О возможности восстановления периодического слова по подсловам фиксированной длины

В. А. Алексеев[♢], Ю. Г. Сметанин[♡]

[♢]Московский физико-технический институт (МФТИ)

[♡]Федеральный исследовательский центр «Информатика и управление» Российской академии наук

63-я научная конференция МФТИ
28 ноября 2020

Содержание

Комбинаторная задача о восстановлении слова по подсловам

В комбинаторике слов

- объект исследования — слова над произвольным алфавитом
- предмет исследований — изучение комбинаторных свойств различных множеств слов

Задача реконструкции слова по его подсловам — задача с неполной информацией, когда по набору последовательных фрагментов слова требуется восстановить это слово.

Пример

У слов 0110 и 1001 одинаковы множества подслов длины 2: $\{00, 01, 10, 11\}$ — и мультимножества подслов длины 2: $\{00, 01, 01, 10, 10, 11\}$.

Постановка задачи

- E_2^n – множество слов длины n из алфавита $\{0, 1\}$
- $|\alpha|$ – сумма элементов слова $\alpha = (a_1 a_2 \dots a_n) \in E_2^n$:
  $|\alpha| = a_1 + a_2 + \dots + a_n$
- λ – пустое слово
- Операция фрагментирования $\langle \alpha \cdot v \rangle$ для слова $\alpha \in E_2^n$ и опорного вектора $v \in E_2^n$:

$$\langle \alpha \cdot v \rangle = \begin{cases} a_i, & v_i = 1 \\ \lambda, & v_i = 0 \end{cases}$$

Задача

Является ли набор слов $X = \{x^1, x^2, \dots, x^N\}, x^i \in E_2^k, i = 1 \dots N$ набором фрагментов некоторого слова $\alpha \in E_2^n$, построенных с помощью операции фрагментирования векторами из $V = \{v^1, v^2, \dots, v^N\}, v^i \in E_2^n, |v^i| = k, i = 1 \dots N$, и найти все возможные решения α.

Литература

В случае полного мультимножества фрагментов ($V = E_2^n$) для однозначного восстановления слова по мультимножеству подслов фиксированной длины k достаточно

- Manvel и др. 1991
$$k \geqslant \left\lfloor \frac{n}{2} \right\rfloor$$
- Scott 1997
$$k \geqslant \bigl(1 + o(1)\bigr)\sqrt{p \log p}$$
- Krasikov и Roditty 1997
$$k \geqslant \left\lfloor \frac{16}{7}\sqrt{n} \right\rfloor + 5$$

Цель работы

Определение (периодическое слово)

Пусть $x^s \equiv \underbrace{xx\ldots x}_{s}$. Слово $\alpha = (a_1 a_2 \ldots a_n), a_i \in \{0, 1\}$ называется периодическим с периодом p, если

$$\alpha = (a_1 a_2 \ldots a_p)^l$$

Цель

Улучшить оценки, полученные в случая произвольного слова, для случая периодического слова.

Теорема

Для однозначного восстановления периодического слова длины n с периодом p по мультимножеству всех его подслов длины k достаточно выполнения условия

$$k \geqslant \left\lfloor \frac{16}{7}\sqrt{p} \right\rfloor + 5$$

Идея — сведение к системе диофантовых уравнений

Слова $\alpha = a_1 \ldots a_n$ и $\beta = b_1 \ldots b_n$

$$\sum_{r=1}^{n} a_r r^j = \sum_{r=1}^{n} b_r r^j \quad 0 \leqslant j \leqslant k-1 \tag{1}$$

Слова α и β имеют одинаковые мультимножества подслов длины $k \Leftrightarrow$ система (1) имеет нетривиальное решение (Krasikov и Roditty 1997). Лишь тривиальное решение будет в случае

$$k \geqslant \left\lfloor \frac{16}{7}\sqrt{n} \right\rfloor + 5$$

Для слов $\alpha = (a_1 a_2 \ldots a_p)^l$ и $\beta = (b_1 b_2 \ldots b_p)^l$:

$$\sum_{r=1}^{p} a_r r^j = \sum_{r=1}^{p} b_r r^j \quad 0 \leqslant j \leqslant k-1$$

$$k \geqslant \left\lfloor \frac{16}{7}\sqrt{p} \right\rfloor + 5$$

Доказательство 1

Лемма

Для любого слова набор его фрагментов вида x^j1 однозначно определяет набор его моментов вида $\sum_{r=1}^{n} a_r r^j$.

Доказательство леммы

$N_\beta(\alpha)$ – число фрагментов, равных β, в слове α. Например, $N_1(\alpha) = \sum_{r=1}^{n} a_r$, $N_{x1}(\alpha) = \sum_{r=1}^{n}(r-1)a_r = \sum_{r=1}^{n} ra_r - \sum_{r=1}^{n} a_r$.

$$N_{x^{k-1}1}(\alpha) = \sum_{r=1}^{n} \binom{r-1}{k-1} a_r$$

$$= \frac{1}{(k-1)!}\sum_{r=1}^{n} r^{k-1}a_r - f_{k-1}\Big(N_1(\alpha), \ldots, N_{x^{k-2}1}(\alpha)\Big)$$

Доказательство 1

$$\sum_{r=1}^{n} a_r r^j \equiv s_j(\alpha) \quad 0 \leqslant j \leqslant k-1$$

Для $\alpha = (a_1 a_2 \dots a_p)^l$ при $m = 0$: $\sum\limits_{r=1}^{n} a_r = s_0(\alpha) \Leftrightarrow \sum\limits_{r=1}^{p} a_r = \frac{s_0(\alpha)}{l}$.
Для произвольного $m \in [1, k-1]$:

$$\begin{aligned}\sum_{r=1}^{n} a_r r^m &= \sum_{r=1}^{p} a_r r^m + \sum_{r=p+1}^{2p} a_r r^m + \ldots + \sum_{r=(l-1)p+1}^{lp} a_r r^m \\ &= l \cdot \sum_{r=1}^{p} a_r r^m + \sum_{j=1}^{m} \sum_{r=1}^{p} \binom{m}{j} ((l-1)p)^j r^{m-j} a_r \\ &= f_m\left(\sum_{r=1}^{p} a_r r^m, \ldots, \sum_{r=1}^{p} a_r\right)\end{aligned}$$

Теорема

Пусть слово периодическое, начиная с некоторого индекса: $\alpha = a_1 a_2 \ldots a_q (a_{q+1} a_{q+2} \ldots a_{q+p})^l$. *Тогда при*

$$l > q^{P \log P} \quad P \equiv \max(p, q)$$

для однозначного восстановления слова достаточно

$$k \geqslant \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5$$

Доказательство 2. Идея — разделение уравнений

При $m = 0$:

$$s_0(\alpha) = \sum_{r=1}^{n} a_r = a_a + \ldots a_q + l \cdot \sum_{r=q+1}^{q+p} a_r$$

При $q < l$ верно $\frac{a_1 + \ldots a_q}{l} < 1$, поэтому

$$\begin{cases} \sum\limits_{r=q+1}^{q+p} a_r = \left\lfloor \frac{s_0(\alpha)}{l} \right\rfloor \\ \sum\limits_{r=1}^{q} a_r = \left\{ \frac{s_0(\alpha)}{l} \right\} \cdot l \end{cases}$$

Доказательство 2

Для произвольного $m \in [1, k-1]$:

$$\sum_{r=1}^{q} a_r r^m + \sum_{r=q+1}^{n} a_r r^m = s_m(\alpha)$$

$$= \sum_{r=1}^{q} a_r r^m + \sum_{r=q+1}^{q+p} a_r \left((q+r)^m + \ldots + (q+p(l-1)+r)^m\right)$$

$$= \sum_{r=1}^{q} a_r r^m + l \cdot \sum_{r=q+1}^{q+p} a_r r^m + f_m(s_{m-1}, \ldots, s_0)$$

Так как $\sum_{k=1}^{n} k^p \to \frac{n^{p+1}}{p+1}$, то при $\frac{q^k}{k} < l$ разделяются все уравнения.

Поэтому при $l \geqslant q^{\left\lfloor \frac{16}{7}\sqrt{P} \right\rfloor + 5}$, где $P \equiv \max(p, q)$

$$k \geqslant \left\lfloor \frac{16}{7}\sqrt{P} \right\rfloor + 5$$

Заключение

Получены оценки на длину подслова k, достаточной для однозначного восстановления периодического слова по мультимножеству подслов фиксированной длины.

- Для слова $\alpha = (a_1 a_2 \dots a_p)^l$:

$$k \geqslant (1 + o(1))\sqrt{p \log p}$$

$$k \geqslant \left\lfloor \frac{16}{7}\sqrt{p} \right\rfloor + 5$$

- Для слова $\alpha = a_1 a_2 \dots a_q (a_{q+1} a_{q+2} \dots a_{q+p})^l$, при условии $l > q^{P \log P}$, $P \equiv \max(p, q)$:

$$k \geqslant (1 + o(1))\sqrt{P \log P}$$

$$k \geqslant \left\lfloor \frac{16}{7}\sqrt{P} \right\rfloor + 5$$