ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Январь—март 2006. Серия 1. Том 13, № 1. 65–76

УДК 519.114

ХАРАКТЕРИЗАЦИЯ БИНАРНЫХ СЛОВ ПОДСЛОВАМИ*)

В. К. Леонтьев, М. Р. Хошманд Асл

Рассматривается задача об «определимости» бинарного слова длины n по подсловам с учётом кратности вхождения каждого подслова длины k. Подробно изучен случай $k = 2$, и найдена максимальная длина подслов, необходимых для однозначного восстановления любого слова длины n.

Пусть B^* — множество слов конечной длины над алфавитом $B = \{0, 1\}$ и B^n — множество слов длины n. На множестве B^* рассмотрим частичный порядок $\leqslant$, положив $a \leqslant b$, если слово a является подсловом слова b. Частичный порядок хорошо известен и встречается в ряде разделов комбинаторной математики, математической логики и теории вероятностей. Отметим, что термин подслово (subword) в монографии [5] фигурирует как «фактор», а термин фрагмент или подпоследовательность как «подслово». Такие изменения в терминологии имеют определённый алгебраический подтекст, но мы предпочитаем классическую терминологию.

Фрагмент и подслово являются «частями» слова и несут определённую информацию о всём слове. Как измерить эту информацию и в какой мере слово определяется своими фрагментами и подсловами? Эти вопросы можно трансформировать в точные постановки задач и притом не единственным образом. Если слово $a = (\alpha_1\alpha_2 \cdots \alpha_n) \in B^n$, то a имеет ровно $n-k+1$ подслов длины k. Некоторые из этих подслов могут совпадать. Это обстоятельство приводит к первой альтернативе в постановке проблем.

Пусть $V_k(a)$ — мультимножество всех подслов длины k в слове $a \in B^n$ и $V_{\leqslant k}(a) = \bigcup\limits_{r=1}^{k} V_r(a)$. Таким образом, $V_k(a)$ — «окрестность» порядка k слова a.

*)Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проект 05–01–01019).

Примеры.

1. Если $a = (101101) \in B^6$ и $k = 3$, то

$$V_3(a) = \{101, 011, 110, 101\} = \{(101)_2, (011), (110)\},$$

где нижний индекс означает кратность вхождения соответствующего подслова в слово a. При этом полагается, что $(x) = (x)_1$.

2. Пусть $a = (1010)$, $b = (0101)$. Тогда

$$\begin{array}{ll} V_1(a) = \{0_2, 1_2\}, & V_1(b) = \{0_2, 1_2\}, \\ V_2(a) = \{(10)_2, (01)\}, & V_2(b) = \{(01)_2, (10)\}, \\ V_3(a) = \{(101), (010)\}, & V_3(a) = \{(010), (101)\}, \end{array}$$

т. е. $V_1(a) = V_1(b)$, $V_3(a) = V_3(b)$, $V_2(a) \neq V_2(b)$.

Определение. Слова $x, y \in B^n$ называются $k(\leqslant k)$-*эквивалентными*, если $V_k(x) = V_k(y)$ $\left(\bigcup_{r=1}^{k} V_r(x) = \bigcup_{r=1}^{k} V_r(y)\right)$ и обозначаются $x \overset{k}{\sim} y$ $(x \overset{\leqslant k}{\sim} y)$.

Отметим, что без учёта кратности слова $x = (10)^n$ и $\bar{x} = (01)^n$ неразличимы по всем подсловам длины не более $2n - 1$. Минимальная длина подслов, необходимая для различения этих слов, равна длине исходных слов. Если в качестве «частей» слов x и $\bar{x}$ длины $m = 2n$ взять их «фрагменты», то аналогичная задача (без учёта кратности) имеет ответ $m+1$, ибо любые два слова длины n различимы по фрагментам длины не более $m + 1$ [2, 3].

Основная цель настоящего исследования — рассмотрение задачи об определении слова по подсловам с учётом кратности их вхождения. В дальнейшем изложении будем использовать результаты из [1, 2].

Обозначим через $W_k(a)$ — класс k-эквивалентности, порождённый словом $a \in B^n$. Формально $W_k(a) = \{x \mid x \in B^n \quad x \overset{k}{\sim} a\}$. Сначала найдём описание класса $W_k(a)$ в терминах булевой алгебры. Используя формализм из [2], рассмотрим характеристическое множество

$$V^{(k)} = \{v_1, v_2, \cdots, v_{n-k+1}\},$$

где $V^{(k)}$ — это множество всех подслов длины k слова $N_n = (12 \ldots n)$, т. е. $v_1 = (12 \ldots k)$, $v_2 = (23 \ldots k+1)$, $\ldots$, $v_{n-k+1} = (n-k+1\, n-k+2\, \ldots\, n)$. В терминах характеристического множества $V^{(k)}$ процесс порождения всех подслов длины k слова $x = (x_1 x_2 \ldots x_n)$ можно описать следующим образом:

$$
\begin{aligned}
&< x, v_1 >= (x_1 x_2 \dots x_k),\\
&< x, v_2 >= (x_2 x_3 \dots x_{k+1}),\\
&\qquad\qquad \dots\\
&< x, v_{n-k+1} >= (x_{n-k+1} x_{n-k+2} \dots x_n).
\end{aligned}
$$

Пусть $v = (i_1 i_2 \dots i_k)$, $e = (e_1 e_2 \dots e_k)$. Через $C_n(v, e)$ обозначим множество решений уравнения

$$< x, v >= e, \tag{1}$$

где $x = (x_1 x_2 \cdots x_n)$. Используя дизъюнкцию и конъюнкцию, множество $C_n(v, e)$ можно описать следующим образом.

Лемма 1. *Множество $C_n(v, e)$ состоит из слов $(x_1 x_2 \cdots x_n)$ таких, что $A(v, e) = x_{i_1}^{e_1} x_{i_2}^{e_2} \dots x_{i_k}^{e_k} = 1$.*

Пример. Пусть $v = (123)$, $e = (010)$, $n = 5$. Тогда уравнение (1) имеет вид $< (x_1 x_2 x_3 x_4 x_5), (123) >= (010)$, множество его решений описывается конъюнкцией $A(v, e) = \bar{x}_1 x_2 \bar{x}_3$ и множество $C_n(v, e)$ имеет вид

$$C_n(v, e) = \{01000, 01001, 01010, 01011\}.$$

Алгоритм построения класса эквивалентности $W_k(a)$

Входные данные: характеристическое множество $V^{(k)}$ и слово $a \in B^n$.

1) Находятся слова e_i по формуле $< a, v_i >= e_i,\ 1 \leqslant i \leqslant n - k + 1$.

2) Образуются элементарные конъюнкции A_{ij}, соответствующие уравнениям $< a, v_i >= e_j,\ \ 1 \leqslant i, j \leqslant n - k + 1$, исходя из леммы 1.

3) Строится класс $W_k(a)$ как множество единиц булевой функции, заданной в виде логического перманента матрицы $A_k = \|A_{ij}\|$.

Пример. Пусть $n = 5$, $k = 3$ и $a = (10010)$. Тогда

$$A_3 = \left\| \begin{matrix} x_1\bar{x}_2\bar{x}_3 & \bar{x}_1\bar{x}_2 x_3 & \bar{x}_1 x_2 \bar{x}_3 \\ x_2\bar{x}_3\bar{x}_4 & \bar{x}_2\bar{x}_3 x_4 & \bar{x}_2 x_3 \bar{x}_4 \\ x_3\bar{x}_4\bar{x}_5 & \bar{x}_3\bar{x}_4 x_5 & \bar{x}_3 x_4 \bar{x}_5 \end{matrix} \right\|$$

и $\operatorname{per} A_3 = x_1\bar{x}_2\bar{x}_3 x_4 \bar{x}_5 \vee \bar{x}_1\bar{x}_2 x_3 \bar{x}_4 \bar{x}_5 \vee \bar{x}_1 x_2 \bar{x}_3 \bar{x}_4 x_5$, что соответствует классу эквивалентности $W_3(a) = \{10010, 00100, 01001\}$.

Класс эквивалентности $W_{\leqslant k}(a)$ слова a можно описать аналогично с помощью логического перманента матрицы $A_{\leqslant k}$, построенной из матриц

A_i следующим образом:

$$A_{\leqslant k} = \left\| \begin{matrix} A_k & 0 & 0 & \dots & 0 \\ 0 & A_{k-1} & 0 & \dots & 0 \\ 0 & 0 & A_{k-2} & \dots & 0 \\ & & \dots & & \\ 0 & 0 & 0 & \dots & A_1 \end{matrix} \right\|.$$

Способы алгебраического описания классов эквивалентности $W_k(a)$ для фрагментов представлены в работах [2, 3]. Эти же методы применимы и для подслов. В этой статье ограничимся случаем $k \leqslant 2$, так как из [2] следует возможность сравнения результатов.

Характеризация $\leqslant 2$-эквивалентности

1) Случай $k = 1$ является тривиальным и выражается следующим образом: $x \overset{1}{\sim} y \rightleftarrows \|x\| = \|y\|$, где $\|x\| = \sum\limits_{i=1}^{n} x_i$ — вес слова $x \in B^n$.

2) В случае $k = 2$ при $x \overset{2}{\sim} y$ имеет место соотношение

$$V_2(x) = V_2(y). \tag{2}$$

Пусть $a, x \in B^*$. Через $t_a(x)$ обозначим кратность вхождения слова a в слово x в качестве подслова. Условие (2) можно записать в виде:

$$t_{00}(x) = t_{00}(y), \quad t_{01}(x) = t_{01}(y), \quad t_{10}(x) = t_{10}(y), \quad t_{11}(x) = t_{11}(y).$$

При этом выполняется очевидное соотношение

$$t_{00}(x) + t_{01}(x) + t_{10}(x) + t_{11}(x) = |x| - 1, \tag{3}$$

где $|x|$ — длина слова x, так как число подслов длины два в слове x равно $|x| - 1$.

Теперь найдём выражение функции $t_{\alpha\beta}(x)$ через «параметры» слова x. Пусть $x = (x_1 x_2 \dots x_n) \in B^n$. Тогда справедливы следующие соотношения:

$$\begin{aligned} t_{00}(x) &= \sum_{i=1}^{n-1} (1 - x_i)(1 - x_{i+1}), & t_{01}(x) &= \sum_{i=1}^{n-1} (1 - x_i) x_{i+1}, \\ t_{10}(x) &= \sum_{i=1}^{n-1} x_i (1 - x_{i+1}), & t_{11}(x) &= \sum_{i=1}^{n-1} x_i x_{i+1}. \end{aligned} \tag{4}$$

Ясно, что полиномы $t_{\alpha\beta}(x)$ не являются независимыми, и из (4) легко следуют формулы

$$\begin{aligned} t_{00}(x) &= (n-1) - 2\|x\| + (x_1 + x_n) + t_{11}(x), \\ t_{01}(x) &= \|x\| - x_1 - t_{11}(x), \\ t_{10}(x) &= \|x\| - x_n - t_{11}(x). \end{aligned} \tag{5}$$

Слово x представим в виде «серий» $x = \gamma^{m_1}\bar{\gamma}^{m_2}\cdots\gamma^{m_r}$. Здесь $\gamma \in \{0,1\}$ и $\bar{\gamma}$ — логическое отрицание γ. Числа m_i — это длины серий, которые удовлетворяют следующим условиям:

$$\sum_{i=1}^{r} m_i = n, \quad m_i \geqslant 1, \quad 1 \leqslant i \leqslant r.$$

Число r-серий слова x обозначим через $S(x)$.

Пример. Если $x = 110001011$, то $x = 1_2 0_3 1 0 1_2$ и $S(x) = 5$.

Серии γ^{m_1}, γ^{m_3}, ... называются *нечётными*, а серии $\bar{\gamma}^{m_2}$, $\bar{\gamma}^{m_4}$, $\cdots$ — *чётными*. Число нечётных серий обозначается через $S^1(x)$, а чётных — через $S^0(x)$.

Очевидно, что

$$\begin{aligned} &S^0(x) = S^1(x) = r/2 && \text{при } r \equiv 0 \pmod 2, \\ &S^0(x) = (r-1)/2,\ S^1(x) = (r+1)/2 && \text{при } r \equiv 1 \pmod 2. \end{aligned} \tag{6}$$

Так как $t_{11}(1^m) = m - 1$, то при $\gamma = 1$ из (4) и (6) получаем

$$\begin{aligned} &t_{11}(x) = \sum_{i=1}^{\lfloor (r+1)/2 \rfloor} (m_{2i-1} - 1) = \|x\| - S^1(x), \\ &t_{00}(x) = \sum_{i=1}^{\lfloor r/2 \rfloor} (m_{2i} - 1) = n - \|x\| - S^0(x), \\ &t_{10}(x) = t_{01}(x) = (r-1)/2 && \text{при } r \equiv 1 \pmod 2, \\ &t_{10}(x) = t_{01}(x) + 1 = r/2 && \text{при } r \equiv 0 \pmod 2. \end{aligned} \tag{7}$$

Аналогичные формулы можно получить и для $\gamma = 0$, так как $t_a(x) = t_{\bar{a}}(\bar{x})$, где $\bar{y}$ — логическое отрицание слова y.

Формулы (5), (7) показывают, что все функции $t_{\alpha\beta}(x)$ могут быть выражены через $t_{11}(x)$, число серий $S(x)$ и «крайние» буквы x_1 и x_n слова x. Поэтому справедлив следующий критерий 2-эквивалентности слов из B^n.

Теорема 1. *Слова x и y из B^n являются 2-эквивалентными тогда и только тогда, когда выполняются условия*:

1) $S(x) = S(y)$,

2) $\|x\| - \|y\| = x_1 - y_1 = x_n - y_n$.

Доказательство. Если $x = (x_1 x_2 \cdots x_n)$, $y = (y_1 y_2 \cdots y_n)$ и $x \overset{2}{\sim} y$, то из (7) получаем

$$\|x\| - S^1(x) = \|y\| - S^1(y), \quad n - \|x\| - S^0(x) = n - \|y\| - S^0(y).$$

Отсюда следует, что $S(x) = S(y)$. Далее из (5) следует, что

$$\begin{aligned} t_{10}(x) = \|x\| - x_n - t_{11}(x) = \|y\| - y_n - t_{11}(y), \\ t_{01}(x) = \|x\| - x_1 - t_{11}(x) = \|y\| - y_1 - t_{11}(y). \end{aligned} \tag{8}$$

Так как $t_{11}(x) = t_{11}(y)$, то из (8) следует условие 2).

Если выполнены условия 1) и 2), то из (7) получаем

$$t_{10}(x) = t_{10}(y), \quad t_{01}(x) = t_{01}(y). \tag{9}$$

Далее из условия 2) и соотношений (5), (9)следует, что

$$t_{11}(x) - t_{11}(y) = \|x\| - x_n - t_{10}(x) - (\|y\| - y_n - t_{10}(y)) = 0.$$

Аналогично получаем, что $t_{00}(x) = t_{00}(y)$, учитывая (3).

Примеры.

1) Пусть $x = (100110)$ и $y = (001011)$. Тогда $S(x) = S(y) = 4$, $\|x\| = \|y\| = 3$. Далее $x_6 - y_6 = -1$, $x_1 - y_1 = 0$. Поэтому $x \overset{2}{\not\sim} y$. Действительно $t_{10}(x) = 2,\ t_{10}(y) = 1$.

2) Если $x = (100100)$ и $y = (101000)$, то $V_2(x) = V_2(y) = \{(00)_2, (01), (10)_2\}$. В тоже время $S(x) = S(y) = 4$, $\|x\| = \|y\| = 2$, $x_6 - y_6 = x_1 - y_1 = 0$.

3) Если $x = (100101)$ и $y = (010100)$, то $S(x) = S(y) = 5$, $\|x\| = 3$, $\|y\| = 2$, $x_6 - y_6 = x_1 - y_1 = 1$. Таким образом $x \overset{2}{\sim} y$. Легко видеть, что

$$V_2(x) = V_2(y) = \{(00), (01)_2, (10)_2\}.$$

Теорема 1 позволяет провести проверку 2-эквивалентности слов из B^n за линейное по входу время, так как вычисление $S(x)$ и $\|x\|$ может быть сделано с линейной сложностью от n. Отметим также, что в [4] предложен алгоритм сложности $O(n^4)$ установления 2-эквивалентности слов длины n над произвольным конечным алфавитом.

Теорема 1 показывает, что над словами из класса эквивалентности $W_2(x)$ можно проводить определённые преобразования, не выводящие за пределы этого класса. Далее используются некоторые из таких преобразований для описания случаев, когда бинарное слово однозначно определяется своими подсловами длины 2.

1. Итак, пусть $x = \gamma^{m_1}\bar{\gamma}^{m_2}\cdots\gamma^{m_r}$ и $\bar{S}(x) = (m_1, m_2, \cdots, m_r)$ — вектор серий слова x. Теорема 1 утверждает, что серии одинаковой чётности можно произвольно переставлять между собой и эти перестановки не выводят за пределы класса $W_2(x)$. Точнее, пусть S_r^0 — подгруппа симметрической группы S_r, действующая отдельно на «чётные» и «нечётные» элементы множества $\{1, 2, \dots, r\}$. Тогда действие S_r^0 на слово x определяется следующим образом. Пусть $g \in S_r^0$. Тогда

$$g(x) = \gamma^{m_{g(1)}}\bar{\gamma}^{m_{g(2)}}\cdots\gamma^{m_{g(r)}}.$$

Пример. Если $x = (0_2 1 0_3 1_4)$ и $g \in S_4^0$, где

$$g = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix},$$

то $g(x) = (0_3 1 0_2 1_4)$. Таким образом, g действует на «чётные» серии тождественным образом, а «нечётные» серии переставляет между собой.

2. Другое преобразование состоит в том, что любые две серии одинаковой чётности можно изменять по длине, но так, чтобы сумма их длин сохранялась и ни одна из серий не исчезла, т. е. не стала нулевой.

Пример. Пусть $\bar{S}(x) = (24421)$ и $T(x)$ — преобразование, изменяющее вторую и четвёртую серию указанным выше способом. Тогда $T(x) = \gamma^2\bar{\gamma}\gamma^4\bar{\gamma}^5\gamma$.

Нетрудно понять, что первое из указанных выше преобразований является частным случаем второго.

С помощью введённых преобразований можно описать все классы эквивалентности, состоящие из одного слова. Будем проводить классификацию по длине серий из $S(x)$.

1) $S(x) = 1$. Тогда $x = \gamma^n$, $V_2(x) = \{(\gamma\gamma)^{n-1}\}$ и $|W_2(x)| = 1$.

2) $S(x) = 2$. Тогда $x = \gamma^p\bar{\gamma}^{n-p}$ и $V_2(x) = \{(\gamma\gamma)^{p-1}, (\gamma\bar{\gamma}), (\bar{\gamma}\bar{\gamma})^{n-p-1}\}$. В этом случае $|W_2(x)| = 1$.

3) $S(x) = 3$. Тогда $x = \gamma^p\bar{\gamma}^q\gamma^r$. Вторым преобразованием слово x можно привести к 2-эквивалентному слову $x' = \gamma^{p'}\bar{\gamma}^q\gamma^1$. Если $|W_2(x)| = 1$, то $x = x'$, т. е. $r = 1$. Аналогично получаем, что $p = 1$. Таким образом, $x = \gamma\bar{\gamma}^{n-2}\gamma$. Отсюда следует, что $V_2(x) = \{(\bar{\gamma}\bar{\gamma})^{n-3}, (\gamma\bar{\gamma}), (\bar{\gamma}\gamma)\}$, и при

$n \geqslant 4$ класс эквивалентности $W_2(x)$ состоит только из слова x. Если же $n = 3$, то $x = \gamma\bar{\gamma}\gamma$ и $x \overset{2}{\sim} \bar{x}$, т. е. $|W_2(x)| \neq 1$.

4) $S(x) \geqslant 4$, т. е. $x = \gamma^{m_1}\bar{\gamma}^{m_2} \cdots \gamma^{m_r}$, где $r \geqslant 4$.

Применяя второе преобразование также как и в предыдущем случае и используя равенство $|W_2(x)| = 1$, можно показать, что $x = \gamma\bar{\gamma}\gamma \underset{r}{\cdots} \gamma$. Отсюда следует, что если $r \equiv 0 \pmod 2$, то $|W_2(x)| = 1$; если же $r \equiv 1 \pmod 2$, то $|W_2(x)| \neq 1$.

Таким образом, справедливо следующее утверждение.

Теорема 2. *Слово $a \in B^n$ однозначно определяется своими подсловами длины 2 если и только если выполняются следующие условия*:

1) $S(a) \leqslant 2$,

2)$a = \gamma\bar{\gamma}^{n-2}\gamma$ *и* $n \geqslant 4$,

3) $a = \gamma\bar{\gamma}\gamma\bar{\gamma}\ldots\gamma\bar{\gamma}$ *и* $n \equiv 0 \pmod 2$ $(n \geqslant 2)$.

Отметим также, что в [4] предложено «алгоритмическое» решение задачи о восстановимости, если это возможно, слова длины n по подсловам длины 2 в случае произвольного конечного алфавита. Сложность такого алгоритма равна $O(n^4)$.

В заключение рассмотрим задачу о проверке $\leqslant 2$ -эквивалентности. Из теоремы 2 легко получить следующее утверждение.

Теорема 3. *Слова x и y из B^n являются $\leqslant 2$ -эквивалентными если и только если выполняются следующие условия*:

1) $S(x) = S(y)$,

2) $\|x\| = \|y\|$,

3) $x_1 = y_1, x_n = y_n$.

Теоремы 1 и 3 показывают, что имеется существенная разница в описании классов эквивалентностей по подсловами длины не более 2 и по фрагментам длины не более 2 [5]. Одно из отличий состоит в том, что 2-эквивалентность по фрагментам влечёт за собой 1-эквивалентность, а 2-эквивалентность по подсловам не имеет таких следствий.

Как было отмечено выше, класс эквивалентности $W_{\leqslant k}(a)$ определяет ту меру точности, с которой можно определить слово $a \in B^n$, если известны все его подслова длины, не превосходящей k. Ясно, что с увеличением k эта «точность» возрастает, что выражается в уменьшении мощности класса эквивалентности. Другими словами, справедливы следующие включения

$$W_n(a) \subseteq W_{n-1}(a) \subseteq \ldots \subseteq W_1(a). \tag{10}$$

Ясно, что цепочка (10) стабилизируется, начиная с некоторого номера, и этот номер определяет ту минимальную длину подслов, которые однозначно определяют слово a. Если этот «минимальный» номер обозначить через $\lambda(a)$, то $\lambda(a)$ — минимальная длина подслов, однозначно определяющих слово a. Ясно, что $|W_{\leqslant\lambda(a)}(a)| = 1$.

Примеры.
1) Если $a = \gamma^n$, то $\lambda(a) = 1$.
2) Если $a = \gamma^{n-1}\bar{\gamma}$, то $\lambda(a) = 2$.
3) Если $\varepsilon_{2n} = (10)^n$, то $\lambda(\varepsilon_{2n}) = 2$.
Пусть $\lambda(n) = \max\limits_{a\in B^n} \lambda(a)$.

Теорема 4. *Справедливо соотношение* $\lambda(n) = \lfloor n/2\rfloor + 1$.

Доказательство. Рассмотрим следующие два слова $a = 1^{m-1}01^{m-2}$ и $b = 1^{m-2}01^{m-1}$. Тогда $|a| = |b| = 2m - 4$ и

$$\begin{aligned} &t_1(a) = t_1(b) = 2m-3,\\ &t_{11}(a) = t_{11}(b) = (m-2)+(m-3) = 2m-5,\\ &\qquad\qquad \cdots\\ &t_{1^{m-2}}(a) = t_{1^{m-2}}(b) = 3. \end{aligned}$$

Все другие подслова слов a и b с двумя и тремя сериями и длины не более $m-2$ входят в a и b с кратностью 1. Поэтому $V_{\leqslant m-2}(a) = V_{\leqslant m-2}(b)$. Следовательно, $\lambda_n(a) \geqslant m-1 = \frac{n}{2}+1$.

Для восстановления слова $x \in B^n$ по подсловам длины не более k мы используем процедуру, описанную в [1].

Пусть $x = (x_1, x_2, \cdots, x_n) \in B^n$. Все подслова слова x длин $k, k-1, k-2, \ldots, 1$ запишем в виде таблицы.

$$\begin{array}{llll} x_1x_2\cdots x_k & x_1x_2\cdots x_{k-1} & & x_1\\ x_2x_3\cdots x_{k+1} & x_2x_3\cdots x_k & & x_2\\ \quad\vdots & \quad\vdots & \cdots & \\ x_{n-k+1}\cdots x_n & x_{n-k+1}\cdots x_{n-1} & & \vdots\\ & x_{n-k+2}\cdots x_n & & \\ & & & x_{n-1}\\ & & & x_n. \end{array} \tag{11}$$

«Реально» расположение строк таблицы (11) может быть совершенно произвольным. В первом «блоке» все строки имеют длину k и их число равно $n-k+1$, во втором «блоке» все строки имеют длину $k-1$ и их число равно $n-k+2$ и т. д. Другим «инвариантом» является сумма элементов в каждом столбце каждого блока. Действительно, как бы мы не меняли строки внутри блока, эта сумма остаётся неизменной. Это позволяет выписать следующую систему уравнений

$$\left\{\begin{array}{ll} x_1+x_2+\cdots+x_{n-k+1}=t_1 & x_1+x_2+\cdots+x_{n-k+2}=r_1 \\ x_2+x_3+\cdots+x_{n-k+2}=t_2 & x_2+x_3+\cdots+x_{n-k+3}=r_2 \\ \cdots & \cdots \\ x_k+x_{k+1}+\cdots+x_n=t_k & x_{k-1}+x_k+\cdots+x_n=r_{k-1}. \end{array}\right. \tag{12}$$

Систему (12) можно выписать, заметив, что столбцы первого блока — подслова длины $n-k+1$ слова $x=(x_1x_2\cdots x_n)$; столбцы второго блока — подслова длины $n-k+2$ слова x и т. д. Таким образом, каждому подслову длины $n-k+i$ слова $x=(x_1x_2\cdots x_n)$ соответствует уравнение из системы (12), левая часть которого есть сумма букв этого подслова, а правая — некоторое фиксированное число. Теперь запишем следующие две подсистемы системы (12)

$$\begin{array}{l} \left\{\begin{array}{l} x_1+x_2+\cdots+x_n=t_1 \\ x_2+\cdots+x_n=t_2 \\ \cdots \\ x_k+\cdots+x_n=t_k \end{array}\right. \\ \left\{\begin{array}{l} x_1+x_2+\cdots+x_n=t_1 \\ x_1+x_2+\cdots+x_{n-1}=r_1 \\ \cdots \\ x_1+\cdots+x_{n-k+1}=r_{k-1} \end{array}\right. \end{array} \tag{13}$$

с общим числом уравнений $2k$.

Из системы (13) находим $x_1=t_1-t_2$, $x_2=t_2-t_3$, $\cdots$, $x_{k-1}=t_{k-1}-t_k$ и $x_n=t_1-r_2$, $x_{n-1}=r_1-r_2$, $\cdots$, $x_{n-k+2}=r_{k-2}-r_{k-1}$. Отсюда следует, что если $k\geqslant\frac{n}{2}+1$, то система (12) имеет единственное решение. Теорема 3 доказана.

Примеры.

1) Пусть $n=6$, $k=4$ и $a=(101001)$. Выпишем все фрагменты слова a длин $1,2,3,4$. Имеем

$$\begin{array}{llll} x_1x_2x_3x_4 & x_1x_2x_3 & x_1x_2 & x_1 \\ x_2x_3x_4x_5 & x_2x_3x_4 & x_2x_3 & x_2 \\ x_3x_4x_5x_6 & x_3x_4x_5 & x_3x_4 & x_3 \\ & x_4x_5x_6 & x_4x_5 & x_4 \\ & & x_5x_6 & x_5 \\ & & & x_6. \end{array}$$

Этой системе соответствует следующее расположение подслов слова a

$$
\begin{array}{llll}
1010 & 101 & 10 & 1\\
0100 & 010 & 01 & 0\\
1001 & 100 & 10 & 1\\
 & 001 & 00 & 0\\
 & & 01 & 0\\
 & & & 1.
\end{array}
\tag{14}
$$

Таблице (14) соответствует система уравнений

$$
\begin{cases}
x_1+x_2+x_3=2 & x_1+\dots+x_4=2 \quad x_1+\dots+x_5=2 \quad x_1+\dots+x_6=3\\
x_2+x_3+x_4=1 & x_2+\dots+x_5=1 \quad x_2+\dots+x_6=2\\
x_3+x_4+x_5=1 & x_3+\dots+x_6=2\\
x_4+x_5+x_6=1.
\end{cases}
$$

Решив эту систему, получаем $x=(101001)=a$.

Приведённые выше рассуждения позволяют получить универсальную верхнюю границу мощности произвольного класса эквивалентности $W_{\leqslant k}(a)$.

Следствие. *Справедливо следующее неравенство*

$$|W_{\leqslant k}(a)| \leqslant 2^{n-2(k-1)},$$

где $a\in B^n$ *и* $k\leqslant n/2+1$.

В заключение отметим следующее обстоятельство. Если сравнить рассмотренную ранее задачу о восстановлении слова по его фрагментам без учёта кратности [3] и задачу о восстановлении слова по его подсловам с учётом кратности вхождения подслов, то, как показывает теорема 4, длина «части» слова, необходимая для получения однозначного решения, в обоих случаях одинакова.

ЛИТЕРАТУРА

1. **Зенкин А. И., Леонтьев В. К.** Об одной неклассической задаче распознавания // Ж. вычисл. матем. и матем. физики. 1984. Т. 24, № 6. С. 925–931.
2. **Леонтьев В. К.** Задачи воччтановления слов по фрагментам и их приложения // Дискрет. анализ и исслед. операций. 1975. Т. 2, № 2. С. 26–48.
3. **Леонтьев В. К., Сметанин Ю. Г.** О восстановлении векторов по набору их фрагментов // Докл. АН СССР. 1988. Т. 302, № 6. С. 1319–1322.

4. **Сметанич Я. С.** О восстановлении слов // Математическая логика, теория алгоритмов и теория множеств. М.: Наука, 1973. С. 183–202. (Труды Математической института им. В. А. Стеклова. Т. 133).
5. **Lothaire M.** Combinatorics on words. Reading, Mass.: Addison-Wesley. Publ. Co, 1983. (Encyclopedia of mathematics and its applications; V. 17)

Адрес авторов:

Вычислительный центр РАН,
Вавилова, 40,
119991 Москва,
Россия.
E-mail: vkleontiev@mtu-net.ru

Статья поступила
21 июня 2005 г.