

О ДЛИНЕ ФРАГМЕНТОВ ДЛЯ ОДНОЗНАЧНОГО ВОССТАНОВЛЕНИЯ ПЕРИОДИЧЕСКОГО СЛОВА ПО ПОЛНОМУ МУЛЬТИМНОЖЕСТВУ ФРАГМЕНТОВ ФИКСИРОВАННОЙ ДЛИНЫ

© 2022 г. В. А. Алексеев^{1,*}, Ю. Г. Сметанин²

Представлено академиком РАН К. В. Рудаковым 18.11.2021

Поступило 23.03.2021 г.

После доработки 27.01.2022 г.

Принято к публикации 17.02.2022 г.

Рассмотрена задача о восстановлении слов из конечного алфавита по мультимножеству их фрагментов одной длины. При этом ставится следующее ограничение на восстанавливаемые слова: они должны быть либо периодическими, либо должны содержать периодическое слово как подслово. Показано, что периодическое слово с периодом p восстановимо мультимножеству фрагментов длины k при $k \geq \left\lfloor \frac{16}{7} \sqrt{p} \right\rfloor + 5$. Для слова, состоящего из периодического префикса с периодом q , повторяющегося m раз, и периодического суффикса с периодом p , повторяющегося l раз, получена оценка $k \geq \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5$, где $P = \max(p, q)$, при условии $l \geq mq^{\left\lfloor \frac{16}{7} \sqrt{p} \right\rfloor + 5}$.

Ключевые слова: слово, фрагмент, подслово, периодическое слово, восстановление по неполной информации, восстановление слова, реконструкция слова, мультимножество подслов

DOI: 10.31857/S2686954322020035

В общем случае задачу о восстановлении объекта некоторой природы по неполной информации о его “частях” можно рассматривать как задачу распознавания образов [1]. В частности, когда объектами выступают слова как последовательности символов над определенным алфавитом, возникает вопрос о возможности восстановления слова по множеству или мультимножеству его подслов. Эта задача лежит в области дискретной математики, называемой комбинаторикой слов [2–6], где исследуется связь между последовательностями символов и множествами их подпоследовательностей.

Можно отметить несколько приложений, которые имеет задача о восстановлении слов по подсловам. Так, в символьном кодировании произвольная информация представляется в виде последовательности символов и слов [7]. Такой спо-

соб представления информации используется, например, при передаче данных по каналам связи [8], когда важно обеспечить надежную передачу такой закодированной информации, исключающую потери, или позволяющую восстановить исходное слово по доступным фрагментам, если потери случаются [9]. Символьное кодирование также используется, например, в области анализа временных рядов [10], в биоинформатике [11].

В работах [12, 13] получены оценки для длины фрагмента, достаточной для того, чтобы произвольное слово можно было восстановить по мультимножеству всех его подслов.

В данной же работе на восстанавливаемые слова накладывается ограничение, состоящее в том, что слова должны содержать периодическое подслово. Это позволяет получить более низкую оценку на достаточную для восстановления слова длину фрагментов. Так, в [17] представлен результат о восстановлении слова с периодическим суффиксом и непериодическим префиксом. Настоящая же работа теоремой 3 о восстановлении слова с периодическим суффиксом и периодическим префиксом усиливает указанный результат, полученный в [17]. Теоремы 1 и 2, связанные с восстановимостью периодических слов, повторяют аналогичные теоремы из [17]. В настоящей работе

¹ Московский физико-технический институт (национальный исследовательский университет), Долгопрудный, Московская обл., Россия

² Федеральный исследовательский центр “Информатика и управление” Российской академии наук, Москва, Россия

*E-mail: vasilij.alekseyev@phystech.edu

рассматривается еще несколько случаев слов, не являющихся периодическими, но содержащих периодическое слово или подслово. А именно, сформулированы и доказаны теоремы о восстановимости для слова с непериодическим подсловом между периодическими префиксом и суффиксом (теорема 4) и для периодического слова с ограничением на повторяющееся подслово (теорема 5). Каждая из этих теорем, в свою очередь, включает результат теоремы 3 как частный случай.

1. ПОСТАНОВКА ЗАДАЧИ

Введем следующие обозначения и определения. Греческими буквами будем обозначать слова, малыми латинскими — символы алфавита.

Пусть E_2^n — множество слов длины n из алфавита $\{0, 1\}$. Для слова $\alpha = (a_1 a_2 \dots a_n) \in E_2^n$ символом $|\alpha|$ обозначается сумма его элементов: $|\alpha| = a_1 + a_2 + \dots + a_n$. Пусть λ означает пустое слово, т.е. слово, длина которого равна нулю. Возведение слова α в степень s является краткой записью для слова, состоящего из s повторов слова α , т.е. $\alpha^s = \beta_1 \beta_2 \dots \beta_s$, где $\beta_i = \alpha, i \in \{1, 2, \dots, s\}$.

Для заданного слова $\alpha = (a_1 a_2 \dots a_n) \in E_2^n$ и заданного опорного вектора $v = (v_1 v_2 \dots v_n)$, где $v_i \in \{0, 1\}$, $i \in \{1, 2, \dots, n\}$ операция фрагментирования $\langle \alpha \cdot v \rangle$ строит слово длины $|v|$ по следующему правилу:

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

Фрагментом, или подсловом, слова $\alpha = (a_1 a_2 \dots a_n) \in E_2^n$ будем называть слово $\tilde{\alpha} \in E_2^n$ следующего вида:

$$\tilde{\alpha} = (a_{i_1} a_{i_2} \dots a_{i_k}), \quad 1 \leq i_1 < i_2 < \dots < i_k \leq n.$$

Пусть $f^*(\alpha)$ — это наименьшая длина фрагментов k , при которой гарантирована однозначность восстановления слова α длины n по мультимножеству его подслов длины k .

З а д а ч а. В общем виде задача о восстановлении слова по мультимножеству подслов формулируется следующим образом. Заданы:

набор опорных векторов $V = \{v^1, v^2, \dots, v^N\}$, $v^i \in E_2^n, |v^i| = k, i \in \{1, 2, \dots, N\}$,

и множество слов $X = \{\chi^1, \chi^2, \dots, \chi^N\}$, $\chi^i \in E_2^k$, $i \in \{1, 2, \dots, N\}$.

Требуется проверить, является ли набор векторов X набором фрагментов фиксированной длины некоторого неизвестного слова $\alpha \in E_2^n$, построенных с помощью операции фрагментирования векторами из V , и найти все возможные решения.

При этом все оценки, получаемые для случая двоичного алфавита, остаются верными и для произвольного алфавита [3], так как в случае алфавита $\{0, 1, \dots, k\}$ задача сводится к набору из k задач в двоичном алфавите с помощью отображений:

$$\begin{cases} \phi_i : x \mapsto \begin{cases} 1, & x = i, \\ 0, & x \neq i, \end{cases} \\ i \in \{0, 1, \dots, k\}. \end{cases}$$

Поэтому далее будут рассмотрены только двоичные слова $\alpha = (a_1 a_2 \dots a_n), a_i \in \{0, 1\}$. Более того, нас будут интересовать периодические слова.

Слово $\alpha = (a_1 a_2 \dots a_n), a_i \in \{0, 1\}$ называется периодическим с периодом p , если

$$\alpha = (a_1 a_2 \dots a_p)^l, \quad a_i \in \{0, 1\}. \quad (1)$$

При этом подслово $\tilde{\alpha} = a_1 a_2 \dots a_p$ будем называть порождающим подсловом для слова α вида (1).

2. СОСТОЯНИЕ ЗАДАЧИ

Для задачи о восстановлении слова по подсловам известно ее сведение к проверке единственности решения диофантовых уравнений определенного вида [2]. Показано, что задача о существовании и единственности решения является NP-полной.

В случае полного мультимножества фрагментов ($V = E_2^n$) установлено [14], что для однозначного восстановления слова α длины n необходима длина $k \geq \exp(\Omega(\log n))$.

Для того же случая полного мультимножества фрагментов известны следующие оценки на достаточную длину [3, 13, 12]: $k \geq \left\lfloor \frac{n}{2} \right\rfloor$,

$$k \geq (1 + o(1))\sqrt{n \log n}, \quad k \geq \left\lfloor \frac{16}{7} \sqrt{n} \right\rfloor + 5.$$

Было изучено несколько частных случаев. Так, показано [15], что для слов, состоящих из l серий, достаточно фрагментов длины l .

3. РЕЗУЛЬТАТЫ

Сформулируем первую теорему.

Т е о р е м а 1. Для однозначного восстановления периодического слова длины n с периодом p по мультимножеству всех его фрагментов длины k достаточно выполнения условия

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

Перед доказательством введем следующее

Определение 1. Пусть $N_\beta(\alpha)$ – это число фрагментов, равных β , в слове α .

В [2] показано, что по $N_\beta(\alpha)$ для всех двоичных слов β длины k однозначно восстанавливаются числа фрагментов всех длин меньше k . Далее сформулируем и докажем лемму, которая непосредственно следует из системы уравнений в [2].

Лемма 1. Для любого слова α по набору его фрагментов вида $x^j 1$ однозначно определяется набор его моментов вида $\sum_{r=1}^n a_r r^j$.

Доказательство. Получим формулы для $N_{x^j 1}(\alpha)$ при разных j :

$$\begin{aligned} N_1(\alpha) &= \sum_{r=1}^n a_r, \\ N_{x1}(\alpha) &= \sum_{r=1}^n (r-1) a_r = \\ &= \sum_{r=1}^n a_r r - N_1(\alpha) = \sum_{r=1}^n a_r r - f_1(N_1(\alpha)), \\ &\dots \\ N_{x^{k-1}1}(\alpha) &= \sum_{r=1}^n \binom{r-1}{k-1} a_r = \frac{1}{(k-1)!} \sum_{r=1}^n a_r r^{k-1} - \\ &- f_{k-1}(N_1(\alpha), N_{x1}(\alpha), \dots, N_{x^{k-2}1}(\alpha)). \end{aligned}$$

Функции $f_2, f_3, \dots, f_{k-1}$ вычисляются на основе комбинаторных соотношений. Например, найдем выражение функции $f_p(N_1(\alpha), \dots, N_{x^{p-1}1}(\alpha))$, $2 \leq p \leq k-1$, через полученные на предыдущих шагах $f_1, \dots, f_{p-1}$ (при этом можно положить $f_0 \equiv 0$):

$$\begin{aligned} N_{x^p 1}(\alpha) &= \sum_{r=1}^n \binom{r-1}{p} a_r = \sum_{r=1}^n \frac{(r-1)!}{p!(r-1-p)!} a_r = \\ &= \sum_{r=1}^n \frac{(r-1) \cdot (r-2) \dots (r-p)}{p!} a_r = \\ &= \sum_{r=1}^n \frac{1}{p!} a_r \left\{ r^p + r^{p-1} \sum_{i_1=1}^p (-i_1) + \right. \\ &\left. + \sum_{\substack{i_1, i_2 \in \{1, 2, \dots, p\} \\ i_1 \neq i_2}} (-i_1)(-i_2) + \dots + r^0 \prod_{i_p=1}^p (-i_p) \right\} = \end{aligned}$$

$$\begin{aligned} &= \frac{1}{p!} \sum_{r=1}^n a_r r^p + \frac{1}{p!} \sum_{l=1}^p \sum_{r=1}^n a_r r^{p-1} \times \\ &\times \sum_{\substack{S \subseteq \{1, \dots, p\} \\ |S|=l}} \prod_{i \in S} (-i) = \frac{1}{p!} \sum_{r=1}^n a_r r^p + \\ &+ \frac{1}{p!} \sum_{l=1}^p (N_{x^{p-l} 1}(\alpha) + f_{p-1}(N_1(\alpha), \dots, N_{x^{p-l-1} 1}(\alpha))) \times \\ &\times \sum_{\substack{S \subseteq \{1, \dots, p\} \\ |S|=l}} \prod_{i \in S} (-i) = \frac{1}{p!} \sum_{r=1}^n a_r r^p + f_p(N_1(\alpha), \dots, N_{x^{p-1} 1}(\alpha)). \end{aligned}$$

Так как, по предположению, все $f_1, \dots, f_{p-1}$ известны, то известна и $f_p(N_1(\alpha), \dots, N_{x^{p-1} 1}(\alpha))$.

Теперь мы можем перейти к доказательству теоремы 1.

Доказательство. Итак, доказательство базируется на результатах, полученных в [13] для достаточной для однозначной реконструкции слова длины фрагментов в случае произвольных слов. Оценка величины $f^*(\alpha)$ в [13] получена на основе анализа следующей системы уравнений:

$$\begin{cases} \sum_{r=1}^n a_r r^j = s_j(\alpha), \\ 0 \leq j \leq k-1, \end{cases} \quad (2)$$

для которой при $f^*(\alpha) \leq (1 + o(1))\sqrt{n \log n}$ решение единственно.

Пусть слово α состоит из $l = \frac{n}{p}$ периодов:

$\alpha = (a_1 a_2 \dots a_p)^l$. Тогда нулевое уравнение $\sum_{r=1}^n a_r = s_0(\alpha)$ можно переписать в виде $\sum_{r=1}^p a_r = \frac{s_0(\alpha)}{l}$. И для произвольного индекса k' от 1 до $k-1$ включительно:

$$\begin{aligned} \sum_{r=1}^n a_r r^{k'} &= \sum_{r=1}^p a_r r^{k'} + \sum_{r=p+1}^{2p} a_r r^{k'} + \dots + \sum_{r=(l-1)p+1}^{lp} a_r r^{k'} = \\ &= \sum_{r=1}^p a_r r^{k'} + \sum_{r=1}^p a_r (p+1)^{k'} + \dots + \sum_{r=1}^p a_r ((l-1)p+r)^{k'} = \\ &= \sum_{r=1}^p a_r r^{k'} + \sum_{r=1}^p \sum_{j=0}^{k'} \binom{k'}{j} p^j r^{k'-j} a_r + \dots + \\ &+ \sum_{r=1}^p \sum_{j=0}^{k'} \binom{k'}{j} ((l-1)p)^j r^{k'-j} a_r = l \cdot \sum_{r=1}^p a_r r^{k'} + \\ &+ \sum_{j=1}^{k'} \sum_{r=1}^p \binom{k'}{j} ((l-1)p)^j r^{k'-j} a_r = \\ &= f_{k'} \left(\sum_{r=1}^p a_r r^{k'}, \dots, \sum_{r=1}^p a_r \right), \end{aligned}$$

где f_k — линейная функция, и далее доказательство теоремы в точности повторяет доказательство из [13].

Сформулируем более сильное утверждение.

Теорема 2. Для однозначного восстановления периодического слова длины n с периодом p по мультимножеству всех его фрагментов длины k достаточно выполнения условия

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

Доказательство. Доказательство основано на [12], где авторы доказывают возможность однозначной реконструкции слова по его подсловам не из анализа системы уравнений (2), как было проделано в [13], но из анализа похожей системы следующего вида:

$$\begin{cases} \sum_{r=1}^n a_r r^j = \sum_{r=1}^n b_r r^j, \\ 0 \leq j \leq k-1, \end{cases} \quad (3)$$

выписанной для двух слов $\alpha = a_1 \dots a_n$ и $\beta = b_1 \dots b_n$. Доказывается, что слова α и β имеют одинаковые мультимножества подслов длины k тогда и только тогда, когда система (3) имеет нетривиальное решение $(a_1, \dots, a_n, b_1, \dots, b_n)$. Таким образом, доказательство сводится к поиску условия, при котором система диофантовых уравнений (3) имеет только тривиальное решение. Авторы [12] затем ссылаются на [16], где получен следующий результат:

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

Для случая же периодических слов ранее в данной работе при доказательстве теоремы 1 показано, как систему (3) можно переписать в переменных только $a_1, \dots, a_p, b_1, \dots, b_p$, где p — период слова. Таким образом, оценка на достаточную для однозначного восстановления исходного слова длину подслов получается равной

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

Реконструкция слова остается возможной и тогда, когда само слово непериодическое, но состоит из нескольких подслов, по крайней мере одно из которых периодическое. Далее мы рассмотрим несколько случаев таких слов. Каждый случай будет оформлен в виде отдельной теоремы, доказательство которой будет опираться на уже доказанное ранее для периодического слова (теорема 2).

Теорема 3. Пусть слово α состоит из двух подслов: префикса и суффикса. Причем оба подслова периодические:

$$\alpha = (a_1 a_2 \dots a_q)^m (a_{q+1} a_{q+2} \dots a_{q+p})^l. \quad (4)$$

Тогда при $l \geq m q \left\lfloor \frac{16}{7} \sqrt{p} \right\rfloor + 5$, где $P \equiv \max(p, q)$, для однозначного восстановления слова (4) достаточно длины фрагментов $k \geq \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5$.

Доказательство. Найдем $s_0(\alpha)$:

$$\sum_{r=1}^n a_r = m \sum_{r=1}^q a_r + l \sum_{r=q+1}^{q+p} a_r = s_0(\alpha).$$

При $m q < l$ получаем $\frac{m(a_1 + \dots + a_q)}{l} < 1$, поэтому

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

Рассмотрим случай $s_1(\alpha)$:

$$s_1(\alpha) = \sum_{r=1}^{qm} a_r r + \sum_{r=qm+1}^{qm+p l} a_r r. \quad (5)$$

Распишем первую сумму в выражении для $s_1(\alpha)$ (5):

$$\begin{aligned} \sum_{r=1}^{qm} a_r r &= a_1 \cdot \sum_{j=0}^{m-1} (qj+1) + \dots + a_q \cdot \sum_{j=0}^{m-1} (qj+q) = \\ &= \sum_{r=1}^q a_r \left(mr + \frac{qm(m-1)}{2} \right) = m \sum_{r=1}^q a_r r + \frac{qm(m-1)}{2} \sum_{r=1}^q a_r. \end{aligned}$$

Аналогично для второй суммы из выражения (5) получаем:

$$\begin{aligned} \sum_{r=qm+1}^{qm+p l} a_r r &= a_{q+1} \sum_{j=0}^{l-1} (qm + pj + 1) + \\ &+ \dots + a_{q+p} \sum_{j=0}^{l-1} (qm + pj + p) = \\ &= \sum_{r=q+1}^{q+p} a_r \left(lr + \frac{2qml + pl(l-1)}{2} \right) = \\ &= l \sum_{r=q+1}^{q+p} a_r r + \left(qml + \frac{pl(l-1)}{2} \right) \sum_{r=q+1}^{q+p} a_r. \end{aligned}$$

И в итоге имеем:

$$s_1(\alpha) = m \sum_{r=1}^q a_r r + l \sum_{r=q+1}^{q+p} a_r r + f_1(s_0).$$

При условии $m q (q+1)/2 < l$ уравнения для префикса и суффикса разделяются, как и в случае с $s_0(\alpha)$. Перейдем теперь к случаю произвольного k' , такого что $2 \leq k' \leq k-1$:

$$s_{k'}(\alpha) = \sum_{r=1}^{qm} a_r r^{k'} + \sum_{r=qm+1}^{qm+pl} a_r r^{k'}.$$

Распишем первую сумму:

$$\begin{aligned} \sum_{r=1}^{qm} a_r r^{k'} &= \sum_{r=1}^q a_r (r^{k'} + \dots + (q(m-1) + r)^{k'}) = \\ &= \sum_{r=1}^q a_r \sum_{h=1}^m (r + q(m-h))^{k'} = \\ &= \sum_{r=1}^q a_r \sum_{h=1}^m \sum_{j=0}^{k'} \binom{k'}{j} r^j (q(m-h))^{k'-j} = \\ &= m \sum_{r=1}^q a_r r^{k'} + g_{k'}(s_{k'-1}, \dots, s_0). \end{aligned}$$

В последнем переходе мы оставили отдельно член с показателем степени при r , равном k' (при $j = k'$). Члены, куда r входит в меньшей степени, мы представили с помощью линейной функции $g_{k'}$.

Аналогично для второй суммы имеем:

$$\begin{aligned} \sum_{r=qm+1}^{qm+pl} a_r r^{k'} &= \sum_{r=q+1}^{q+p} a_r ((q(m-1) + r)^{k'} + \dots + \\ &+ (q(m-1) + p(l-1) + r)^{k'}) = \\ &= \sum_{r=q+1}^{q+p} a_r \sum_{h=1}^l \sum_{j=0}^{k'} \binom{k'}{j} (q(m-1) + r)^j (p(l-h))^{k'-j} = \\ &= \sum_{r=q+1}^{q+p} a_r \sum_{h=1}^l \sum_{j=0}^{k'} \binom{k'}{j} \binom{j}{s} (q(m-1))^s r^{j-s} (p(l-h))^{k'-j} = \\ &= l \sum_{r=q+1}^{q+p} a_r r^{k'} + h_{k'}(s_{k'-1}, \dots, s_0). \end{aligned}$$

Где снова в последнем переходе мы оставили отдельно член с показателем степени при r , равном k' (при $j = k'$ и $s = 0$). Члены, куда r входит в меньшей степени, мы представили с помощью линейной функции $h_{k'}$.

Таким образом, объединяя результаты, полученные для обеих сумм, мы можем вернуться к вычислению $s_{k'}(\alpha)$:

$$\begin{aligned} s_{k'}(\alpha) &= \sum_{r=1}^{qm} a_r r^{k'} + \sum_{r=qm+1}^{qm+pl} a_r r^{k'} = \\ &= \left(m \sum_{r=1}^q a_r r^{k'} + g_{k'}(s_{k'-1}, \dots, s_0) \right) + \\ &+ \left(l \sum_{r=q+1}^{q+p} a_r r^{k'} + h_{k'}(s_{k'-1}, \dots, s_0) \right) = \\ &= m \sum_{r=1}^q a_r r^{k'} + l \sum_{r=q+1}^{q+p} a_r r^{k'} + f_{k'}(s_{k'-1}, \dots, s_0), \end{aligned}$$

где за $f_{k'}$ мы обозначили линейную функцию от $s_{k'-1}, \dots, s_0$, являющуюся суммой функций $g_{k'}$ и $h_{k'}$.

Так как $\sum_{k=1}^n k^p \rightarrow \frac{n^{p+1}}{p+1}$, то при $m \cdot \frac{q^k}{k} < l$ разделяются все уравнения и получается две системы. Поэтому, по доказанному ранее для случая периодического слова (теорема 2), при $l \geq m q^{\lfloor \frac{16}{7} \sqrt{P} \rfloor + 5}$, где $P \equiv \max(p, q)$, размер подслов ограничивается следующим образом: $k \geq \lfloor \frac{16}{7} \sqrt{P} \rfloor + 5$.

Следствие 1. При $m = 1$ слово α имеет вид слова с непериодическим префиксом и периодическим суффиксом:

$$\alpha = a_1 a_2 \dots a_q (a_{q+1} a_{q+2} \dots a_{q+p})^l.$$

При длине суффикса, большей длины префикса: $l \geq q^{\lfloor \frac{16}{7} \sqrt{P} \rfloor + 5}$ — для однозначного восстановления слова α достаточно подслов длины $k \geq \lfloor \frac{16}{7} \sqrt{P} \rfloor + 5$ при $P \equiv \max(p, q)$.

Теорема 4. Пусть слово α состоит из трех подслов: периодического префикса $\bar{\alpha}$, периодического суффикса $\hat{\alpha}$ и непериодического корня $\tilde{\alpha}$:

$$\alpha = \bar{\alpha} \tilde{\alpha} \hat{\alpha} = (a_1 a_2 \dots a_q)^m a_{q+1} a_{q+2} \dots \times \times a_{q+s} (a_{q+s+1} a_{q+s+2} \dots a_{q+s+p})^l. \quad (6)$$

Тогда при $m \geq s^{k^*}$ и $l \geq (m-1)q^{k^*} + (q+s)^{k^*}$, где $k^* = \lfloor \frac{16}{7} \sqrt{P} \rfloor + 5$ и $P \equiv \max(p, q, s)$, для однозначного восстановления слова (6) достаточно $k \geq k^*$.

Доказательство. Доказательство опирается на теоремы 2 и 3.

Рассмотрим $s_0(\alpha)$:

$$s_0(\alpha) = \sum_{r=1}^{qm+s+pl} a_r = m \sum_{r=1}^q a_r + \sum_{r=q+1}^{q+s} a_r + l \sum_{r=q+s+1}^{q+s+p} a_r.$$

Тогда при $m q + s < l$ разделяются уравнения для двух подслов: $\bar{\alpha} \tilde{\alpha}$ и $\hat{\alpha}$.

Рассмотрим k' , такой что $1 \leq k' \leq k-1$. Опуская подробные выкладки, аналогичные уже приведенным в доказательстве теоремы 3, получаем:

$$\begin{aligned} s_{k'}(\alpha) &= m \sum_{r=1}^q a_r r^{k'} + \sum_{r=q+1}^{q+s} a_r r^{k'} + \\ &+ l \sum_{r=q+s+1}^{q+s+p} a_r r^{k'} + f_{k'}(s_{k'-1}, \dots, s_0), \end{aligned} \quad (7)$$

где $f_{k'}$ — линейная функция. Таким образом, при выполнении условия $l \geq m q^{k_1^*} + (q+s)^{k_1^*} - q^{k_1^*}$, где $k_1^* = \lfloor \frac{16}{7} \sqrt{P_1} \rfloor + 5$ и $P_1 = \max(p, qm + s)$, мы получаем разделимость уравнений для слов $\bar{\alpha} \tilde{\alpha}$ и $\hat{\alpha}$, а по-

тому и восстановимость целого слова при ограничении на длины подслов $k \geq k_1^*$.

Однако мы также можем отдельно рассмотреть вопрос о восстановимости слова $\bar{\alpha}\bar{\alpha}$. Действительно, по следствию 1, получаем, что для реконструкции слова $\bar{\alpha}\bar{\alpha}$ по множеству его подслов достаточно, чтобы $m \geq s^{k_2^*}$, где $k_2^* = \left\lfloor \frac{16}{7}\sqrt{P_2} \right\rfloor + 5$, $P_2 = \max(q, s)$. Тогда длина подслов ограничивается как $k \geq k_2^*$.

Объединяя условия с k_1^* и k_2^* , получаем, что слово α восстановимо по подсловам при $m \geq s^{k^*}$, $l \geq (m-1)q^{k^*} + (q+s)^{k^*}$, где $k^* = \left\lfloor \frac{16}{7}\sqrt{P} \right\rfloor + 5$ и $P \equiv \max(p, q, s)$. Длина подслов при этом $k \geq k^*$.

З а м е ч а н и е. Длина корня $s = 0$ дает случай, уже рассмотренный в теореме 3.

С л е д с т в и е 2. Длина корня $s = 1$ равносильна наличию одного символа между двумя периодическими подсловами в слове α . В таком случае α восстановимо по подсловам при $l \geq (m-1)q^{k^*} + (q+1)^{k^*}$, где $k^* = \left\lfloor \frac{16}{7}\sqrt{P} \right\rfloor + 5$ и $P \equiv \max(p, q)$. Длина подслов k ограничивается соотношением: $k \geq k^*$.

С л е д с т в и е 3. Пусть в слове α вида (6) порождающие подслова у префикса и суффикса одинаковы:

$$\alpha = (a_1 a_2 \dots a_q)^m a_{q+1} a_{q+2} \dots a_{q+s} (a_{q+s+1} a_{q+s+2} \dots a_{q+s+p})^l, \\ q = p, \quad a_i = a_j, \\ (i, j) \in \{(1, q+s+1), \dots, (q, q+s+p)\}.$$

Пусть $k^* = \left\lfloor \frac{16}{7}\sqrt{P} \right\rfloor + 5$ и $P \equiv \max(q, s)$. Тогда при $l+m \geq s^{k^*}$ слово α однозначно восстановимо при $k \geq k^*$.

Д о к а з а т е л ь с т в о. Уравнение (7) в случае слова α указанной структуры примет вид:

$$s_{k'}(\alpha) = (m+l) \sum_{r=1}^q a_r r^{k'} + \sum_{r=q+1}^{q+s} a_r r^{k'} + f_{k'}(s_{k'-1}, \dots, s_0).$$

А потому при $m+l \geq (q+s)^{k^*} - q^{k^*}$ получаем разделимость уравнений для порождающего подслова префикса (суффикса) и для корня.

Т е о р е м а 5. Пусть слово α имеет вид

$$\alpha = ((a_1 a_2 \dots a_q)^m (a_{q+1} a_{q+2} \dots a_{q+p})^l)^N, \quad (8)$$

т.е. образовано повторением подслова, которое, в свою очередь, состоит из периодического префикса и суффикса. Тогда при $l \geq m q^{k^*}$, где $k^* = \left\lfloor \frac{16}{7}\sqrt{P} \right\rfloor + 5$

и $P \equiv \max(p, q)$, для однозначного восстановления слова (8) достаточно $k \geq k^*$.

Д о к а з а т е л ь с т в о. Начнем с вычисления $s_0(\alpha)$:

$$s_0(\alpha) = \sum_{r=1}^{N(qm+pl)} a_r = Nm \sum_{r=1}^q a_r + Nl \sum_{r=q+1}^{q+p} a_r.$$

При $m < l$ получаем разделимость уравнений:

$$\begin{cases} \sum_{r=q+1}^{q+p} a_r = \left\lfloor \frac{s_0(\alpha)}{Nl} \right\rfloor, \\ \sum_{r=1}^q a_r = \left\lfloor \frac{s_0(\alpha)}{Nm} \right\rfloor \cdot \frac{l}{m}. \end{cases}$$

Введя обозначение $\tilde{n} \equiv qm + pl$, выразим теперь $s_{k'}(\alpha)$ при k' , таком что $1 \leq k' \leq k-1$:

$$\begin{aligned} s_{k'}(\alpha) &= \sum_{r=1}^{N\tilde{n}} a_r r^{k'} = \sum_{j=0}^{N-1} \left[\left(\sum_{i=0}^{m-1} \sum_{r=\tilde{n}j+qi+1}^{\tilde{n}j+qi+q} a_r r^{k'} e^{k'} \right) + \right. \\ &\quad \left. + \left(\sum_{i=0}^{l-1} \sum_{r=\tilde{n}j+qm+pi+1}^{\tilde{n}j+qm+pi+p} a_r r^{k'} e^{k'} \right) \right] = \\ &= \sum_{j=0}^{N-1} \left[\left(\sum_{i=0}^{m-1} \sum_{r=1}^q a_r (r + \tilde{n}j + qi)^{k'} \right) + \right. \\ &\quad \left. + \left(\sum_{i=0}^{l-1} \sum_{r=q+1}^{q+p} a_r (r + \tilde{n}j + qm + pi - q)^{k'} \right) \right] = \\ &= Nm \sum_{r=1}^q a_r r^{k'} + Nl \sum_{r=q+1}^{q+p} a_r r^{k'} + f_{k'}(s_{k'-1}, \dots, s_0). \end{aligned}$$

Таким образом, при $\frac{m}{l} \cdot \frac{q^k}{k} < 1$ все уравнения разделяются и получается две системы. По доказанной ранее для периодического слова теореме 2, при $l \geq m q^{k^*}$, где $k^* = \left\lfloor \frac{16}{7}\sqrt{P} \right\rfloor + 5$ и $P \equiv \max(p, q)$, слово α восстановимо при $k \geq k^*$.

З а м е ч а н и е. Если $N=1$, то получаем результат, доказанный ранее в теореме 3 для слова, состоящего фактически из одного повтора подслова с периодическими суффиксом и префиксом.

4. ЗАКЛЮЧЕНИЕ

В работе получена оценка на длину подслова k , достаточную для однозначного восстановления периодического слова длины n с периодом p по мультимножеству его подслов фиксированной длины: $k \geq \left\lfloor \frac{16}{7}\sqrt{p} \right\rfloor + 5$. Также получены оценки на k для следующих слов, не являющихся периодическими, но содержащими периодические подслова. Так, для слов, состоящих из периодического префикса и периодического суффикса,

доказана возможность однозначного восстановления при достаточной длине суффикса. Показано, что слова, состоящие из периодического префикса, периодического суффикса и непериодического корня, могут быть однозначно восстановлены по мультимножеству фрагментов длины k при выполнении двух условий: малом по сравнению с префиксом корне, с одной стороны, и малом по сравнению с суффиксом префиксе, с другой. Доказана оценка для k в случае периодического слова, порожденного подсловом, в свою очередь, состоящим из двух частей: периодического суффикса и префикса.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00150.

СПИСОК ЛИТЕРАТУРЫ

1. Журавлёв Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978. Т. 33. С. 5–68.
2. Леонтьев В.К., Сметанин Ю.Г. О восстановлении векторов по набору их фрагментов // ДАН. 1988. Т. 302. № 6. С. 1319–1322.
3. Manvel B. et al. Reconstruction of sequences // Discrete Mathematics. 1991. V. 94. № 3. P. 209–219.
4. Леонтьев В.К. Распознавание двоичных слов по их фрагментам // ДАН. 1993. Т. 330. № 4. С. 434–436.
5. Lothaire M. Combinatorics on words. T. 17. Cambridge: Cambridge University Press, 1997.
6. Levenshtein V.I. Efficient reconstruction of sequences from their subsequences or supersequences // Journal of Combinatorial Theory. Series A. 2001. V. 93. № 2. P. 310–332.
7. Lin J. et al. Experiencing SAX: a novel symbolic representation of time series // Data Mining and knowledge discovery. 2007. V. 15. № 2. P. 107–144.
8. Batu T. et al. Reconstructing strings from random traces // Departmental Papers (CIS). 2004. P. 173.
9. Kannan S., McGregor A. More on reconstructing strings from random traces: insertions and deletions // Proc. International Symposium on Information Theory, 2005. ISIT 2005. IEEE. 2005. P. 297–301.
10. Smetanin Y., Ul'yanov M. Determining the characteristics of Kolmogorov complexity of time series: an approach based on symbolic descriptions. 2013.
11. Gusfield D. Algorithms on strings, trees, and sequences: Computer science and computational biology // Acm Sigact News. 1997. V. 28. № 4. P. 41–60.
12. Krasikov I., Roditty Ye. On a reconstruction problem for sequences // J. Combinatorial theory. Series A. 1997. V. 77. № 2. P. 344–348.
13. Scott A.D. Reconstructing sequences // Discrete Mathematics. 1997. V. 175. № 1–3. P. 231–238.
14. Dudik M., Schulman L.J. Reconstruction from subsequences // J. Combinatorial Theory. Series A. 2003. V. 103. № 2. P. 337–348.
15. Леонтьев В.К. Задачи восстановления слов по фрагментам и их приложения // Дискретный анализ и исследование операций 2.2. 1995. С. 26–48.
16. Borwein P., Erde'lyi T., Ko's G. Littlewood-type problems on $[0, 1]$ // Proc. London Math. Society. 1999. V. 79. № 1. P. 22–46.
17. Алексеев В.А., Сметанин Ю.Г. О возможности восстановления периодического слова по подсловом фиксированной длины // Чебышевский сборник. 2021. Т. 22. № 1 (77). С. 57–66.

ON THE WORD FRAGMENT LENGTH FOR UNAMBIGUOUS RECONSTRUCTION OF A PERIODIC WORD FROM THE COMPLETE MULTISSET OF FRAGMENTS OF FIXED LENGTH

V. A. Alekseev^a and Y. G. Smetanin^b

^a Moscow Institute of Physics and Technology (MIPT) (National Research University), Dolgoprudny, Moscow Region, Russian Federation

^b Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow, Russian Federation
Presented by Academician of the RAS K.V. Rudakov

We consider the problem of word reconstruction, where a word is to be reconstructed from a multiset of its fragments of the fixed length. Words consist of symbols from a finite alphabet. Furthermore, we impose the following restriction on the words being reconstructed: they either must be periodic or must contain a periodic word as a subword. It is shown that a periodic word with a period p can be reconstructed from a multiset

of its fragments of length k where $k \geq \left\lfloor \frac{16}{7} \sqrt{p} \right\rfloor + 5$. For a word consisting of a periodic prefix with a period q ,

repeated m times, and a periodic suffix with a period p , repeated l times, the estimate becomes $k \geq \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5$,

where $P = \max(p, q)$, subject to $l \geq mq^{\left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5}$.

Keywords: word, fragment, subword, periodic word, reconstruction from incomplete information, word reconstruction, multiset of subwords