CEUR Workshoop Proceedings. 2021. Vol. 2866. P. 61–73. URL: http://ceur-ws.org/Vol-2866/ceur_61-73Rogushina6.pdf (дата звернення: 20.10.2022).

DOI https://doi.org/10.30525/978-9934-26-277-7-201

SYSTEM OF SEARCHING FOR VAGUE DUPLICATES IN ELECTRONIC TEXTS

СИСТЕМА ПОШУКУ НЕЧІТКИХ ДУБЛІКАТІВ В ЕЛЕКТРОННИХ ТЕКСТАХ

Rozlomii I. O.
PhD,
Senior Lector at the Department
of Information Technologies
Bohdan Khmelnytsky
National University of Cherkasy
Cherkasy, Ukraine

Розломій І. О.
кандидат технічних наук,
старший викладач кафедри
інформаційних технологій
Черкаський національний
університет
імені Богдана Хмельницького
м. Черкаси, Україна

Veretelnyk V. V.
PhD,
Docent at the Department
of Information Technologies
Bohdan Khmelnytsky
National University of Cherkasy
Cherkasy, Ukraine

Веретельник В. В.
кандидат технічних наук, доцент,
доцент кафедри
інформаційних технологій
Черкаський національний
університет
імені Богдана Хмельницького
м. Черкаси, Україна

Hrushovyi V. O.
Master at the Department
of Information Technologies
Bohdan Khmelnytsky
National University of Cherkasy
Cherkasy, Ukraine

Грушовий В. О.
магістр кафедри
інформаційних технологій
Черкаський національний
університет
імені Богдана Хмельницького
м. Черкаси, Україна

Сучасний розвиток інформаційних технологій та мережі Інтернет надав широким колам користувачів доступ до великих об’ємів

інформації. Це призвело до бурхливого зростанню кількості дубльованої і запозиченої інформації. Особливо це помітно в самій мережі інтернет (новинні сайти, блоги, соціальні мережі, дані в інформаційно-пошукових системах), у сфері освіти та засобах масової інформації. Іноді запозичений контент зустрічається у наукових колах [1, с. 194].

Повідомлення, що публікуються одним джерелом, часто багаторазово передруковується іншими (в початковому вигляді або з невеликими змінами). В результаті, при виконанні автоматичного збору документів із багатьох джерел у текстовій колекції, що формується, накопичуються ідентичні або близькі за змістом документи – дублікати. З цієї причини задача по знаходженню подібних запозичень набуває підвищеної актуальності. Робота присвячена проблемі виявлення дублікатів та нечітких дублікатів у текстах науково-технічної інформації.

Для пошуку дублікатів у вихідних текстах найбільшого розвитку набули метод шинглів та метод N-грам. Застосування таких засобів є ефективним для аналізу наукової інформації (статей, доповідей на конференціях, дисертацій).

Для аналізу запозичених фрагментів у вихідних текстах пропонується узагальнений та модифікований підхід, що поєднує метод структурного аналізу кодів (токени), метод шинглів.

Однією з важливих особливостей електронної документації є наявність фрагментів тексту, що повторюються, що значно ускладнює написання документа, оскільки у разі відсутності додаткових інструментів процес його розробки та супроводу може виявитися досить трудомістким і забрати чимало часу. Також, якщо не відстежувати наявність повторів у тексті, стає можливим зниження якості документа. Тому важливим завданням стає спрощення та часткова автоматизація процесу пошуку та рефакторингу таких повторів. Найбільшу популярність отримав метод (алгоритм) шинглів та метод N-грам.

Метод шинглів заснований на поданні текстів у вигляді множини послідовностей фіксованої довжини, що складаються з сусідніх слів. При значному перетині таких множин документи будуть схожі один на одного. Одна з модифікацій методу, що отримала назву «супершинглів» і «мегашинглів», використовується для швидкого виявлення подібних документів.

Іншим сигнатурним підходом, що базується на лексичних принципах, є метод «опорних» слів. Для документів складаються за

певними правилами набори опорних слів, за якими будуються сигнатури документів. Збіг сигнатур говорить про подібність самих документів. Ця група методів, незважаючи на велику складність реалізації, показує найкращі результати у виявленні схожих документів.

Для практичного використання описаних підходів у завданнях виявлення нечітких дублікатів у текстах нині існує досить велика кількість сервісів, що дозволяють, так чи інакше, виявити запозичений контент.

Більшість із розглянутих систем використовує у своїй роботі метод «шинглів». За дослідженнями [2, с. 1217] цей метод демонструє високу точність виявлення дубльованих текстів. Проте через особливості реалізації результати перевірки в кожній системі сильно відрізняються від інших. Мінусом методу є неможливість обробки синонімів [3, с. 766]. Це значний недолік існуючих систем. Існує багато засобів синонімізації текстів. Використання подібних засобів може звести нанівець роботу систем перевірки текстів, наприклад, на плагіат.

Таким чином, для ефективного виявлення нечітких дублікатів системи повинні вміти обробляти стоп-слова, застосовувати методи лематизації, здійснювати заміну літер з англійської на українську та вміти обробляти синоніми. Крім того, в універсальних системах має бути підтримка пошуку як в Інтернеті, у внутрішній базі. Звіт про перевірку має бути детальний і містити відомості про знайдені збіги з відображенням списку джерел.

Метод шинглів має високу точність знаходження відсотку подібності досліджуваного тексту до колекції текстів, тобто наскільки відсотків один з документів містить набір тексту з іншого.

Методи пошуку подібних послідовностей використовуються не тільки в дослідженні повторень в електронних текстах, а й в інших галузях наук. Метод N-gram [4, с. 912] застосовується в таких науках як математика, біологія, криптографія та музиці. Суть даного методу полягає в розбитті послідовності даних на n-грами.

Для реалізації системи було обрано метод шинглів та метод N-грам. Метод шинглів показує високу точність пошуку нечітких дублікатів, тому що він порівнює хеш-суму рядку. Метод N-грам, має застосування в багатьох галузях наук, для дослідження послідовності закономірностей даних. Цей метод легко реалізується та показує точний результат своєї роботи. Основними критеріями власного методу це, простота реалізації та швидкодія.

Перший етап даного методу – канонізація тексту, шляхом відкидання всіх непотрібних елементів які не беруть участь у роботі методів пошуку дублікатів, відбувається формування даних для методу.

На другому етапі потрібно порівнювати один файл з іншим, а щоб убезпечити роботу коректно, тому що під час такого створення формату слів, випадково може існувати 2 або більше слів, які можуть формувати одинаковий формат. Тому було прийнято рішення, порівнювати послідновність таких слів, від 2...N, де N – кількість слів у послідовності. Формула зхожості виглядає наступним чином (1):

$$R(T1,T2) = \frac{LENGTH(T1 \cap T2)}{LENGTH(T1)*100}, \quad (1)$$

де T1, T2 – перший та другий текст відповідно, LENGTH – кількість послідовностей.

Результатом, методу є відсоток запозичень із першого тексту в другому. Можна відмітити, що формування даних для методу зменшує розмір вихідного файлу до 10 разів, тому що йому не потрібні майже всі символи, які є в текстах.

Перший етап роботи системи пошуку нечітких дублікатів – це передача вхідних даних в систему. Ці дані, являються окремими налаштуваннями методів пошуку нечітких дублікатів. Наприклад, метод шинглів, може отримати кількість слів в одному шинглі, або кількість випадкових хеш-сум шингла. Таким чином, можна дослідити поведінку системи з різними вхідними даними.

Наступний етап, після того як система, отримала вхідні дані, здійснюється ініціалізація системи, а саме методів пошуку. Далі виконується кожен із методів одночасно для одного контрольного варіанту електронного тексту, після того як методи завершили свою роботу, вони створюють звіт про результат пошуку.

У блоці збору результатів, для кожного методу формується час на виконання та відсоток запозичень із кожного файлу, і на основі цих даних можна показати наскільки кожен метод оптимально за часом визначає чи є електронний текст запозиченим.

Розроблене програмне забезпечення може бути рекомендоване для студентів очної/заочної освіти для попереднього автоматичного виявлення дублікатів (запозичень), виключивши втручання людини-оператора (експерта), і тим самим підвищити якість даних в інформаційних масивах дипломних робіт.

References:

1. Hajishirzi, H., Yih, W. T., & Kolcz, A. (2010, July). Adaptive near-duplicate detection via similarity learning. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 419–426).

2. Sharapova, E. V., & Sharapov, R. V. (2014). System of fuzzy duplicates detection. In Applied Mechanics and Materials (Vol. 490, pp. 1503–1507). Trans Tech Publications Ltd.

3. Wang, Z., Zuo, C., & Deng, D. (2022, June). TxtAlign: Efficient Near-Duplicate Text Alignment Search via Bottom-k Sketches for Plagiarism Detection. In Proceedings of the 2022 International Conference on Management of Data (pp. 1146–1159).

4. Mishra, A. R., Panchal, V. K., & Kumar, P. (2020). Similarity Search based on Text Embedding Model for detection of Near Duplicates. *International Journal of Grid and Distributed Computing*, 13(2), 1871–1881.