Жадная гипотеза в задаче о надстроке. Задача 2

15 марта 2023 г.

$\forall$ s, t: s, t - строки, введем функции:

1. $len(s)$ - возвращает длину строки s
2. $overlap(s, t)$ - возвращает пересечение s и t, то есть наибольший суффикс s, который является префиксом t

Назовем множество строк T - "множеством шифрования"для S, если:
1) строки в T состоят только из символов "0"и "1".
2) Все строки в T имеют одинаковую длину.
3) В множестве T столько элементов, сколько различных символов в строках множества S.
4) $len(overlap(s, t)) = 0 \ \forall s, t \in T : s \neq t$.
5) Пересечение конкатенации любых строк из T: a и конкатенации любых строк из T: b (причем последняя добавленная строка в a не равна первой добавленной строке в b) имеет длину 0.

Введем функции, $\forall s : s \in S$, для множества строк S: T - "множество шифрования": Тогда пусть f: множество символов строк из S $\mapsto$ T - биекция, которая ставит в соответсвие какой-то символ и какую-то строку из множества T.

1. encrypt(s) - возвращает строку $t = f(s_1) + f(s_2) + ... + f(s_{len(s)})$
2. decrypt(s) - наоборот, дешифрует строку s.

Шифрование и дешифрования происходят однозначно, так как f - биекция.

Теорема 1: Пусть $S = s_1, s_2, ..., s_n$ - множество строк, T - "множество шифрования"для S.

Множество $S_{encrypt} = encrypt(S_1), encrypt(S_2), ..., encrypt(S_n)$.

Тогда после шага алгоритма 1 для слияния будут выбраны такие строки, что после выполнения слияния при шифровке строк из S получится множество $S_{encrypt}$.

Доказательство:

1. Если len(overlap(s, t)) = x, где $s, t \in S$, то len(overlap(encrypt(s), encrypt(t))) = x * m. Действительно, так как len(overlap(a, b)) = 0 (для любых строк $a, b \in T$), то пересечением может быть только зашифрованное пересечение s и t (по пункту 4 и 5 для T), длина которого равна $x \cdot \ m$.

2. Пусть для слияния на S алгоритм 1 выбрал строки a, b: a = A + B, b = B + C, где A, B, C - строки, B = overlap(a, b), то для $S_{encrypt}$ он выберет строки encrypt(a), encrypt(b) (по пункту 1). Получим новые строки: $s* = A + B + C$ и $s*_{encrypt} = encrypt(A) + encrypt(B) + encrypt(C)$ (по пункту 4 и 5 для T). Тогда после слияния для множеств S и $S_{encrypt}$ условие Теоремы 1 выполнено.

Что и требовалось доказать

Теорема 2: Пусть $S = s_1, s_2, ..., s_n$ - множество строк, T - "множество шифрования"для S.

Множество $S_{encrypt} = encrypt(S_1), encrypt(S_2), ..., encrypt(S_n)$.

Тогда после выполнения алгоритма 1 для S и $S_{encrypt}$ мы получим строки s и $s_{encrypt}$ соответсвенно, причем $s_{encrypt}$ = encrypt(s).

Доказательство:

1. По теореме 1, после каждого шага алгоритма строки будут равны (при шифровке), значит в конеченом итоге мы получим строки, равные при шифровке.

Что и требовалось доказать

Теорема 3: Пусть $S = s_1, s_2, ..., s_n$ - множество строк, T - "множество шифрования"для S.

Множество $S_{encrypt} = encrypt(s_1), encrypt(s_2), ..., encrypt(s_n)$.

s_{opt} - оптимальная надстрока для S, а s_{enopt} - оптимальная надстрока для $S_{encrypt}$. Тогда $len(s_{opt}) \cdot m \geq len(s_{enopt})$. m - длина строк в T.

Доказательство:

1. Рассмотрим $s_1 = encrypt(s_{opt})$, она содержит в качестве подстрок все строки из $S_{encrypt}$, ведь s_1 содержала все строки из S, но $s_i = encrypt(S_i) \forall i \in \mathbf{N} : i <= n$. Но $len(s_1) * m = len(s_{enopt})$, значит $len(s_{opt}) \cdot m \geq len(s_{enopt})$ (оптимальная строка точно не длинее s_1).

Что и требовалось доказать

Тогда по теореме 2 и 3 и условию задачи, если для множества S найдется "множество шифрования".

То тогда $S_{encrypt} = encrypt(S_1), encrypt(S_2), ..., encrypt(S_n). len(S_{enres}) \geq \alpha \cdot len(S_{enopt})$ $(len(s_{res})) \cdot m = len(s_{enres}), len(s_{opt})) * m \geq len(s_{enopt}), len(s_{res}) = \alpha * len(s_{opt})))$. Где S_{enres} - строка, которую получит алгоритм 1. S_{enopt} - оптимальная надстрока для $S_{encrypt}$. Т.е. утверждение задачи будет доказано.

Докажем, что для любого набора строк S найдется "множество шифрования". $S = s_1, s_2, ..., s_n$

Пусть множество T задается следующим правилом (всего различных символов в строках S - n): 1) всего n элементов. $T = t_1, t_2, ..., t_n$ 2) $\forall i \in \mathbf{N} : i <= n : t_i = 0 + B + 11$, A = 000...0 (len(A) = n + 1), B = A, но заменим i-ый элемент в B на 1.

Например, для n = 3: T = 0100011, 0010011, 0001011.

Докажем, что T - "множество шифрования"для S.

1) Верно по правилу создания T

2) Все строки имеют длину 4 + n

3) Верно по правилу создания T

4) $len(overlap(s, t)) = 0$ $\forall s, t \in T : s \neq t$. Рассмотрим строки s, t из T ($s \neq t$). Так как в T все "1"(не считая последних) расположены на уникальном положении относительно нулей и последних двух символов (нельзя начать overlap с "1". Так как все строки начинаются с "0". Но так как начиная с любого нуля мы в overlap'е обязаны получить "11"в конце (они есть во всех строках), то единственный вариант ненулевого overlap - если строки совпадают), то overlap неравных строк из T всегда равен пустой строке.

5) Пересечение не может начинаться с "1". Так как все строки начинаются с "0". В пересечении мы обязательно получим "11"в конце пересечения, но так как "11"встречается только в конце строк, то первый "0"в пересечении должен стоять в позиции начала какой-то из добавленных строк, но так как последняя добавленная в первую итоговую строку не равна первой добавленной строке во вторую итоговую строку, то подобное пересечение может быть только нулевым.

Получили, что T - "множество шифрования"для S. Значит все условия выполнены и утверждение задачи доказан для любого множества строк S, удовлетворяющих условию задачи.

Что и требовалось доказать