DOI: 10.34185/1991-7848.itmm.2020.01.049

ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МОДЕЛІ ГРАФОВОГО ПРЕДСТАВЛЕННЯ ТЕКСТІВ ДЛЯ ВИЯВЛЕННЯ ЗАПОЗИЧЕНЬ

Куроп'ятник О. С.

Дніпровський національний університет залізничного транспорту імені академіка В. Лазаряна, Україна

Ключові слова: ВИЯВЛЕННЯ ТЕКСТОВИХ ЗАПОЗИЧЕНЬ, АНТИПЛАГІАТ, КОНСТРУКТИВНО-ПРОДУКЦІЙНЕ МОДЕЛЮВАННЯ, ЧАСОВА ЕФЕКТИВНІСТЬ, ФУНКЦІОНАЛЬНА ЕФЕКТИВНІСТЬ.

Вступ. Для виявлення плагіату, а в більш широкому смислі – запозичень, все частіше застосовуються технічні засоби – системи антиплагіату. Їх основна функція – порівняння тексту з базою та обчислення відсотку унікальності або запозичень.

Головними вимогами до роботи таких систем є точність ідентифікації запозичень, відсутність хибних спрацювань, швидкість, можливість корегування параметрів роботи.

Для оцінки ефективності роботи антиплагіату можуть бути обчислені часова та функціональна ефективність та проведено їх аналіз за SR-показниками [1].

В роботі представлено результати розробки системи виявлення запозичень на основі конструкутивно-продукційної моделі (КПМ) графового представлення тексту та дослідження показників її часової та функціональної ефективності.

Основний матеріал. На основі КПМ графового представлення тексту [2] розроблено програмне забезпечення (ПЗ) для виявлення запозичень. Модель є конструктором, що включає розширюваний носій, сигнатуру операцій та інформаційне забезпечення конструювання (ІЗК). ІЗК визначає онтологію носія, мету, правила, умови початку та завершення конструювання.

Основна ідея моделювання полягає у формуванні орієнтованого графа із навантаженими вершинами та дугами. Вага вершини – символ тексту; вага дуги – множина номерів циклів у графі, яким належить дана дуга.

ПЗ реалізує побудову графа відповідно алгоритму, який складається з таких кроків:

1. створити стартову вершини (СВ) з навантаження, рівним першому символу тексту. Стартову вершину позначити як поточну (ПВ).

2. переглянути всі символи тексту. Для кожного символу виконати одну з послідовностей дій:

2.1. якщо символ співпадає з навантаженням СВ, додати у список суміжності ПВ стартову, вказати для неї номер маршруту рівний кількості циклів у графі, збільшити кількість циклів на один, позначити СВ як поточну;

2.2. якщо символ співпадає з навантаженням суміжної вершини, додати у список суміжності ПВ суміжну вершину, вказати для неї номер маршруту рівний кількості циклів у графі, позначити суміжну вершину як ПВ;

2.3. якщо символ не співпадає з навантаженням СВ і суміжних вершин, створити нову вершину, додати її у список суміжності ПВ, нову вершину позначити як ПВ.

Для зменшення об’єму пам’яті, необхідного для обробки та збереження графів, може бути виконано їх стиснення, в результаті якого навантаження вершин можуть складати декілька символів. Для тексту будується кілька графових представлень: одне відповідає всьому тексту, а інші – фрагментам, що починаються з різних слів, перші букви яких є унікальними в тексті. Таким чином для тексту утворюється набір графів.

Розроблене ПЗ на основі даної моделі є додатком для виявлення запозичень у текстових фрагментах [2] та системою виявлення запозичень у структурованих документах.

Проведено експериментальні дослідження ефективності використання КПМ графового представлення текстів для виявлення запозичень.

Комп’ютерні експерименти для дослідження часової ефективності операції побудови графового представлення тексту дозволили встановити поліноміальну, близьку до лінійної залежність показників часу від розміру вхідного тексту. Експериментально визначено час побудови графового представлення текстів розділів структурованих документів, який у середньому склав близько 62 секунд для 152 тис. символів (для кваліфікаційних робіт ОС Бакалавр спеціальності 121 «Інженерія ПЗ»).

Визначено значення показників часової ефективності операції зіставлення текстових фрагментів і структурованих документів на основі графового представлення, її складових та характер залежності показників від розміру тексту у символах. Операція зіставлення має чотири складові:

попередня обробка, побудова (отримання) графового представлення, порівняння та оцінка результату.

Часова ефективність операції порівняння текстових фрагментів (до 14 тис. символів) один до одного не перевищує 0.2 секунди, та залежить від розмірів фрагментів та структури графа.

Встановлено лінійну залежність часової ефективності операції зіставлення структурованих документів (текстів) від їх розміру та розміру бази та близьку до лінійної – за складовим операції. Аналіз складових операції зіставлення показав, що основний час (близько 94%) витрачається на отримання наборів графів. Середній час зіставлення структурованого документу (середній розмір 172 тис. символів) зі всіма наявними у базі роботами складає від 11 до 65 секунд при базі від 0,6 до 3,8 млн. символів.

Всі експерименти виконано на ПК з такими технічними характеристиками: процесор Intel Pentium(R) Dual Core CPU, кеш L1 коду/ L1 даних/ L2 – 2*32/2*32/1024 Кб, тактова частота/частота системної шини/частота пам'яті – 2,3 ГГц /400 МГц/400 МГц, час доступу до ОП (читання/запис) 5751/4253 Мб/с, операційна система – MS Windows 7 Ultimate SP1. Для збереження графових наборів використано локальний веб-сервер xampp v 3.2.2 (компоненти Apache, MariaDB).

Досліджено функціональну ефективність використання КПМ графового представлення тексту для виявлення запозичень. Під функціональною ефективністю будемо розуміти здатність програми виявляти запозичення, що може бути виміряна в кількості знайдених запозичених мовних одиниць та відсотку запозичень. Під мовною одиницею будемо розуміти ланцюжок символів мінімальної довжини, який буде враховано при підрахунку кількості та відсотку запозичень.

Для визначення функціональної ефективності необхідно виконати зіставлення N-текстів та порівняти їх з результатами ручної перевірки на наявність запозичень або аналогічних програм-антиплагіатів.

Проведений аналіз аналогів дозволив виявити перешкоди у їх використанні для експериментальних досліджень ефективності. Умовно їх можна поділити на дві групи:

– відсутність можливості: використання однакових баз для зіставлення та/або подання файлів і пакетів; формування загального файлу результатів

роботи для їх подальшого аналізу; завдання мінімальної довжини фрагменту, який буде вважатися запозиченням;

– відмінності понять та методів: визначення поняття «слово», на якому базується поняття запозиченого фрагменту; різні методи попередньої обробки, що призводить до зміни кількості слів та інших мовних одиниць, що в подальшому має вплив на відсоток виявлених запозичень; різні методи врахування повторюваних фраз.

Для дослідження функціональної ефективності розробленого програмного забезпечення було використано програму WCopyfind [3], у вихідний код якої було додано функціонал для формування загального файлу результатів порівняння.

Експериментальна база. Перша серія: 64 створених текстових файлів у форматі docx за тематикою «Розробка ПЗ» за матеріалами Wikipedia (розміром від 16 – 24 Кб, від 2 – 14 тис. символів). Друга серія: 40 текстових файлів у форматі docx, які є технічними завданнями до розробки програмного забезпечення (розміром від 48 – 290 Кб, від 18 – 28 тис. символів). Документи мають структуру, позначену форматуванням за рівнями, автозміст, формули та таблиці, текстові поля.

Проведені експерименти показали, що існує розбіжність у результатах роботи розробленої програми та аналогу. Для її оцінки обчислено ступінь переваги ефективності роботи за аналогією до S-оцінки ефективності алгоритмів [1]. Сумарно для двох серій експериментів виконано понад 4.5 тис. порівнянь – S не перевищує 5%.

Аналіз текстів та виявлених запозичених фрагментів показав, що розбіжності зумовлені:

– різною інтерпретацією поняття слова;

– різницею у кількості слів в документах після попередньої обробки аналогом та розробленим ПЗ;

– впливом форматування. Так, наприклад наявність посилань забезпечує додаткові слова для аналога.

Висновки. Експериментальні дослідження ефективності КПМ графового представлення текстів, яка покладено в основу роботи системи виявлення запозичень, дозволили:

– встановити лінійну залежність часу операції зіставлення документів від їх розмірів та визначити, що близько 94% часу витрачається на отримання графів;

– виявити ряд факторів, що є перешкодою для порівняння ефективності з аналогами, та причини відмінності результатів роботи програм.

Отже, доцільним є подальша робота над покращенням механізмів збереження та відтворення графів та оптимізацією відповідного програмного коду, а також дослідження методів попередньої обробки текстів та поняття «слово», що є головним фактором впливу на розбіжність результатів.

Література

1. Шинкаренко В. И. Экспериментальные исследования алгоритмов в программно-апааратных средах [Текст]: монография. – Д.: Изд-во Днепропетр. нац. ун-та ж.-д. трансп. им. акад. В. Лазаряна. – 2009. – 275 с.

2. Куроп'ятник О. С. Конструктивне та об'єктно-орієнтоване моделювання текстів для виявлення запозичень //Системні технології. – 2019. – №. 4. – С. 34-47.

3. The Plagiarism Resource Site. WCopyfind [Електронний ресурс] – Режим доступу: https://plagiarism.bloomfieldmedia.com/software/wcopyfind/. – Заг. з екрану. – Перевірено: 12.02.2020.

EXPERIMENTAL INVESTIGATIONS OF TEXT GRAPH REPRESENTATION MODEL EFFICIENCY FOR BORROWINGS DETECTION

Olena Kuropiatnyk

Abstract. The paper deals with investigation of time and functional efficiency of the developed software system for text borrowings detection. Base of this system is constructive-synthesizing text graph representation model. The experiment revealed a linear relationship between the time of the text borrowing check operation and the size of the text base for comparison. The conducted experiments showed that there is a difference in the results of the checking text documents by the developed system and analogue. For its estimation the degree of functional efficiency advantage is calculated by similarly to S-estimation of efficiency of algorithms. The reasons for the difference in results are identified. Attention is drawn to obstacles in analogues use for experimental efficacy investigations.

Keywords: TEXT BORROWING DETECTION, ANTI-PLAGIARISM, CONSTRUCTION-SYNTHESIZING MODELING, TIME EFFICIENCY, FUNCTIONAL EFFICIENCY.

References

1. Shynkarenko V. I. Experimental investigations of algorithms in software and hardware environments [Text]: monograph. – D.: Publishing house Dnepropetr. Nat. University of Railway transp. named after Acad. V. Lazaryan. - 2009. – 275 p.

2. Kuropiatnyk O. Constructive and object-oriented modeling text for detection of text borrowings. System technologies, no. 4 (123), pp. 34–47 (2019). doi: 10.34185/1562-9945-4-123-2019-04

3. The Plagiarism Resource Site. WCopyfind [online] – Access mode: https://plagiarism.bloomfieldmedia.com/software/wcopyfind/. – Title from screen. – Checked: 12.02.2020.