МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №4

по дисциплине «Построение и анализ алгоритмов»

Тема: Алгоритм КМП

Студент гр. 7304 _____ Абдульманов Э.М

Преподаватель _____ Филатов А.Ю

г. Санкт-Петербург
2019

Цель работы

Освоить алгоритм Кнута – Морриса – Пратта, разобрать работу префикс – функции, как примитивного вида, так и оптимизированного. Реализовать КМП на языке программирования с++.

Задачи

1. Реализуйте алгоритм КМП и с его помощью для заданных шаблона PP (|P|≤15000|P|≤15000) и текста TT (|T|≤5000000|T|≤5000000) найдите все вхождения PP в TT.
2. Заданы две строки AA (|A|≤5000000|A|≤5000000) и BB (|B|≤5000000|B|≤5000000).
   Определить, является ли A циклическим сдвигом B (это значит, что A и B имеют одинаковую длину и A состоит из суффикса B, склеенного с префиксом B). Например, defabc является циклическим сдвигом abcdef.

Теоретические сведения

Определение префикс функции:

Дана строка $s[0 \dots n-1]$. Требуется вычислить для неё префикс-функцию, т.е. массив чисел $\pi[0 \dots n-1]$, где $\pi[i]$ определяется следующим образом: это такая наибольшая длина наибольшего собственного суффикса подстроки $s[0 \dots i]$, совпадающего с её префиксом (собственный суффикс — значит не совпадающий со всей строкой). В частности, значение $\pi[0]$ полагается равным нулю.

Математически определение префикс-функции можно записать следующим образом:

$$\pi[i] = \max_{k=0 \dots i} \{\ k\ :\ s[0 \dots k-1] = s[i-k+1 \dots i]\ \}.$$

Например, для строки "abcabcd" префикс-функция равна: $[0, 0, 0, 1, 2, 3, 0]$, что означает:

- у строки "a" нет нетривиального префикса, совпадающего с суффиксом;
- у строки "ab" нет нетривиального префикса, совпадающего с суффиксом;
- у строки "abc" нет нетривиального префикса, совпадающего с суффиксом;
- у строки "abca" префикс длины 1 совпадает с суффиксом;
- у строки "abcab" префикс длины 2 совпадает с суффиксом;
- у строки "abcabc" префикс длины 3 совпадает с суффиксом;
- у строки "abcabcd" нет нетривиального префикса, совпадающего с суффиксом.

Ход работы

1. Была решена задача нахождения всех вхождений подстроки в строку с помощью алгоритма Кнута – Морриса – Пратта. Алгоритм разделяется на два этапа. Это построение массива $\pi[i]$ с помощью префикс-функции и нахождения вхождений подстроки в строку. Префикс – функция работает следующим образом:

- Считать значения префикс-функции $\pi[i]$ будем по очереди: от $i = 1$ к $i = n - 1$ (значение $\pi[0]$ просто присвоим равным нулю).
- Для подсчёта текущего значения $\pi[i]$ мы заводим переменную j, обозначающую длину текущего рассматриваемого образца. Изначально $j = \pi[i - 1]$.
- Тестируем образец длины j, для чего сравниваем символы $s[j]$ и $s[i]$. Если они совпадают — то полагаем $\pi[i] = j + 1$ и переходим к следующему индексу $i + 1$. Если же символы отличаются, то уменьшаем длину j, полагая её равной $\pi[j - 1]$, и повторяем этот шаг алгоритма с начала.
- Если мы дошли до длины $j = 0$ и так и не нашли совпадения, то останавливаем процесс перебора образцов и полагаем $\pi[i] = 0$ и переходим к следующему индексу $i + 1$.

Второй этап работает следующим образом:

Образуем строку S+|+T, где символ | — это разделитель, который не должен нигде более встречаться. Посчитаем для этой строки префикс-функцию. Теперь рассмотрим её значения, кроме первых $n + 1$ (которые, как видно, относятся к строке s и разделителю). По определению, значение $\pi[i]$ показывает наидлиннейшую длину подстроки, оканчивающейся в позиции i и совпадающего с префиксом. Но в нашем случае это $\pi[i]$ — фактически длина наибольшего блока совпадения со строкой s и оканчивающегося в позиции i. Больше, чем n, эта длина быть не может — за счёт разделителя. А вот равенство $\pi[i] = n$ (там, где оно достигается), означает, что в позиции i оканчивается искомое вхождение строки s

Таким образом, если в какой-то позиции i оказалось $\pi[i] = n$, то в позиции $i - (n + 1) - n + 1 = i - 2n$ строки t начинается очередное вхождение строки s в строку t.

2. Была решена задача определения, является ли A циклическим сдвигом B. Для этого понадобилась префикс функция. Префикс –функция вычисляется для значение (A+|+B), тем самым последний индекс массива $\pi[i]$ будет иметь значение длины суффикса B, который равен префиксу A.

Затем вычисляется значение префикс функции для строки(B+|+A) и так же проверяем последний элемент массива, который равен значению длины префикса B,который равен суффиксу A. Если мы сложим эти два значение, мы получим число, которое нам покажет является ли A циклическим сдвигом B. Это работает следующим образом. Если это число равно длине строки A, следовательно является и мы выводит первое вхождение, иначе не является.

Примеры работы программы

1. Поиск вхождений подстроки в строке.

```
ab
abab
0,2
```

2. Проверка является ли A цикличным сдвигом B.

```
defabc
abcdef
 3
```

Вывод

В ходе данной лабораторной работы был реализован алгоритм Кнута – Морриса – Пратта, который ищет все вхождения подстроки в строке. Для его реализации понадобилась специальная функция – Префикс-Функция. Ее задача заключалась в том, чтобы вычислить, на сколько можно сместить счетчик L (считчик, который показывает с каким символом из подстроки мы работаем) при расхождение символов из строки и подстроки. Сложность данного алгоритма по времени это O(n+m), а по памяти O(n). Временная сложность префикс – функции O(n). Данный алгоритм может работать в онлайн режиме, обрабатывать поступающий символ подстроки в реальном времени. Это возможно из-за того, что для определение $\pi[i]$ нужна только информация о предыдущих ячейках этого массива.