ЕФЕКТИВНИЙ ПОШУК НАБЛИЖЕНИХ ПІДРЯДКІВ ДЛЯ МАЛЕНЬКИХ АЛФАВІТІВ

Вечур О. В., Насонов Є. О.
Харківський національний університет радіоелектроніки, Харків, Україна

Наближений пошук підрядків є прикладною задачею та постає у багатьох задачах, таких як фільтрація спаму, антивірусні застосування, біоінформатичні задачі, пошук у базах даних, тощо. Задача має багато різних постановок та математичних формалізацій, кожна з яких краще описує конкретну задачу. Зокрема, одним з найважливіших аспектів визначення задачі є вибір метрики, за якою ми визначаємо, чи є рядки схожими. Досить часто ми маємо змогу працювати з моделлю, де розмір алфавіту дуже малий. Так, наприклад, задачі біоінформатики працюють з алфавітом розміру 4. Найперші алгоритми наближеного пошуку підрядків, які базуються на алгоритмі Бойера-Мура, працюють тим швидше, чим більше алфавіт [1]. Було досліджено інші алгоритми, які використовують експоненційний від розміру алфавиту та кількості допустимих помилок попередній підрахунок табличних значень [2]. Такі підходи складно узагальнити для пошуку багатьох підрядків у великому рядку, та вони не використовують сучасні можливості паралелізму виконання програм.

Метою доповіді є розробка швидкого алгоритму для наближеного пошуку підрядків для задач з маленьким алфавітом, де за відстань між рядками в першу чергу береться відстань Геммінга. Пропонується узагальнення алгоритму для відстані Левенштейна.

В доповіді визначається задача наближеного пошуку підрядків за відстанню Геммінга та наводиться алгоритм на основі підходу multi-Volnitsky. Наведено порівняльні графіки з іншими алгоритмами, які вирішують дану задачу, що доказує швидкіть алгоритму, хоча він і використовує більше пам’яті та часу на попередній підрахунок даних.

Також доповідається про перевагу алгоритму, яка дозволяє шукати водночас багато підрядків, що неможливо з використанням алгоритмів на основі Бойера-Мура, та перевагу можливості ефективної реалізації паралельної версії алгоритму. Розглядається узагальнення алгоритму для відстані Левенштейна та проблематика зростання часової та просторової складності такого узагальнення.

Список літератури

1. Tarhio, Jorma & Ukkonen, Esko. (1993). Approximate Boyer–Moore String Matching. SIAM J. Comput.. 22. 243-260. 10.1137/0222018.

2. Salmela, Leena & Tarhio, Jorma & Kalsi, Petri. (2010). Approximate Boyer-Moore String Matching for Small Alphabets. Algorithmica. 58. 591-609. 10.1007/s00453-009-9286-3.