Алгоритм перевірки тестових завдань на основі синтаксичного методу

Віталій Яковина[1], Тетяна Смірнова[2]

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

2. Кафедра напівпровідникової електроніки, Національний університет "Львівська політехніка", УКРАЇНА, м.Львів, вул. С.Бандери, 12

The description of the near-duplicate revealing algorithm is presented in this paper. It is proposed to use the algorithm for open quiz tasks checking.

Ключові слова – тестування, нечіткий дублікат, алгоритм, синтаксичний метод.

I. Вступ

Тестування як форма контролю та оцінювання рівня знань та умінь широко застосовується в педагогічній діяльності [1]. Разом з тим при використанні комп'ютерного тестування знань перевірка відповідей у відкритій формі (без яких неможливе повноцінне оцінювання набутих компетенцій особистості) і надалі здійснюється вручну. У попередній роботі [2] автори здійснили огляд та аналіз методів виявлення нечітких дублікатів з метою використання їх в автоматизованій системі перевірки тестових завдань. Ця стаття присвячена опису алгоритму знаходження нечітких дублікатів, який пропонується використати в такій автоматизованій системі. Цей алгоритм базується на синтаксичному методі "шинглів" (англ. shingle – ґонт, черепиця) [3].

II.Опис алгоритму

Відповідь студента на питання у відкритій формі порівнюється з еталонною відповіддю. 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, а відповідь є правильною. 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. Узагальнена блок-схема такого алгоритму навдена на рис. 1.

Принцип алгоритму шинглів полягає в порівнянні випадкової вибірки контрольних сум "шинглів" (підпослідовностей) двох текстів між собою. Збільшення кількості шинглів для порівняння характеризується ростом операцій, що критично позначиться на продуктивності.

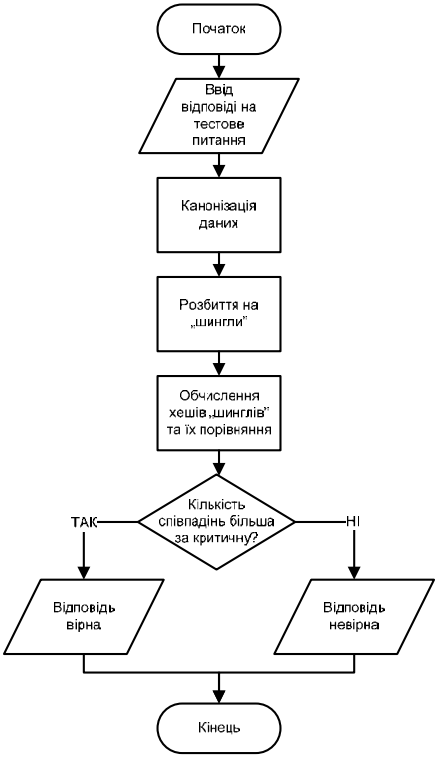

Рис. 1. Блок-схема алгоритму знаходження нечітких дублікатів.

Канонізація тексту призводить оригінальний текст до єдиної нормальної форми. Текст очищається від прийменників, сполучників, розділових знаків, HTML тегів і т.ін. Так само на етапі канонізації тексту можна приводити іменники до називного відмінку, однини, або залишати від них тільки корінь.

Висновок

У роботі наведено опис алгоритму знаходження нечітких дублікатів, яких пропонується використати для автоматизованої системи перевірки тестових завдань.

Література

1. Аванесов В. С. Научные проблемы тестового контроля знаний / В. С. Аванесов. – М. : Исслед. центр, 1994. – 135 с.
2. Яковина В.С., Камінський Р.І., Смірнов В.О. Огляд методів виявлення нечітких дублікатів для автоматизованої перевірки тестових завдань // Матеріали V Міжнародної конференції молодих вчених "Комп'ютерні науки та інженерія" CSE-2011, Львів, 2011, Р. 366–367.
3. 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.