УДК 004.738.52; 371.26

[1)]В. Яковина, [2)]Т. Смірнова, [1)]В. Смірнов
Національний університет “Львівська політехніка”,
[1)]кафедра програмного забезпечення,
[2)]кафедра напівпровідникової електроніки

ПРО МОЖЛИВІСТЬ ВИКОРИСТАННЯ МЕТОДІВ ПОШУКУ НЕЧІТКИХ ДУБЛІКАТІВ ДЛЯ АВТОМАТИЗАЦІЇ ПЕРЕВІРКИ ТЕСТОВИХ ЗАВДАНЬ

Розглянуто основні методи пошуку нечітких дублікатів засновані як на синтаксичному, так і на лексичному підходах. Показано, що за відповідної модифікації ці методи можуть бути використані для автоматизації перевірки відповідей на тестові завдання у відкритій формі.

Ключові слова: нечіткий дублікат, пошук, тестування, перевірка відповіді.

The main methods of near-duplicates searching based on both syntactic and the lexical approach are reviewed. It is shown that by appropriate modification of these methods, they can be used to automate the checking of quiz answers in an open form.

Key words: near-duplicate, search, quiz, answer checking.

Вступ

Одним з важливих етапів навчального процесу є вимірювання та порівняння навчальних досягнень студентів з метою забезпечення належної якості освіти. Педагогічна тестологія, як наука про тести, покликана займатися розробленням тестів для об'єктивного контролю рівня знань, тих, хто навчається [1]. Ключовими поняттями тестології, як однієї з методичних теорій, є тест, зміст і форма завдань, надійність і валідність результатів виміру [1, 2].

Традиційний тест являє собою стандартизований метод діагностики рівня і структури підготовленості. Такий тест визначається як система завдань визначеного змісту, зростаючої складності, специфічної форми, що дозволяє якісно й ефективно встановити рівень і оцінити структуру підготовленості учнів [1–3]. Завдання в тестовій формі характеризується як педагогічний засіб не тільки контролю рівня підготовленості, а й навчання і розвитку особистості [4].

Розвиток комп'ютерних технологій та інформатизації усіх сфер діяльності суспільства спричинив появу систем автоматизованого контролю знань осіб, котрі навчаються. Такі системи можуть бути інтегрованими з системою дистанційного навчання (наприклад Moodle [5]), або ж бути окремими засобами інформаційного забезпечення навчального процесу (наприклад, система OpenTest [6]). Такі засоби дозволяють значно полегшити проведення та оцінювання тестування, однак важливий клас тестових питань – питання у відкритій формі [5] – все ще потребує, щоб викладач оцінював, а це значно знижує ефективність автоматизації проведення тестування.

Разом з тим існують добре розроблені алгоритми пошуку подібних документів, які використовуються, починаючи від пошуку подібних файлів у файловій системі [7], до пошуку подібних документів в мережі Інтернет [8]. Зазвичай дублікати документів визначаються на основі відношення подібності на парах документів: два документи подібні, якщо деяка числова міра їх схожості перевищує деякий поріг.

За аналогією, можна вважати, що процес перевірки тестових завдань у відкритій формі якраз і полягає у пошуку “плагіату”, і чим більший ступінь подібності відповіді до еталонної, тим вище вона оцінюється. (Цілком очевидно, що повинна існувати верхня межа подібності, у разі

перевищення якої з'являється підозра про цілковите списування з еталона – підручника, лекції тощо.) Таким чином метою цієї роботи є огляд та аналіз можливості використання алгоритмів пошуку нечітких дублікатів для автоматизованої перевірки тестових завдань у відкритій формі.

Проблема виявлення нечітких дублікатів

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 та 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 у разі незаконного копіювання інформації тощо.

Основною перешкодою для успішного розв'язання цієї задачі є гігантський обсяг даних, що зберігаються в базах сучасних пошукових машин. Такий обсяг робить практично неможливим (за розумний час) її "пряме" розв'язання, шляхом попарного порівняння текстів документів. Тому останнім часом велика увага приділяється розробленню методів зниження обчислювальної складності створюваних алгоритмів за рахунок вибору різних евристик (наприклад, хешування певного фіксованого набору "значущих" слів або речень документа, семплювання набору підрядків тексту, використання дактилограм та ін.) [8–11].

У разі застосування наближених підходів спостерігається зменшення (іноді вельми значне) показника повноти виявлення дублів. Важливим фактором, що впливає на точність і повноту визначення дублікатів, є виділення змістовної частини за допомогою надійного розпізнавання елементів оформлення документів та їх подальшого видалення [12].

По відношенню подібності обчислюються кластери подібних документів. Спочатку, після зняття HTML-розмітки, документи, як лінійні послідовності слів (символів), перетворюються у множини. Тут двома основними схемами (що визначають весь можливий спектр змішаних методів) є синтаксичний і лексичний метод. До синтаксичного належить метод шинглювання [9], в якому документ у результаті представляється набором хеш-кодів. Цей метод знайшов застосування у таких популярних пошукових системах як Google і AltaVista. У лексичних методах [10] велика увага приділяється побудові словника – набору дескриптивних слів, відомими його різновидами є такі, як I-match і метод ключових слів Іллінського [10]. На другому етапі з документа, представленого множиною синтаксичних або лексичних ознак, вибирається підмножина ознак, що утворить короткий опис (образ) документа. На третьому етапі визначається відношення подібності на документах, за допомогою деякої метрики схожості, яка зіставляє двом документам число в інтервалі [0, 1], і деякого параметра – порогу, вище від якого знаходяться документи дублікати. На основі відношення подібності документи об'єднуються в кластери нечітких дублікатів. Визначення кластера також може змінюватись [12].

Синтаксичні методи виявлення нечітких дублікатів

Одними з перших досліджень в області знаходження нечітких дублікатів є роботи Манбера (Manber) [7] та Хайнтце (Heintze) [13]. У цих роботах для побудови вибірки використовуються пост-послідовності сусідніх букв. Дактилограма файла [7] або документа [13] включає всі текстові підрядки фіксованої довжини. Чисельне значення дактилограм обчислюється за допомогою алгоритму випадкових поліномів Карпа–Рабіна [14]. Критерієм схожості двох документів використовують відношення числа спільних підрядків до розміру файла або документа. Манбер використовував цей підхід для знаходження схожих файлів (утиліта sif), а Хайнтце – для виявлення нечітких дублікатів документів (система Koala).

У 1997 р. Бродер (Broder) та ін. [12, 15] запропонували новий синтаксичний метод оцінювання подібності між документами, заснований на зображенні документа у вигляді множини послідовностей фіксованої довжини k, які складаються з сусідніх слів. Такі послідовності були названі "шинглами" ("гонт", англ. shingle). Два документа вважалися схожими, якщо їх множини "шинглів" істотно перетиналися. Оскільки кількість "шинглів" приблизно дорівнює довжині

документа в словах, тобто є достатньо великою, автори [12, 15] запропонували два методи семплювання для отримання репрезентативних підмножин.

Перший метод залишав тільки ті “шингли”, чиї дактилограми, які обчислюють за алгоритмом Карпа–Рабіна [14], 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 [14] за допомогою взаємно-однозначних і незалежних функцій. Потім 84 “шингли” розбиваються на 6 груп по 14 (незалежних) “шинглів” у кожній. Ці групи називаються “супершинглами”.

Для ефективної перевірки збігу не менше двох “супершинглів” (і відповідно, підтвердження гіпотези про подібність змісту) кожен документ представляється усіма можливими попарними поєднаннями з 6 “супершинглів”, які називаються “мегашинглами”. Два документи подібні за змістом, якщо у них збігається хоча б один “мегашингл”.

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

Результати досліджень, проведених за сприяння компанії Яндекс [16], показали, що методи породження приватних замкнених множин є ефективним способом визначення подібності документів одночасно з породженням кластерів подібних документів. На результати синтаксичних методів виявлення дублікатів значний вплив здійснює параметр “довжина шингла”, а значного впливу використання методу “мінімальні елементи в *n* перестановках” на якість результатів виявлено не було [16].

В [16] показано необхідність порівняння методів кластеризації, що використовують замкнені множини ознак з алгоритмами, заснованими на пошуку мінімальних розрізів вершин дводольних графів, в яких множини вершин відповідають множинам документів та множинам ознак [17, 18]. Ці методи є спорідненими, оскільки замкнені множини документів виражаються через мінімальні розрізи такого роду дводольних графів.

Лексичні методи виявлення нечітких дублікатів

Інший сигнатурний підхід, заснований вже не на синтаксичних, а на лексичних принципах, запропонував Чоудурі (Chowdhury) та ін. в 2002 р. і удосконалений у 2004 р. [19, 20]. Основна ідея такого підходу полягає в обчисленні дактилограми I-Match (хеш-функції SHA-1) на основі перетину словника колекції документів та множини різних слів документа для зображення 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 документа. Для подолання вказаного недоліку вихідний алгоритм був модифікований [20].

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, який запропонував С. Ільїнський та ін. [10]. Цей метод дозволяє, почавши з вибірки в сотні тисяч слів, залишити набір з 3–5 тисяч, розрахунок сигнатур по яких з застосуванням повнотекстового індексу здійснюється надзвичайно швидко та ефективно [10].

Висновки

У роботі розглянуто основні методи пошуку нечітких дублікатів, засновані як на синтаксичному, так і на лексичному підходах. Показано, що основні модифікації алгоритмів здійснюються на третьому етапі методу пошуку нечітких дублікатів – кластеризації та пошуку

відношення подібності документів. Проведено аналогії між пошуком нечітких дублікатів у мережі Інтернет та перевіркою тестових завдань у відкритій формі (типу “ессе”) та показано принципову можливість використання розглянутих методів для автоматизації перевірки відповідей на тестові завдання у відкритій формі шляхом порівняння відповіді з еталонною. Подальші дослідження будуть спрямовані на встановлення впливу параметрів синтаксичних алгоритмів, зокрема алгоритмів на основі “шинглів” на ефективність пошуку нечітких дублікатів та адаптації цих алгоритмів до задачі автоматизованої перевірки тестових питань у відкритій формі.

1. Аванесов В. С. Научные проблемы тестового контроля знаний. – М. : Исслед. центр, 1994. – 135 с. 2. Челышкова Н. Б. Теория и практика конструирования педагогических тестов : учеб. пособие / Н. Б. Челышкова. – М. : Логос, 2002. – 432 с. 3. Майоров А. Н. Теория и практика создания тестов для системы образования. (Как выбирать, создавать и использовать тесты для целей образования) / А. Н. Майоров. – М. : Интеллект-центр, 2001. – 296 с. 4. Ингекамп К. Педагогическая диагностика / К. Ингекамп. – М. : Педагогика, 1991. – 240 с. 5. Яковина В.С. Методи та засоби організації тестування у віртуальному навчальному середовищі Львівської політехніки // Вісник Нац. ун-ту “Львівська політехніка” Інформатизація вищого навчального закладу. – № 703 (2011). – С. 65–68. 6. Шкіль О.С. Єдиний тестовий сервер вищого навчального закладу як основа якісного проведення тестувань // Вісник Нац. ун-ту “Львівська політехніка” Інформатизація вищого навчального закладу. – № 703 (2011). – С. 54–59. 7. U. Manber. Finding Similar Files in a Large File System. // WTEC'94 Proceedings of the USENIX Winter 1994 Technical Conference, p. 2. 8. W. Pugh, M.H. Henzinger Detecting duplicate and near-duplicate files // US Patent 6658423 (2003). 9. A. Broder, Identifying and Filtering Near-Duplicate Documents // LNCS, Vol. 1848 (2000). – pp. 1–10. 10. 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). 11. G. Grahne and J. Zhu, Efficiently Using Prefix-trees in Mining Frequent Itemsets // Proceedings of FIMI'03 Workshop on Frequent Itemset Mining Implementations, 2003. – pp. 125–134. 12. 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. 13. N. Heintze. Scalable document fingerprinting. // Proc USENIX Workshop on Electronic Commerce (1996). – pp. 191–200. 14. Д. Гасфилд. Строки, деревья и последовательности в алгоритмах. – СПб.: Невский диалект, 2003. – 656 с. 15. A. Broder. On the resemblance and containment of documents. // Proceedings of the Compression and Complexity of Sequences 1997. – pp. 21–29. 16. Игнатов Д.И., Кузнецов С.О. О поиске сходства Интернет-документов с помощью частых замкнутых множеств признаков // Труды 10-й национальной конференции по искусственному интеллекту с международным участием (КИИ’06). – М.:Физматлит, 2006, Т.2. – С.249–258. 17. I.S. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning // In Knowledge Discovery and Data Mining. – pp. 269-274, 2001. 18. Y.Zhao and G. Karypis, Empirical and Theoretical Comparison of Selected Criterion Functions for Document Clustering // Machine Learning, Vol. 55 (2004). – pp. 311–331. 19. 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. 20. 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.