Поиск абелевых строк наибольшей длины

И. Збань
Научный руководитель: В. Аксёнов

УНИВЕРСИТЕТ ИТМО

5 июня 2017 г.

Постановка задачи

Задача: Нахождение наибольшей общей абелевой подстроки и поиск абелевых подквадратов.

Постановка задачи

Задача: Нахождение наибольшей общей абелевой подстроки и поиск абелевых подквадратов.

Мотивация:

Постановка задачи

Задача: Нахождение наибольшей общей абелевой подстроки и поиск абелевых подквадратов.

Мотивация:

- Быстроразвивающаяся область, много публикаций за последнее время

Постановка задачи

Задача: Нахождение наибольшей общей абелевой подстроки и поиск абелевых подквадратов.

Мотивация:

- Быстроразвивающаяся область, много публикаций за последнее время
- Актуальность: подзадачи в бионформатике (gene clusters), фильтры в задаче поиска образца

Постановка задачи

Задача: Нахождение наибольшей общей абелевой подстроки и поиск абелевых подквадратов.

Мотивация:

- Быстроразвивающаяся область, много публикаций за последнее время
- Актуальность: подзадачи в бионформатике (gene clusters), фильтры в задаче поиска образца
- Близость с известной задачей 3SUM

Краткое описание

Работа состоит из следующих пунктов:

Краткое описание

Работа состоит из следующих пунктов:

- Оценка алгоритма решения 3SUM+ для монотонных множеств на примере задачи о количестве абелевых подквадратов

Краткое описание

Работа состоит из следующих пунктов:

- Оценка алгоритма решения 3SUM+ для монотонных множеств на примере задачи о количестве абелевых подквадратов
- Анализ задачи LCAF для частного случая бинарного алфавита

Краткое описание

Работа состоит из следующих пунктов:

- Оценка алгоритма решения 3SUM+ для монотонных множеств на примере задачи о количестве абелевых подквадратов
- Анализ задачи LCAF для частного случая бинарного алфавита
- Решение задачи LCAF для общего случая

Количество абелевых подквадратов

Задача о количестве абелевых подквадратов сводится к 3SUM+

Количество абелевых подквадратов

Задача о количестве абелевых подквадратов сводится к 3SUM+

A=B=$\{(c_a(i),\ c_b(i))\}$, C=$\{2c_a(i),2c_b(i)\}$

где $c_a(i), c_b(i)$ — количество букв a и b на префиксе длины i

Количество абелевых подквадратов

Задача о количестве абелевых подквадратов сводится к 3SUM+

A=B=$\{(c_a(i), c_b(i))\}$, C=$\{2c_a(i), 2c_b(i)\}$

где $c_a(i), c_b(i)$ — количество букв a и b на префиксе длины i

и число подстрок — $(\#3SUM^{+}$(A,B,C)-(n+1))/2

Сравнение алгоритмов на простой строке

Картиночка на которой видно, что квадрат работает быстро, а 1.86 — медленно

Сравнение алгоритмов на случайном тесте

Картиночка, на которой видно, что квадрат работает быстро, а 1.86 — ооооочень медленно

binary LCAF

Картинка с матожиданием LCAF случайных бинарных строк

Доказана оценка сверху, что LCAF ограничена линейной функцией, тем самым опровергнута посылка из первоисточника

general LCAF

Используя персистентные деревья с limited node copying предложен алгоритм LCAF в общем случае за $(\mathcal{O}(n^2 \log \sigma), \mathcal{O}(n))$

Идея — построить персистентный массив вектора Парея для строки-конкатенации обеих данных строк, посчитать некий хеш от каждого корня, и проверить, были ли одинаковые версии, соответствующие обеим строкам

general LCAF

Используя персистентные деревья с limited node copying предложен алгоритм LCAF в общем случае за $(\mathcal{O}(n^2 \log \sigma), \mathcal{O}(n))$

Идея — построить персистентный массив вектора Парея для строки-конкатенации обеих данных строк, посчитать некий хеш от каждого корня, и проверить, были ли одинаковые версии, соответствующие обеим строкам

Схема вычисления хеша

Тут какая-то непонятная картинка

Сравнение алгоритмов LCAF

Год	Авторы	Время	Память
2013	StringMasters	-	-
2015	Кто-то	$\mathcal{O}(n^2\sigma)$	$\mathcal{O}(n\sigma)$
2016	Кто-то	$\mathcal{O}(n^2\sigma)$	$\mathcal{O}(n)$
2016	SPIRE	$\mathcal{O}(n^2 \log^2 n \log^* n)$	$\mathcal{O}(n \log^2 n)$
2017	Я	$\mathcal{O}(n^2 \log \sigma)$	$\mathcal{O}(n)$

Вопросы?

Спасибо за внимание.