складність цих алгоритмів.

1. Плагіат у сучасному суспільстві

Завдяки цифровій ері, обсяг цифрових ресурсів у Всесвітній павутині збільшується. При швидкому зростанні цих ресурсів, можливість порушення авторських прав та плагіат також зростають. Щоб вирішити цю проблему дослідники почали працювати над виявленням плагіату між різними мовами з 1990 р. Це було новаторським методом виявлення копій у цифрових документах[1].

2. Виявлення плагіату

Плагіат може відбуватися між двома однаковими або двома різними мовами. На основі мовної однорідності або неоднорідності текстових документів, що порівнюються, виявлення плагіату можна розділити на два основних типи[4].

1. Виявлення одномовного плагіату: цей тип виявлення стосується однорідних текстів плагіату, наприклад, українська-українська. Більшість методів виявлення відносяться до цієї категорії[2].

2. Виявлення міжмовного плагіату: цей підхід виявлення може виконуватись у неоднорідних текстах плагіату, українська-англійська. Є лише невелика кількість способів розпізнавання даного плагіату через труднощі у пошуку близькості між двома текстовими сегментами для різних мов.

2.1. Знаходження подібності для порівняння документів або сегментів тексту

Щоб виявити плагіат, нам слід виміряти подібність між двома документами. Для цього більшість дослідників використовують наступні два типи метрик подібності[3].

1. Показник подібності рядків (String Similarity Metric): це метрика, яка вимірює відстань між двома текстовими рядками для приблизної відповідності рядків.

2. Метрика векторної схожості (Vector Similarity or Cosine similarity Metric): коефіцієнт подібності двох не нульових векторів у предгільбертовому просторі, який обчислюється як косинус кута між ними.

2.2. Методи виявлення плагіату

Виявлення плагіату в текстовому документі, з високою точністю, є складним завданням. Два десятиліття дослідники повідомляють про велику кількість методів для вирішення цього завдання. Деякі відомі методи будуть висвітлені далі.

1. Відстань Левенштейна (Levenshtein distance): у теорії інформації і комп'ютерній лінгвістиці міра відмінності двох послідовностей символів (рядків). Обчислюється як мінімальна кількість операцій вставки, видалення і заміни, необхідних для перетворення одної послідовності в іншу.

2. Відстань Джаро-Вінклера (Jaro–Winkler distance): є рядковою метрикою, що вимірює відстань редагування між двома послідовностями. Це варіант, запропонований у 1990 р. Вільямом Е. Вінклером з метрики відстані Джаро

3. Пошук найдовшої спільної підпослідовності (longest common subsequence, LCS) це завдання пошуку послідовності, яка є підпослідовністю кількох послідовностей. Часто завдання визначається як пошук всіх найбільших спільних підпослідовностей.

4. N-грама (N-gram) це послідовність з n елементів. З семантичної точки зору, це може бути послідовність звуків, складів, слів або букв. На практиці частіше зустрічається N-грами як ряд слів, стійкі словосполучення називають колокацію. Послідовність з двох послідовних елементів часто називають біграм, послідовність з трьох елементів називається триграма. Не менш чотирьох і вище елементів позначаються як N-грами, N замінюється на кількість послідовних елементів.

5. Міра Жаккара (Jaccard index) це бінарна міра подібності, запропонована Полем Жаккаром в 1901 році. Запропонований метод здобув поширення і нині використовується для оцінки подібності скінченних множин, в інформатиці, для пошуку подібних документів, плагіату

Висновок.

В даній роботі було розглянуто методи виявлення та класифікації плагіату. На основі отриманих даних, було зроблено висновок на рахунок складності деяких алгоритмів пошуку збігів та плагіату в текстах. Було розроблено таблицю для візуалізації отриманих результатів дослідження.

Таблиця 1. Порівняння результатів

Назва алгоритму	Тип виміру	Чи нормаізований	Складність
Levenshtein	distance	ні	$O(m*n)^1$
Jaro–Winkler	similarity distance	так	$O(m*n)$
LCS	distance	ні	$O(m*n)^{1,2}$
N-gram	distance	так	$O(m*n)$
Jaccard index	similarity distance	так	$O(m+n)$

- де m і n – довжина порівнювальних лексем.

Література.

1. Чиркин Е.С. Системы автоматизированной проверки на неправомерные заимствования // Вестник ТГУ. – 2013. − №12. − С. 164-171.
2. Петренко В.С. Поняття плагіату [Електронний ресурс]. — Режим доступу: http://www.clj.nuoua.od.ua/archive/14/29.pdf.
3. Закон України “Про авторське право і суміжні права” [Електронний ресурс]. — Режим доступу: https://zakon.rada.gov.ua/laws/show/3792-12.
2. Зеленков Ю.Г, Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов [Електронний ресурс]. – Режим доступу: http://rcdl2007.pereslavl.ru/papers/paper_65_v1.pdf