The survey of near-duplicate revealing techniques for application to automated quiz review

Vitaly Yakovyna, Rostyslav Kaminskii, Vsevolod Smirnov

Software Department, Lviv Polytechnic National University, UKRAINE, Lviv, S. Bandery street 12, E-mail: yakovyna@lp.edu.ua

The problem of near-duplicates detecting is one of the most important and difficult problem for analyzing web data and information search on the Internet.

At the same time, while using the computer knowledge testing the open answers review is still carried out manually, thus resulting in subjective evaluation and significant time consumption. Thus the purpose of this paper is to review and analyze the near-duplicate detection methods for use in automated quiz review system.

One of the first studies on finding fuzzy duplicates have been carried out by U. Manber [1] and N. Heintze [2]. In these works the post-sequence of adjacent letters were used to build the sample. The file [1] or document [2] fingerprint includes all text substrings of fixed length. Numerical values of fingerprint are calculated using the Karp-Rabin algorithm of random polynomials [3].

In 1997 A. Broder et al. [4, 5] proposed a new, "syntactic" method of estimating similarity between documents based on the presentation of a document as a set of sequences of fixed length *k*, consisting of neighboring words. Such sequences are called shingles. Two documents were considered to be duplicates if their shingles sets substantially overlap.

Another signature-based approach based on not syntactic, but lexical principles, was proposed by A. Chowdhury et al. in 2002 and improved in 2004 [6, 7].

A similar approach is described in US Patent 6,658,423 by W. Pugh et al. [8] affiliating at Google Inc. The authors offer a comprehensive dictionary of the document to be divided into a fixed number of lists of words using a hash function. Then for each list the fingerprint is calculated and two documents are duplicates if they have at least one common fingerprint.

Another signature approach, also based on lexical principles is a method of "descriptive words" proposed by S. Ilyinsky et al. [9]. This method, starting with a set of hundreds of thousands of words, keep a set of 3–5 thousand signatures for extremely fast and efficient calculations [9].

Being modified appropriately, these methods can be used to automate the answers review by comparison with a standard response (e.g. with textbook). The optimal parameters of the algorithm, as well as the similarity threshold require further research.

Огляд методів виявлення нечітких дублікатів для автоматизованої перевірки тестових завдань

Віталій Яковина, Ростислав Камінський, Всеволод Смірнов

Кафедра програмного забезпечення, Національний університет "Львівська політехніка", УКРАЇНА, м.Львів, вул.С.Бандери, 12, E-mail: yakovyna@lp.edu.ua

Описано основні синтаксичні та лексичні методи виявлення нечітких дублікатів та показано перспективу їх використання для автоматизованої перевірки тестових завдань у відкритій формі.

Ключові слова – нечіткий дублікат, тест, аналіз змісту документів.

I. Вступ

Проблема виявлення нечітких дублікатів є однією з найбільш важливих і важких задач аналізу веб-даних і пошуку інформації в інтернеті. Актуальність цієї проблеми визначається різноманітністю додатків, у яких необхідно враховувати "схожість", наприклад, текстових документів – це і поліпшення якості індексу і архівів пошукових систем за рахунок видалення надлишкової інформації, і об'єднання новин в сюжети на основі подібності цих повідомлень за змістом, і фільтрація спаму (як поштового, так і пошукового), встановлення порушень авторських прав при незаконному копіюванні інформації (проблема плагіату), та ряд інших.

Разом з тим, при використанні комп'ютерного тестування знань перевірка відповідей у відкритій формі і надалі здійснюється вручну, що приводить до суб'єктивізму оцінювання та значних витрат часу викладача. Таким чином метою цієї роботи є огляд та аналіз методів виявлення нечітких дублікатів з метою використання їх в автоматизованій системі перевірки тестових завдань.

II. Синтаксичні методи визначення нечітких дублікатів

Одними з перших досліджень в області знаходження нечітких дублікатів є роботи U. Manber [1] та N. Heintze [2]. У цих роботах для побудови вибірки використовуються пост-послідовності сусідніх букв. Дактилограма файлу [1] або документа [2] включає всі текстові підрядки фіксованої довжини. Чисельне значення дактилограм обчислюється за допомогою алгоритму випадкових поліномів Карпа–Рабіна [3]. В якості критерію схожості двох документів використовується відношення числа спільних підрядків до розміру файлу або документа. U. Manber використовував цей підхід для знаходження схожих файлів (утиліта sif), а N. Heintze – для виявлення нечітких дублікатів документів (система Koala).

У 1997 році A. Broder та ін. [4, 5] запропонували новий, "синтаксичний" метод оцінки подібності між документами, заснований на представленні документа у вигляді множини послідовностей фіксованої довжини *k*, які складаються з сусідніх слів. Такі послідовності були названі "ґонтом" або "черепицею" (англ. shingle). Два документа вважалися схожими, якщо їх множини "ґонтин" істотно перетиналися. Оскільки кількість "ґонтин" приблизно дорівнює довжині документа в словах, тобто є достатньо великою, авторами [4, 5] були запропоновані два методи семплювання для отримання репрезентативних підмножин.

Перший метод залишав тільки ті "ґонтини", чиї дактилограми, які обчислюють за алгоритмом Карпа–Рабіна [3], ділилися без залишку на деяке число *m*. Основний недолік цього методу – залежність вибірки від довжини документа.

У другому методі для кожного ланцюжка обчислюються 84 дактилограми за алгоритмом Карпа–Рабіна [3] за допомогою взаємно-однозначних і незалежних функцій. Потім 84 "ґонтини" розбиваються на 6 груп по 14 (незалежних) "ґонтин" у кожній. Ці групи називаються "суперґонтинами".

Для ефективної перевірки збігу не менше 2-х "суперґонтин" (і відповідно, підтвердження гіпотези про подібність змісту) кожен документ представляється усіма можливими попарними поєднаннями з 6 "суперґонтин", які називаються "мегаґонтинами". Два документи подібні за змістом, якщо у них збігається хоча б одна "мегаґонтина".

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

III. Лексичні методи визначення нечітких дублікатів

Інший сигнатурний підхід, заснований вже не на синтаксичних, а на лексичних принципах, був запропонований A. Chowdhury та ін. в 2002 р. і удосконалено в 2004 р. [6, 7].

Основна ідея такого підходу полягає в обчисленні дактилограми I-Match (хеш-функції SHA-1) на основі перетину словника колекції документів та множини різних слів документу для представлення змісту документів.

Два документи вважаються схожими, якщо у них збігаються I-Match сигнатури. Перевагою алгоритму є його висока ефективність для порівняно невеликих за розміром документів. Основний недолік – нестійкість до невеликих змін змісту документу. Для подолання вказаного недоліку вихідний алгоритм був модифікований [7].

Схожий підхід описаний в патенті США [8]. Автор пропонує повний словник документа розбити на фіксовану кількість списків слів за допомогою будь-якої функції хешування. Потім для кожного списку обчислюється дактилограма і два документи вважаються подібними, якщо вони мають хоча б одну спільну дактилограму.

Ще одним сигнатурним підходом, також заснованим на лексичних принципах (тобто використанні словника), є метод "опорних" слів, запропонований С. Ільїнським та ін. [9]. Цей метод дозволяє, почавши з вибірки в сотні тисяч слів, залишити набір з 3–5 тисяч, розрахунок сигнатур по яким з застосуванням повнотекстового індексу здійснюється надзвичайно швидко та ефективно [9].

Висновок

В роботі розглянуто основні методи виявлення нечітких дублікатів засновані як на синтаксичному, так і на лексичному підходах. За відповідної модифікації ці методи можуть бути використані для автоматизації перевірки відповідей на тестові завдання у відкритій формі шляхом порівняння відповіді з еталонною (наприклад з підручником). Оптимальні параметри алгоритму, так само як і поріг подібності потребують подальших досліджень.

Література

[1] U. Manber. Finding Similar Files in a Large File System. // WTEC'94 Proceedings of the USENIX Winter 1994 Technical Conference, p. 2.

[2] N. Heintze. Scalable document fingerprinting. // Proc USENIX Workshop on Electronic Commerce (1996), pp. 191–200.

[3] Д. Гасфилд. Строки, деревья и последовательности в алгоритмах. – СПб.: Невский диалект, 2003. – 656 с.

[4] A. Broder, S. Glassman, M. Manasse and G. Zweig. Syntactic clustering of the Web. // Computer Networks and ISDN Systems, Vol. 29 (1997), Issues 8–13, pp. 1157–1166.

[5] A. Broder. On the resemblance and containment of documents. // Proceedings of the Compression and Complexity of Sequences 1997, pp. 21–29.

[6] A. Chowdhury, O. Frieder, D. Grossman, M. McCabe. Collection statistics for fast duplicate document detection. // ACM Transactions on Information Systems, Vol. 20 (2002), Issue 2, pp. 171–191.

[7] A. Kolcz, A. Chowdhury, J. Alspector. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization. // Proc. 10th ACM Int. Conference on Knowledge discovery and data mining (KDD'04), pp. 605–610.

[8] W. Pugh, M. Henzinger. Detecting duplicate and near-duplicate files. // US Patent 6658423, 2003.

[9] S. Ilyinsky, M. Kuzmin, A. Melkov, I. Segalovich. An efficient method to detect duplicates of Web documents with the use of inverted index. // Proc. 11th Int. World Wide Web Conference (WWW'2002).