Программирование, 11-й класс

Листок 16. Конечные детерминированные автоматы, префикс- и z-функция, поиск подстроки.

Постановка задачи

Пусть даны две строки: $S[1..m]$ и $T[1..n], m < n$. Требуется найти индекс k в строке T, такой что

$$T[k+i-1] = S[i], i = 1 \dots m$$

Используемые термины

Префиксом строки $S = s_1 s_2 \dots s_n$ называется строка X вида $s_1 s_2 \dots s_k, k \leqslant n$. Если $k < n$, то префикс называется собственным. Обозначение: $X \sqsubset S$ (или X "начинает"строку S).

Например, для строки $abcdefg$ строки $a, ab, abc, abcd, abcde, abcdef, abcdefg$ являются префиксами и все, кроме последней – собственными префиксами.

Суффиксом строки $S = s_1 s_2 \dots s_n$ называется строка Y вида $s_k s_{k+1} \dots s_n, 1 \leqslant k \leqslant n$. Если $k > 1$, то суффикс называется собственным. Обозначение: $Y \sqsupset S$ (или Y "заканчивает"строку S).

Например, для строки $abcdefg$ строки $g, fg, efg, defg, cdefg, bcdefg, abcdefg$ являются суффиксами и все, кроме последней – собственными суффиксами.

Конечные автоматы

Конечным детерминированным автоматом M называется пятёрка $(S, s_0, \Sigma, \delta, F)$, где:

- S – множество внутренних состояний автомата
- $s_0 \in S$ – начальное состояние
- Σ – конечное множество входных символов (алфавит)
- $\delta : S \times \Sigma \to S$ – функция переходов
- $F \subseteq S$ – множество допустимых состояний

Слово *детерминированный* означает, что функция переходов δ определена однозначно на множестве $S \times \Sigma$, т.е. из каждого состояния выходит только одна стрелка с определённой буквой.

Определим Σ^* как множество строк конечной длины, составленных из символов алфавита Σ, включая пустую строку.

Пусть дана строка $w \in \Sigma^*, w = (a_1 a_2 \dots a_n)$. Тогда строка w является *допустимой* для автомата M, если существует такая последовательность состояний $q_0 q_1 \dots q_{n-1}, q_n \in S$:

- $q_0 = s_0$
- $q_{i+1} = \delta(q_i, a_{i+1}), i = 0 \dots n-1$
- $q_n \in F$

Все остальные строки будем называть *недопустимыми*.

Построение функции переходов конечного автомата для поиска образца P длины m

Определим суффикс-функцию $\sigma: \Sigma^* \to \{0, 1, \dots, m\}$ как длину наибольшего префикса P_k строки P, одновременно являющегося суффиксом строки x:

$$\sigma(x) = max\{k : P_k \sqsupset x\}$$

Тогда для множества состояний автомата $S = \{0, 1, \dots, m\}$, начального состояния $s_0 = \{0\}$, конечного состояния $s_m = \{m\}$, алфавита Σ функция переходов вычисляется по формуле:

$$\delta(q, a) = \sigma(P_q a),$$

где строка $P_q a$ получается приписыванием символа $a \in \Sigma$ справа к префиксу P_q.

Префикс-функцией строки S называется массив $\pi[0 \dots n-1]$, где $\pi[i]$ — длина максимального собственного суффикса строки $S[0 \dots i]$, являющегося её префиксом. $\pi[0]$ по определению считается равной нулю.

Z-функцией строки S называется массив $z[0 \dots n-1]$, где $z[i]$ — длина максимального префикса подстроки, начинающейся с позиции x в строке S, который одновременно является и префиксом всей строки S.

$z[0]$ считается по определению равной нулю.

Наивный метод вычисления $\pi[i]$ (цикл по длине строки; цикл по длине префикса-суффикса; сравнение подстрок) даёт оценку времени работы $O(n^3)$.

Свойства префикс-функции и z-функции позволяют построить алгоритм их вычисления с линейным временем работы.

Задачи

1. Опишите конечный детерминированный автомат и постройте функцию переходов, для которых строки из указанных множеств S (и только они) являются допустимыми, т.е. автомат заканчивает чтение строки в состоянии $q \in F$:
   (a) $\Sigma = \{a, b, c\}, S = \{(abc)^n\}$, где $n > 1$.
   (b) $\Sigma = \{a, b, c\}, S = \{a^n b^m c^k\}$, где $n, m, k \geqslant 1$.
   (c) $\Sigma = \{0, 1\}, S = \{$строки, заканчивающиеся на нечётное количество нулей.$\}$
   (d) $\Sigma = \{0, 1\}, S = \{$строки, заканчивающиеся ровно на два нуля$\}$.
   (e) $\Sigma = \{0, 1\}, v, w \in \Sigma^*, S = \{v011w\}$
   (f) $\Sigma = \{0, 1\}, S = \{$строки, в которых не менее 2-х нулей$\}$.
   (g) $\Sigma = \{0, 1\}, S = \{$строки, в которых не более 2-х нулей$\}$.
   (h) $\Sigma = \{0, 1\}, S = \{$строки, содержащие чётное количество нулей и единиц$\}$
   (i) $\Sigma = \{0, 1\}, S = \{$строки, содержащие двоичное представление чисел, кратных 6$\}$
2. Опишите конечный детерминированный автомат и изобразите функцию переходов для строк, содержащих:
   (a) `aabab` (проиллюстрируйте работу этого конечного автомата при поиске подстроки в строке `aaababaabaababaab`)
   (b) `ababaca`
   (c) `ababbabbababbababbabb`
3. Напишите программу для построения таблицы переходов конечного детерминированного автомата, принимающего данную строку.
4. Выпишите префикс-функцию для строк:
   (a) `asdf`
   (b) `aaaaab`
   (c) `bbbcbbb`
   (d) `bbccaebbcabd`
   (e) `abcaeabcabd`
   (f) `абракадабра`
5. Напишите программу для наивного вычисления префикс-функции.
6. (1323) Вычисление префикс-функции (программа должна пройти как минимум первые 84 теста).
7. Выпишите z-функцию для строк
   (a) `abacaba`
   (b) `aabcaabxaaz`
   (c) `xabxyabxyabxz`
8. (1324) Вычисление Z-функции.
9. (99) Напишите программу для поиска подстроки S ($|S| = n$) в строке T ($|T| = m$). Время работы $O(n+m)$, требуемая память $O(n)$.

   Указание: рассмотрите строку $S +' \#' + T$, где '#' – символ, заведомо не встречающийся в строках S и T и префикс-функцию на такой строке.
10. (101) Задача про циклическую подстроку.
11. (1619) Задача про подсчёт смайликов.