УДК 681.31

К.т.н., доцент Тесленко О.К., к.т.н., доцент Замятін Д.С., магістрант Подолян О.Д.

Національний технічний університет України «Київський політехнічний інститут»

ДОСЛІДЖЕННЯ АДЕКВАТНОСТІ АНАЛІЗУ ЗАПОЗИЧЕНЬ В ІНФОРМАЦІЙНИХ ОБ'ЄКТАХ В ЗАЛЕЖНОСТІ ВІД ПАРАМЕТРІВ ІЄРАРХІЧНОГО АДАПТИВНОГО ПОРІВНЯННЯ

Abstract

Oleksandr C. Teslenko, assoc. prof., PhD; Denis S. Zamyatin, assoc. prof., PhD; Oleksii Podolian, student

Study of adequacy of data entities' originality analysis in relation to hierarchical adaptive comparison parameters

This paper studies the impact of hierarchical adaptive comparison parameters on the results of data entities' analysis. The hierarchical adaptive comparison algorithm is described briefly and is used as a basis for the research. Optimal set of parameters for texts of the same kind as analysed is proposed. The ways for further research are suggested as well.

Вступ

Масове застосування комп'ютерних інформаційних технологій поряд із багатьма позитивними рисами несе в собі і додаткові проблеми, до яких належать значно розширені можливості використання несанкціонованих запозичень в інформаційних об'єктах. В суспільній практиці це призводить до розширення можливостей порушень авторських прав, а в освіті – до погіршення якості навчання.

Одним із перспективних методів виявлення запозичень в інформаційних об'єктах є метод ієрархічного адаптивного порівняння [1]. Суть методу полягає в наступному. В двох послідовностях символів, що порівнюються, шукають однакові підпослідовності довжиною не менше d символів. Якщо дві такі підпослідовності знайдені, то вони вважаються збігом і беруться до уваги. Для пошуку цих підпослідовностей із першої (A) послідовності з кроком x, а з другої (B) – з кроком y обираються підпослідовності символів довжиною c. Параметри обираються так, щоб $x \leq d - c + 1$. Це гарантує, що будь-яка підпослідовність довжиною в d символів з A цілком міститиме принаймні одну підпослідовність в c символів з B, якщо вони обиралися з кроком $y = 1$. Значення параметру c

визначається розрядністю даних в командах порівняння процесора. При використанні спеціалізованих пристроїв для визначення факту входження підпослідовності довжиною c символів в підпослідовність із c^* символів, де $c^* > c$, значення параметру y може бути більшим за 1.

На нижньому рівні ієрархії інформаційний об'єкт розглядається як неструктурована послідовність символів певного алфавіту. Більш високі рівні ієрархії визначаються типом інформаційного об'єкта. Збіги елементів більш високого рівня ієрархії визначаються результатом виявленого збігу (наприклад, в процентному відношенні) на попередньому рівні ієрархії.

Постановка задачі

Метою роботи є дослідження та аналіз методу ієрархічного адаптивного порівняння при різних наборах параметрів на спеціально підготовлених текстах для визначення оптимальних параметрів з точки зору адекватності результатів.

Методика проведення дослідження

Дане дослідження проводилось тільки для текстових даних як таких, для яких характерне широке використання запозичень, а також можлива відносно проста програмна реалізація технології.

Для таких даних в загальному випадку було виділено три рівні ієрархії: нижній оперує окремими символами; середній, на якому розглядаються речення; верхній, який оперує окремими абзацами.

Для виконання дослідження було створено програму, яка реалізує даний алгоритм та дозволяє змінювати значення необхідних параметрів. Вона має три режими роботи: «без розбиття тексту», «з розбиттям на речення» та «з розбиттям на абзаци». Дослідження проводилось для всіх трьох режимів.

У зв'язку з цим введено два додаткових параметри: «відсоток збігів для абзаців» (позначимо p) та «відсоток збігів для речень» (s). Вони характеризують величину збігу між двома фрагментами відповідного рівня ієрархії, при якому вони вважаються співпадаючими, тобто фіксується запозичення.

У даній роботі досліджено вплив параметрів с та d на результати роботи в режимі без розбиття тексту, а також в режимах з урахуванням ієрархічної структури документа при різних значеннях параметрів p та s.

Для автоматизації дослідження реалізовано програмний режим, який автоматично порівнює два задані файли при різних значеннях параметрів з

заданого діапазону та з заданим кроком. Це дозволяє зосередити зусилля на аналізі результатів та мінімізувати участь людини в їх генерації. Усі кінцеві данні зберігаються в форматі електронної таблиці, що дає широкі можливості для їх аналізу, наприклад, засобами фільтрації ПЗ для роботи з такими даними.

Дослідження проводилось на двох файлах. Перший є типовим прикладом студентського реферату, а другий був отриманий шляхом злиття першого файлу з іншим текстом. При цьому від першого файлу було залишено лише певний відсоток, який і вважається коректним результатом роботи програми. Деякі абзаци були об'єднані, інші розбиті на декілька нових, також частково змінено порядок слідування. В такий спосіб була промодельована ситуація швидкої зміни тексту без вникання у суть з метою усунення візуальної подібності до оригінала.

Результати дослідження

Перший текст, який був використаний для дослідження, містить 8254 символів. Другий текст містить 8140 символів, з них – 4596 з першого тексту. Таким чином, рівень запозичення складає 56,5%.

Дослідження проводилось при зміні параметрів: *c* – значення з набору {1, 2, 4, 8}, *d* змінювалось від 10 до 70 з кроком 2, *p* та *s* приймали значення від 50 до 100, з кроком 5.

Всього було проведено 2852 тестів. В режимі без розбиття тексту результати змінювались в діапазоні від 56% до 62%, при розбитті на параграфи результати були з набору {18, 28, 37, 59}, при розбитті на речення – з набору {56, 57}.

Аналіз отриманих результатів

Отримані результати демонструють відсутність впливу параметра *c* на результати порівняння у всіх режимах. Це цілком очікувано і пояснюється характером формування вхідних текстів. Зміни було внесено на великих порівняно із значеннями *c* і *d* відстанях, що нівелює вплив цих параметрів на якість порівняння. Тому можна рекомендувати обирати доволі велике значення *d*. Хибні результати, як то 62%, отримані саме при *d = 10*, що теж очікувано, адже такий короткий фрагмент може співпасти в різних текстах випадково. Значення *c* на практиці визначається апаратурою.

При розбитті тексту на абзаци адекватні результати отримано лише при *p =50*. Цей режим дуже чутливий до розбиття абзаців на декілька нових.

Найкращі результати на використовуваних даних дав режим розбиття на речення. Наявність двох фіксованих значень результату пояснюється тим, що блоки, коротші за d в режимах з розбиттям не розглядаються. Тому для великих d (більше 64) результат склав на один відсоток більше.

Перспективи практичних застосувань

Отримані результати дозволяють застосовувати алгоритм ієрархічного адаптивного порівняння на практиці, використовуючи значення параметрів, які є оптимальними, або близькими до таких з точки зору якості результатів та часу роботи. Розроблена програма може бути використана як для практичного застосування для контролю оригінальності текстів, так і для подальших досліджень методики, що реалізована.

Висновки

У ході дослідження було отримано практичні результати роботи алгоритму для широкого діапазону вхідних параметрів. Отримані результати не розбігаються із теоретично очікуваними, натомість конкретизують їх. В подальшому доцільно дослідити вплив параметрів алгоритму та режимів роботи програми (наприклад, рівень ієрархії, що використовується) на якість результатів використання методу та його швидкодію. Доцільно також провести дослідження для текстів з різних предметних областей та текстів з різним характером внесених змін.

Література

1. *В.П.Тарасенко, А.Ю.Михайлюк, О.К.Тесленко, О.С.Осипов.* Автоматизація оцінки оригінальності інформації // Наукові записки українського науково-дослідного інституту зв’язку. - 2007.- №1. -С. 95-100.