УДК 004.421

К.т.н., доцент Петрашенко А.В., магістрант Шелест І.Є.

Національний технічний університет України
«Київський політехнічний інститут»

ПРОГРАМНІ ЗАСОБИ КОНТРОЛЮ ОРИГІНАЛЬНОСТІ ТЕКСТІВ

Вступ

У наш час мільйони документів стали доступними в електронній формі, створено тисячі інформаційних систем та електронних бібліотек. В такій ситуації постає задача контролю оригінальності текстів та вилучення дублікатів із загального масиву документів. Зрозуміло, що абсолютно ідентичні тексти можна легко виявити, достатньо зберігати контрольну суму кожного тексту та за нею виявляти всі дублікати. Проте цей метод не ефективний, якщо в текст-дублікат було внесено незначні зміни (переставлено окремі фрази, речення в тексті тощо).

Для вирішення цієї проблеми Уді Манбер у 1994 році запропонував ідею, а у 1997 році Андрій Бродер повністю доопрацював та описав її, назвавши цей метод «алгоритмом шинглів». Також існує метод «описуючих слів», запропонований групою розробників пошукової системи «Яндекс» [1]. Незважаючи на широке застосування цих методів не існує визначених підходів врахування специфіки української мови, тому вдосконалення методів автоматизованого контролю дублікатів документів є актуальною науково-технічною задачею.

Постановка задачі

Метою дослідження є розробка узагальненого алгоритму функціонування програмних засобів контролю оригінальності текстів, що виявляє однакові частини текстів, враховуючи специфіку української мови.

Метод «описових (дискриптивних) слів»

Сутність базового алгоритму полягає у наступному: спочатку формується контрольна вибірка слів (близько 2000-3000 слів) з максимальної кількості документів. Ця вибірка повинна бути такою, щоб за її допомогою можна було достатньо повно описати практично будь-який документ та цей опис не повинен бути надмірним. Таким чином, для

формування вибірки потрібно відкинути слова, які найбільш та найменш вживані, так звані стоп-слова. Також у вибірку не потрапляють прикметники, оскільки вони не несуть змістовного навантаження. Для кожного слова встановлюється його гранична частота появи. Далі кожен документ зіставляється з вибіркою і розраховується вектор, розмірність якого дорівнює кількості слів у вибірці. Компоненти вектора можуть набувати двох значень – 0 або 1: 0 – якщо слова з вибірки немає в документі або його частота вживання менша за обрану граничну, 1 – якщо слово зустрічається в документі з частотою, що є більшою або дорівнює граничній. Далі документи перевіряються на дублювання шляхом зіставлення їх векторів. Цей двійковий вектор вважається нечітким цифровим підписом документа. Кожен вектор однозначно визначає клас схожих документів [1].

Метод «шинглів»

Для кожного десятислів'я тексту розраховується контрольна сума («шингл»). Десятислів'я обробляються з перекриттям, так, щоб жодне слово не втрачалося при подальшому аналізі. Далі з усієї кількості контрольних сум (очевидно, що їх стільки ж, скільки слів в документі мінус 9) відбираються тільки ті, які діляться на деяке число, наприклад на 25. Оскільки значення контрольних сум розподілено рівномірно, спосіб формування вибірки жодним чином не прив'язаний до змісту тексту. Очевидно, що повтор навіть одного десятислів'я – вагома ознака дублювання, якщо ж їх багато, скажімо, більше половини, то з певним ступенем впевненості можна стверджувати, що копія знайдена. Адже один «шингл», що співпав, у вибірці відповідає приблизно 25 десятислів'ям, що співпали в тексті. Таким чином, можна визначати відсоток перекриття текстів та виявляти всі його джерела [2][3].

Узагальнений алгоритм порівняння текстів

На основі вже існуючих методів контролю оригінальності текстів розроблено узагальнений алгоритм, що складається з наступних етапів.

1. Нормалізація текстів: вилучення знаків пунктуації та спеціальних символів, зведення слів до базової форми (використовуючи словник української мови).
2. Розбиття тексту на «шингли».
3. Розрахунок контрольних сум «шинглів» (можливі різні алгоритми хешування: CRC32, FNV, MD5, SHA1, SHA2 тощо).
4. Запис отриманих контрольних сум у базу даних чи файл.
5. Пошук однакових частин тексту у різних файлах.

Для отримання більш точних результатів у запропонованому узагальненому алгоритмі можна варіювати наступні дані:

- довжина підрядка для формування «шинглу» (обрана довжина повинна звести до мінімуму випадкові збіги тексту);
- алгоритм для розрахунку контрольних сум;
- кількість контрольних сум, що обираються для порівняння, та на практичному застосуванні алгоритму виявлення, які варіанти кожного з пункту будуть давати найкращі результати.

Ілюструє розроблений алгоритм наступна схема (рис.1).

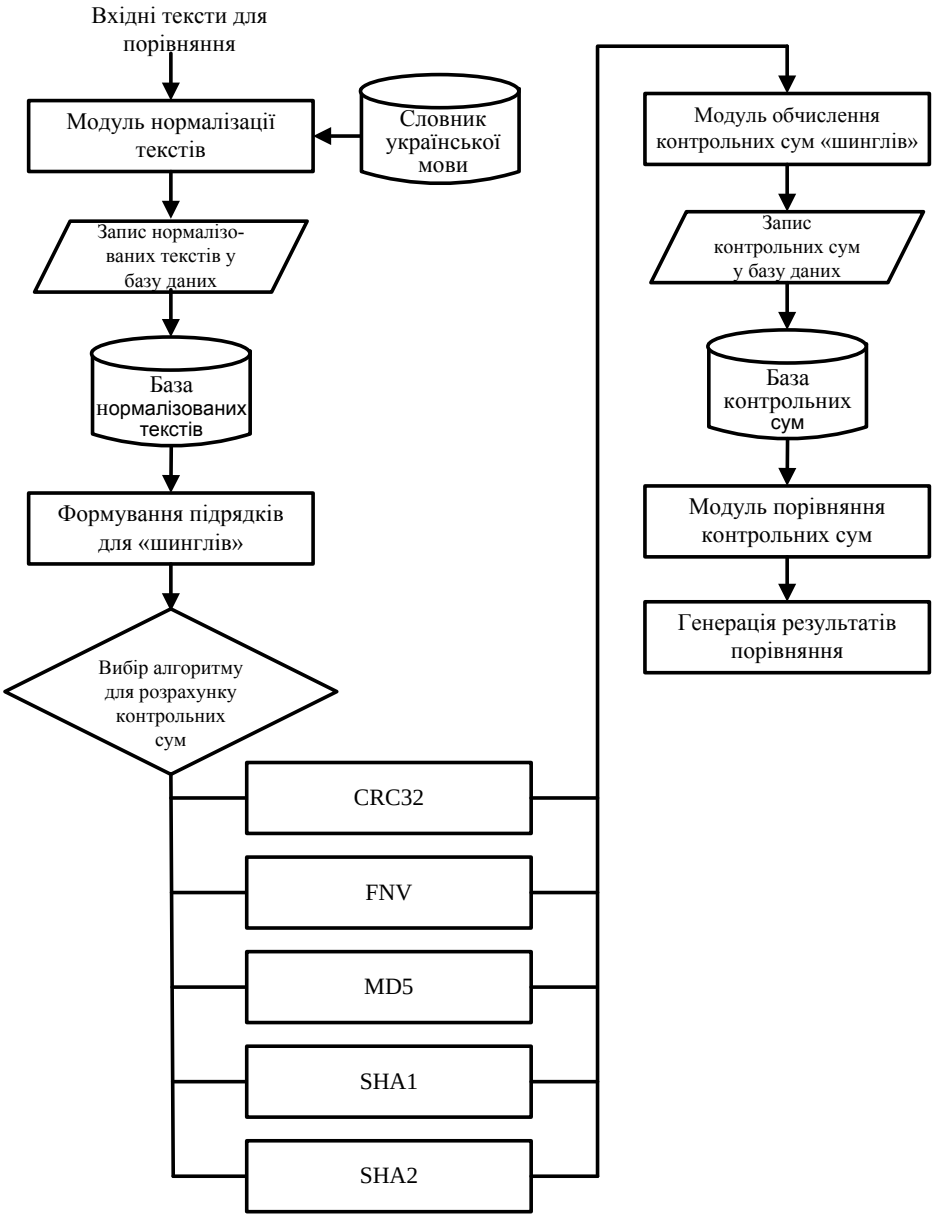

Рис.1. Узагальнена схема функціонування програмних засобів контролю оригінальності текстів

Методи контролю оригінальності текстів можуть застосовуватись [4]:

1) у пошукових системах для виявлення дублікатів документів, щоб поліпшити результати пошуку;

2) для кластеризації документів за їх подібністю;

3) для формування архіву електронних видань;

4) для фільтрування «спаму» в електронній пошті;

5) при встановленні порушення авторських прав, незаконного копіювання інформації тощо.

Висновки

У ході дослідження розроблено узагальнений алгоритм порівняння текстів та вказано основні його етапи, аналіз яких приведе до підвищення ефективності вже відомих методів вирішення даної задачі.

Крім того, розроблено узагальнену схему функціонування програмних засобів. Однією з головних переваг запропонованої схеми є використання словника української мови, оскільки всі попередні реалізації методів контролю оригінальності текстів базувались на виявленні текстів-дублікатів, що написані переважно російською чи англійською мовами.

Література

1. *Ilyinsky S., Kuzmin M., Melkov A., Segalovich I.* An efficient method to detect duplicates of Web documents with the use of inverted index // WWW Conference, 2002. http://www2002.org/CDROM/poster/187/ .
2. *Andrei. Z. Broder, Steven. C. Glassman, and Mark. S. Manasse* Syntactic.
1. Clustering of the Web. In Proceedings of the Sixth World Wide Web Conference, 1997. http://www.std.org/~msm/common/clustering.html.
2. *Manber U.* Finding similar files in a large file system // Proceedings of the USENIX Winter 1994 Technical Conference, pages 1–10, San Fransisco, 1994.
3. *Зеленков Ю.Г, Сегалович И.В* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2007, Переславль, Россия, 2007. – Том 1, С. 166-174.