ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Апрель—июнь 1995. Том 2, № 2, 26–48

УДК 519.12

ЗАДАЧИ ВОССТАНОВЛЕНИЯ СЛОВ ПО ФРАГМЕНТАМ И ИХ ПРИЛОЖЕНИЯ*)

В. К. Леонтьев

Рассматриваются задачи восстановления слов по фрагментам и их приложения в распознавании образов и теории информации. Приводится ряд новых результатов, характеризующих возможности распознавания и декодирования в терминах комбинаторных и геометрических конструкций.

Введение

В работе обсуждаются постановки задач и результаты в области исследования, объединяемой проблемой восстановления слов по их фрагментам. Интерес к этой тематике возник сравнительно недавно, и исследуемые понятия, задачи и методы их решения находятся в стадии формирования. В более широком контексте задачи существования и восстановления объекта из определенного класса по некоторой информации о множестве его «частей» и взаимосвязи этих «частей» между собой можно рассматривать как задачи распознавания образов и теории информации [1–3]. Такой подход оказывается полезным, и в настоящей работе обсуждаются новые модели, возникающие в связи с приложениями задач восстановления слов к распознаванию речи и теории надежного приема сообщений, когда известны лишь некоторые их фрагменты.

В теоретическом плане рассматриваемые задачи о восстановлении слов приводят к необходимости исследования новых математических конструкций, отражающих взаимосвязи свойств последовательностей символов и множеств их подпоследовательностей [4–10]. В работе приводится ряд результатов, а также формулируются комбинаторные задачи, являющиеся фундаментом проводимых исследований. При этом фактически содержание публикуемой работы несколько шире, чем указано в ее заголовке: в ней освещается ряд общих вопросов «комбинаторики слов».

Некоторые из приводимых ниже результатов содержатся в работах [5, 6, 11–13], другие же являются новыми и публикуются впервые.

*) Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (код проекта 93-01-00449).

§ 1. Задачи о покрытии на множестве слов

Пусть E^n — множество двоичных слов длины n и A^n — множество двоичных слов длины не более n. На множестве A^n рассмотрим частичный порядок $\prec$, приняв по определению $a \prec b$, если последовательность букв слова a является подпоследовательностью последовательности букв слова b. Другими словами, слово a может быть получено из b вычеркиванием в нем некоторых букв. В этом случае будем говорить, что слово a является *фрагментом* слова b [6, 10]. Структура этого частично упорядоченного множества довольно сложная, и приведенные ниже результаты есть то немногое, что удалось о нем выяснить.

Диаграммой Хассе конечного частично упорядоченного множества X называется *ориентированный граф*, вершинами которого являются элементы из X и две вершины x, y соединены дугой, исходящей из x и заходящей в y, если $x \succ y$ и не существует $z \in X$ такого, что $x \succ z \succ y$.

Следующие утверждения можно рассматривать как геометрические факты, относящиеся к диаграмме Хассе множества A^n с порядком $\prec$.

Лемма 1. *Для любого двоичного слова b длины k имеется* $\sum_{i=k}^{n} \binom{n}{i}$ *двоичных слов a длины n, удовлетворяющих условию $a \succ b$.*

Доказательство, применение и различные содержательные интерпретации этого утверждения можно найти в работах [5, 14, 15].

Чтобы получить более подробную информацию о диаграмме Хассе, воспользуемся представлением двоичного слова $a \in E^n$ в виде серий: $a = \gamma^{t_1}\bar{\gamma}^{t_2}\ldots\gamma^{t_{r-1}}\bar{\gamma}^{t_r}$, где $\gamma \in \{0,1\}$ и $\bar{\gamma}$ — логическое отрицание буквы γ. Слово γ^t называется *серией*, а число t — *длиной серии*. Например, пусть $a = 10011100$. Тогда $a = 10^2 1^3 0^2$.

Лемма 2. *Пусть $a \in E^{n-1}$. Тогда число слов b из E^n таких, что $b \succ a$, равно числу серий в слове a.*

Из лемм 1 и 2 следует, что в диаграмме Хассе введенного частичного порядка степень исхода каждой вершины r-го уровня равна $r+2$, а степень захода вершины, соответствующей слову a, равна числу серий этого слова. Распределение степеней захода описывается следующей простой леммой.

Лемма 3. *На r-м уровне диаграммы Хассе имеется* $2\binom{r-1}{k-1}$ *вершин со степенью захода, равной k.*

Нетрудно показать, что «средняя» степень захода вершин r-го уровня равна $(r+1)/2$. Таким образом, примерная картина r-го уровня диаграммы Хассе рассмотренного порядка выглядит так: степень исхода каждой вершины равна $r+2$, а степень захода в среднем равна $(r+1)/2$.

Следует также отметить, что построение диаграмм Хассе индивидуальных слов связано с вычислением так называемых параметров Уитни, т. е. числа различных фрагментов слова a одной и той же длины. По-видимому, нахождение явного вида параметров Уитни в общем случае представляет собой серьезную комбинаторную проблему [15, 16].

Пусть $t(a,k)$ — число различных фрагментов длины k в слове a. Приведем некоторые простые оценки для величины $t(a,k)$, которые могут оказаться полезными.

Положим по определению $t(a,0)=0$, и пусть ab — конкатенация слов $a=\alpha_1\ldots\alpha_n$ и $b=\beta_1\ldots\beta_n$, т. е. $ab=\alpha_1\alpha_2\ldots\alpha_n\beta_1\beta_2\ldots\beta_n$.

Утверждение. *При любых a и b справедливы неравенства*

$$\max_p\{t(a,p)t(b,k-p)\}\leqslant t(ab,k)\leqslant\sum_{p=0}^{k}t(a,p)t(b,k-p).$$

В частности, если число серий в слове a равно m и $k\leqslant|a|-1$, то $t(a,k)\geqslant m$.

Отметим, что представляет интерес нахождение экстремумов функции $t(a,k)$ следующего типа.

Пусть слово a состоит из m серий, а длина i-й серии равна t_i, $1\leqslant i\leqslant m$, т. е. $a=\gamma^{t_1}\bar{\gamma}^{t_2}\ldots\gamma^{t_m}$, и $t(k,t_1,\ldots,t_m)$ обозначает число фрагментов длины k в слове a. Тогда имеем $t(a,k)=t(k,t_1,\ldots,t_m)$. Положим

$$t(k,m,n)=\min(k,t_1,\ldots,t_m),\tag{1}$$

где минимум берется по всем наборам $\{t_1,\ldots,t_m\}$ натуральных чисел таким, что

$$\sum_{i=1}^{m}t_i=n,\quad t_i\geqslant 1,\ 1\leqslant i\leqslant m.$$

Ясно, что функция (1) представляет собой лишь один из многочисленных примеров функций шенноновского типа, связанных с $t(a,k)$.

В ряде случаев представляет интерес и является непростой задача вычисления параметров Уитни специальных слов.

Пример. Пусть $\xi_n=1010\ldots10$ — слово длины n и $t_{n,k}$ — число слов длины k, являющихся фрагментами слова ξ_n.

Утверждение. *При любых n и k справедливо соотношение*

$$t_{n,k}=\binom{k-1}{2k-n-1}+2\sum_{r=2k-n+1}^{k}\binom{k-1}{r-1}.\tag{2}$$

Это утверждение дает ответ на один из вопросов, сформулированных в работе [15].

В формуле (2) суммирование начинается с нуля, если нижний предел для r отрицательный, и $\binom{n}{m} = 0$ при $m > n$.

Слово ξ_n обладает многими интересными свойствами и, в частности, определенным свойством универсальности, так как при $k \leqslant \lfloor n/2 \rfloor$ слово ξ_n содержит в качестве фрагментов все слова длины k и при $n = 2k$ слово ξ_n является словом минимальной длины, содержащим в качестве фрагментов все слова длины k. Нетрудно видеть, что ξ_{2k} не является единственным универсальным словом минимальной длины, и поэтому возникает задача описания всех таких слов длины $2k$. В связи с «монотонностью» свойства универсальности представляет интерес описание всех тупиковых универсальных слов, т. е. слов, теряющих свойство универсальности после удаления любой буквы.

Рассмотрим следующие задачи.

Задача о покрытии. Дан набор слов $\{a_1, a_2, \ldots, a_m\}$ в алфавите $\{0, 1\}$, вообще говоря, разной длины. Требуется найти слово a минимальной длины, которое в качестве фрагментов содержит все слова a_i, $1 \leqslant i \leqslant m$. В работе [17] эта задача названа задачей о *минимальной общей надпоследовательности.*

Пример. Пусть $a_1 = 1010$, $a_2 = 010$, $a_3 = 111$. Нетрудно проверить, что слово $a = 10101$ является минимальной общей надпоследовательностью слов a_1, a_2, a_3.

Другим примером минимальной общей надпоследовательности является приведенное выше слово ξ_{2k}, универсальное для множества всех двоичных слов длины k.

Для множества двоичных слов длины n, каждое из которых содержит ровно k единиц, минимальная общая надпоследовательность найдена в работе [9]. Упомянем также известную недоказанную гипотезу о том, что длина минимальной общей надпоследовательности для множества всех перестановок n-элементного множества равна $n^2 - 2n + 4$ при любом $n \geqslant 3$ (см., например, [18]).

Задача о кратном покрытии. Дан набор слов $a_1, a_2, \ldots, a_m$ с кратностями $\nu_1, \nu_2, \ldots, \nu_m$. Требуется найти слово a минимальной длины, в котором каждое слово a_i, $1 \leqslant i \leqslant m$, в качестве фрагмента встречается не менее ν_i раз.

Отметим, что для сформулированных задач о покрытии интерес представляет и нахождение всех минимальных и всех тупиковых покрытий, т. е. покрытий, перестающих быть таковыми после удаления любой буквы.

Близкой к задаче о покрытии в приведенной выше формулировке является задача о *максимальной общей подпоследовательности* [17].

Задача о максимальной общей подпоследовательности. Дан набор слов $\{a_1, a_2, \ldots, a_m\}$ в алфавите $\{0,1\}$. Требуется найти слово a максимальной длины, являющееся фрагментом каждого слова a_i, $1 \leqslant i \leqslant m$.

Задачи о минимальной общей надпоследовательности и максимальной общей подпоследовательности тесно связаны, и в работе [18] приведен алгоритм типа динамического программирования для решения первой из этих задач. В общем виде обе задачи являются NP-полными [18]. Однако в отличие от задачи о покрытии множества системой его подмножеств для сформулированной выше задачи имеется возможность дать простые и достаточно хорошие универсальные границы для мощности минимального покрытия.

Пусть $d(a_1, a_2, \ldots, a_m)$ — длина минимальной общей надпоследовательности для множества слов $\{a_1, \ldots, a_m\}$ и

$$\gamma(a_1, a_2, \ldots, a_m) = \max\{|a_1|, |a_2|, \ldots, |a_m|\}.$$

Тогда справедливы очевидные оценки:

$$\gamma(a_1, a_2, \ldots, a_m) \leqslant d(a_1, a_2, \ldots, a_m) \leqslant 2\gamma(a_1, a_2, \ldots, a_m).$$

Общую задачу о кратном покрытии можно записать как нелинейную задачу булева программирования следующего вида.

Пусть

$$x^\sigma = \begin{cases} x & \text{при} \quad \sigma = 1, \\ 1 - x & \text{при} \quad \sigma = 0. \end{cases}$$

Тогда задача о кратном покрытии может быть выписана в следующей форме. Дано $a_r = (\sigma_1^r, \sigma_2^r, \ldots, \sigma_k^r)$. Требуется найти минимальное натуральное r такое, что

$$\sum_{1 \leqslant i_1 < i_2 \cdots < i_k \leqslant z} x_{i_1}^{\sigma_1^{r_1}} x_{i_2}^{\sigma_2^{r_2}} \ldots x_{i_k}^{\sigma_k^{r_k}} \geqslant \nu_r, \quad 1 \leqslant r \leqslant m, \; x_i = \{0,1\}.$$

Пример. Пусть $a_1 = 1010$, $a_2 = 0101$, $a_3 = 111$. Тогда оптимизационная задача состоит в нахождении минимального значения z при ограничениях

$$\sum_{1 \leqslant i_1 < i_2 < i_3 < i_4 \leqslant z} x_{i_1}(1 - x_{i_2})x_{i_3}(1 - x_{i_4}) \geqslant 1,$$

$$\sum_{1 \leqslant i_1 < i_2 < i_3 < i_4 \leqslant z} (1 - x_{i_1})x_{i_2}(1 - x_{i_3})x_{i_4} \geqslant 1,$$

$$\sum_{1 \leqslant i_1 < i_2 < i_3 \leqslant z} x_{i_1}x_{i_2}x_{i_3} \geqslant 1,$$

$$x_i = \{0,1\}.$$

Без дополнительных теоретических и экспериментальных исследований трудно судить о пользе такой формы записи задачи о покрытии. Однако подобная запись оказывается полезной при решении задачи о восстановлении слов по их фрагментам (см. ниже).

Обращаясь к задаче о покрытии, рассмотрим самый простой случай $m = 2$. Длина минимального покрытия $d(a, b)$ слов a, $b \in E^n$ удовлетворяет очевидным неравенствам

$$n + 1 \leqslant d(a, b) \leqslant 2n. \tag{3}$$

Обе границы в (3) являются достижимыми. Чтобы получить нижнюю границу в (3), в качестве a достаточно взять слово ξ_n, а в качестве b — его логическое отрицание, т. е. $b = \bar{\xi}_n$. Верхняя граница достигается на паре $a = 1^n$, $b = 0^n$.

Минимальное число вставок и удалений букв, преобразующих слово a в слово b, определяет расстояние в E^n, обозначаемое через $\rho_L(a, b)$. Оно введено впервые В. И. Левенштейном в связи с изучением кодов с исправлением выпадений, вставок и замещений символов [11].

Пример. Пусть $a = 0110$, $b = 1001$. Тогда слово $c = 101010$ является минимальным покрытием для a и b и $\rho_L(a, b) = 4$ (из a удаляются первый и четвертый символы, а на второе и третье места вставляются соответственно нуль и единица).

Существует простая связь между функциями $\rho_L(a, b)$ и $d(a, b)$.

Лемма 4. *Для любых a, $b \in E^n$ справедливо равенство*

$$d(a, b) - n = \frac{1}{2}\rho_L(a, b).$$

В дальнейшем расстояние $\rho(x, y) = d(x, y) - n$ будем называть *метрикой покрытий.*

Справедлива следующая

Лемма 5. *Если (x, y) — длина* Н.О.П. *для слов x и y из E^n, то*

$$d(x, y) + (x, y) = 2n. \tag{4}$$

Из соотношения (4) следует, что пара слов $a = 1^n$, $b = 0^n$ является единственной в E^n парой слов, расстояние между которыми в метрике покрытий равно n (в метрике Хэмминга число таких пар равно 2^{n-1}). Отметим также, что в отличие от метрики Хэмминга метрика покрытий не инвариантна относительно автоморфизмов. Поэтому, в частности, число точек в шаре фиксированного радиуса в метрике покрытий зависит от центра этого шара.

Слово, состоящее из чередующихся нулей и единиц, назовем *альтернирующим.*

Примеры. 1. Если радиус шара равен единице и центром шара является 0^n, то этот шар совпадает с шаром в метрике Хэмминга и число точек в нем равно $n+1$.

2. Если радиус шара равен единице и центром шара является альтернирующее слово длины n, то число точек в шаре равно $\binom{n+1}{2}+1$.

В общем случае число точек в шаре единичного радиуса зависит от распределения серий в слове, являющемся центром этого шара.

Определение. Альтернирующее подслово b в слове a называется *максимальным*, если в a нет альтернирующего подслова c такого, чтобы слово b оказалось собственным подсловом слова c.

Пример. Слово $\gamma = \alpha_1\alpha_2\alpha_3\ldots\alpha_9 = 100110101$ имеет три максимальных альтернирующих подслова: $\alpha_1\alpha_2 = 10$, $\alpha_3\alpha_4 = 01$, $\alpha_5\alpha_6\alpha_7\alpha_8\alpha_9 = 10101$.

Число максимальных подслов длины i в слове γ обозначим через k_i.

Теорема 1[*]. *Пусть γ — центр шара радиуса единица, а слово γ состоит из m серий и k_i максимальных альтернирующих слов длины i. Тогда число точек в шаре радиуса единица с центром $\gamma \in E^n$ равно*

$$nm + 1 - \sum_{i=2}^{m} k_i \binom{i}{2}.$$

Следствие. *Максимальное число точек в шаре радиуса единица в метрике покрытий асимптотически равно n^2.*

Таким образом, справедливы неравенства

$$n+1 \leqslant |S(\gamma)| < n^2(1+o(1)),$$

где $|S(\gamma)|$ — объем единичного шара с центром в точке $\gamma \in E^n$.

Следует отметить, что нахождение объема шара радиуса t в метрике покрытий — нетривиальная задача, и это обстоятельство связано с упомянутой выше задачей о вычислении функции $t(a,k)$. Ниже мы приводим границы объема шара радиуса t, которые отнюдь не претендуют на универсальность и завершенность.

Пусть $v_t(n,m)$ обозначает объем шара радиуса t с центром в точке $a \in E^n$, состоящей из m серий.

Утверждение. *При любых n и m справедливы неравенства*

$$\sum_{i=0}^{t} \binom{n}{i} \leqslant v_t(n,m) < (n-1)^t \prod_{i=0}^{t-1}(m+i). \tag{5}$$

[*] Этот результат получен автором совместно с Х. Арочей (Куба).

Если m и t фиксированы, а $n \to \infty$, то из (5) *следует, что*

$$\frac{n^t}{t!} \lesssim v_t(n, m) \lesssim m^t n^t.$$

Некоторые неравенства для функции $v_t(n, m)$ можно извлечь из результатов работ [11, 14], а также из других источников. Однако нет полной картины асимптотического поведения величины $v_t(n, m)$ как функции трех параметров.

§ 2. Приложения к распознаванию образов и теории информации

Рассмотрим одну новую модель принятия решений в области теории информации и распознавания образов.

Пусть заданное подмножество M из E^n используется как источник сообщений. При этом число потенциальных получателей сообщений из M есть некоторое фиксированное число q. Предположим, что каждый получатель владеет своей частью информации о неизвестном сообщении $a \in M$. Пусть эта информация задается вектором $\{a_1, a_2, \ldots, a_q\}$. По информации $\{a_1, a_2, \ldots, a_q\}$ и данным о канале (способе получения информации a_i) требуется принять решение о принадлежности переданного сообщения фиксированному классу, т. е. декодировать сообщение. Ясно, что при $q = 1$ мы имеем классическую модель Шеннона [1].

В сформулированной выше модели «параметром» остается «индивидуальность» восприятия, которая обозначена как «своя» часть информации. Рассмотрим два примера, в которых понятию индивидуальная информация придается однозначный смысл.

1. Распознавание слуховых образов по фрагментам. Предположим, что в условиях плохой слышимости переданное в эфир сообщение принимается несколькими станциями. При этом помехи при принятии сообщения выражаются в том, что часть букв сообщения не воспринимается получателями. Таким образом, в данном случае каждый получатель имеет на выходе определенный фрагмент сообщения (слова). Индивидуальность восприятия состоит в том, что фрагменты, полученные разными получателями, образовались путем выпадения разных групп символов.

Пример. Пусть $q = 2$, $a = 000111$ и известно, что каждое сообщение содержит пять букв, т. е. получено из a выпадением одного символа. Тогда приведенное выше ограничение состоит в том, что номера выпавших символов должны быть разными для получателей, хотя сами фрагменты могут совпадать. Например, это произойдет, если первый получатель воспримет сообщение a', полученное из a выпадением первого символа,

а второй получатель — сообщение a'', полученное из a выпадением третьего символа.

Отметим, что рассмотренная модель отличается от хорошо известной в теории кодирования модели канала с выпадением и вставкой символов в части, связанной именно с числом получателей [11].

В качестве одного из самых простых ограничений на канал связи примем следующее: каждое полученное на выходе сообщение есть фрагмент длины $n - t$ исходного слова.

Определение. Множество слов $\mathscr{M} \subseteq E^n$ называется *q-декодируемым*, если любое слово $a \in \mathscr{M}$ однозначно восстанавливается по q фрагментам при принятых выше ограничениях.

Определение. Мультимножество всех фрагментов длины t слова $a \in E^n$ называется *t-окрестностью* слова a. Эта окрестность обозначается через $S_t(a)$.

Замечание. Под пересечением мультимножеств в дальнейшем понимается множество их общих элементов, причем каждый элемент a учитывается в пересечении с кратностью, равной минимальной кратности элемента a в пересекаемых мультимножествах.

Согласно введенным определениям и ограничениям на канал можно сформулировать следующее условие q-декодируемости множества $\mathscr{M}$.

Утверждение. *Множество $\mathscr{M} \subseteq E^n$ является q-декодируемым тогда и только тогда, когда в пересечении $(n - t)$-окрестностей любых элементов $a, b \in M$ содержится не более чем $q - 1$ элементов.*

Интересно сравнить это условие с обычным условием декодируемости для кодов с выпадением и вставкой символов [11].

Следующее утверждение показывает, что при достаточно большом q множество E^n может являться q-декодируемым множеством.

Пусть $t = 1$, т. е. информация получателя представляет собой слова длины $n - 1$.

Теорема 2 [19]. *Множество E^n является q-декодируемым множеством тогда и только тогда, когда $q > \lfloor n/2 \rfloor + 1$.*

Нетрудно привести пример, показывающий необходимость указанного в теореме числа получателей. Пусть $a = 0^{\lfloor n/2 \rfloor}10^{n-\lfloor n/2 \rfloor - 1}$ и $b = 0^{\lfloor n/2 \rfloor - 1}10^{n-\lfloor n/2 \rfloor}$. Тогда фрагменты, полученные удалением по одному нулю из первой серии в слове a и по одному нулю из последней серии в слове b, одинаковы.

Согласно теореме 2 в случае многих получателей возможны результаты принципиально иного плана, чем в случае классической теории кодирования, хотя, конечно, число приемников не вызывает оптимизма.

Однако приведенные ниже примеры, связанные с двоичным симметричным каналом связи, указывают на то, что и эта ситуация не является фатальной. Следуя обычным канонам, введем функцию $A(n, q, t)$, задающую объем максимального по мощности q-декодируемого множества с исправлением не более чем t выпадений символов из слов длины n. Из теоремы 2 следует, что $A(n, q, 1) = 2^n$ при $q > \lfloor n/2 \rfloor + 1$.

Отметим, что в данном варианте можно рассмотреть также и задачу в вероятностной постановке, когда каждый символ слова $\alpha_1\alpha_2 \ldots \alpha_n$ выпадает с вероятностью δ и сохраняется с вероятностью $1 - \delta$. В этом случае на выходе получается вектор $(\beta_1, \beta_2, \ldots, \beta_q)$ и требуется найти «ближайшее» к этому вектору слово из кодового словаря $\mathscr{M}$, что будет соответствовать процедуре декодирования по принципу максимального правдоподобия. В данном случае эта процедура может быть реализована, например, следующим образом.

Декодирование по максимуму правдоподобия.

1. Для множества слов $\{a_1, a_2, \ldots, a_q\}$ из A^n находятся все покрытия длины n, т. е. строится множество всех слов $\mathscr{M}^0 = \{b_1, b_2, \ldots, b_N\} \in E^n$ таких, что каждое b_i содержит в качестве фрагментов все слова $a_1, a_2, \ldots, a_q$.

2. Находится расстояние между множествами $\mathscr{M}$ и $\mathscr{M}^0$, и любое слово a, на котором «реализуется» это расстояние, объявляется искомым сообщением.

2. Кодирование с многими приемниками. Отметим, что задача, аналогичная рассмотренной выше, может быть сформулирована и в классической теории кодирования.

Рассмотрим в качестве источника сообщений некоторое кодовое множество $\mathscr{M} \subseteq E^n$ и предположим, что при передаче по каналу связи некоторые буквы могут меняться на противоположные ($0 \to 1$, $1 \to 0$). Допустим также, что сообщение источника на выходе принимается двумя приемниками и что в слове длины n происходит не более одного искажения. Задача стандартная: как велик может быть объем кодового множества при условии правильного восстановления исходного сообщения?

Ограничения. Сообщения, полученные разными приемниками, возникли из исходного сообщения искажением *разных* позиций.

Таким образом, окрестность кодового слова $a \in \mathscr{M}$ в этом случае состоит из множества пар (a', a''), где a', $a'' \in S_1(a)$ — шар единичного радиуса (в метрике Хэмминга) с центром в слове a.

Нетрудно видеть, что множество $\mathscr{M}$ является декодируемым тогда и только тогда, когда любые два шара радиуса единица с центрами в точках кодового множества пересекаются не более чем в одной точке.

Действительно, пусть по каналу связи было передано слово $a \in \mathcal{M}$. Тогда на выходе получена пара (a', a''), причем $a', a'' \in S_1(a)$. Далее, пара (a', a'') может принадлежать только *одному* шару с центром в точках $\mathcal{M}$, так как в противном случае было бы нарушено условие декодируемости (мощность пересечения была бы равна двум). Поэтому по паре (a', a'') определяются однозначно шар $S_1(a)$, а значит, и его центр.

Отметим далее, что в метрике Хэмминга шары радиуса единица либо не пересекаются, либо пересекаются в двух точках. Поэтому сформулированное выше условие декодируемости для случая одной ошибки и двух приемников переходит в обычное условие непересечения шаров радиуса единица с центрами в точках кодового множества.

Отметим, что если число приемников $q \geqslant 3$, то в случае одной ошибки множество E^n является декодируемым множеством.

Пусть, как обычно, функция $A(n, q, 1)$ обозначает объем максимального декодируемого множества для случая одной ошибки и q приемников. Тогда из приведенных выше соображений вытекают следующие соотношения для функции $A(n, q, 1)$:

$$A(n, 2, 1) = A(n, 1, 1) = A(n, 3), \quad A(n, q, 1) = 2^n \quad (q \geqslant 3).$$

В общем случае, когда имеется t ошибок и q приемников, условие декодируемости выглядит следующим образом.

Условие декодируемости. Множество $\mathcal{M}$ является декодируемым тогда и только тогда, когда в пересечении любых двух шаров радиуса t с центрами в кодовых точках содержится не более $q - 1$ точек.

Таким образом, в случае многих приемников классическое условие плотной упаковки заменяется менее обременительным условием «малости» пересечений. Соответствующая функция $A(n, q, t)$ обозначает, как и выше, объем максимального по мощности декодируемого множества.

Отметим следующий простой комбинаторный факт: при любом $n \geqslant 4$ любые два шара радиуса 2 с центрами в точках, находящихся на расстоянии $d = 4$, пересекаются в шести точках.

Рассмотрим теперь код $\mathcal{M}$ с расстоянием 4 в E^n (например, код Хэмминга с проверкой на четность). В силу отмеченного выше факта любые два шара радиуса 2 с центрами в точках множества $\mathcal{M}$ пересекаются не более чем в шести точках. Поэтому множество $\mathcal{M}$ является декодируемым и справедлива следующая нижняя оценка:

$$A(n, 7, 2) \geqslant \frac{2^n}{n}(1 + o(1)).$$

Для сравнения отметим, что в классическом случае одного приемника для функции $A(n, 1, 2)$ справедливо соотношение

$$A(n, 1, 2) \asymp 2^n / n^2.$$

В случае вероятностной постановки задачи для простоты можно рассмотреть двоичный симметричный канал. Тогда на выходе канала получается набор слов $\{a_1, a_2, \ldots, a_q\}$, каждое из которых возникло из одного и того же слова a путем замены некоторых букв на противоположные с заданной вероятностью δ. Декодирование по максимуму правдоподобия может проводиться обычным способом. Однако в случае «неоднозначности» ближайшего слова в варианте с многими приемниками остается много способов эту неоднозначность исключить.

Замечание. Приведенную нижнюю оценку нетрудно обобщить и получить общую оценку

$$A\left(n, \binom{2t+1}{t}, t\right) \geqslant \frac{2^n}{n^{t-1}},$$

которая по порядку в n раз превосходит мощность наилучшего кода, исправляющего t ошибок. Безусловно, это лишь самые первые результаты такого типа. Здесь, конечно, интересно как получение более точных оценок для мощности кода, так и установление связи этой задачи с другими задачами, например в случае модели коррекции ошибок с помощью переспроса. Важно рассмотреть также разные типы каналов связи. Отметим, что результаты при этом могут сильно различаться, что продемонстрировано в двух предыдущих случаях для канала с выпадениями букв и двоичного симметричного канала.

§ 3. Экономная запись множества слов и проблема реализуемости

В связи с многими проблемами компьютерной математики возникает задача экономной записи слов в памяти ЭВМ и быстрого извлечения их из памяти. Эта задача имеет много различных аспектов, и известен ряд комбинаторных конструкций, относящихся к ее решению. Упомянем такие широко известные конструкции, как код Грея [20], последовательности де Брейна [21] (в некоторых статьях их называют «кольцевыми кодами»), циклы в единичном n-мерном кубе [22], коды Хэмминга. Рассмотрим один из аспектов этой задачи.

Проблема реализуемости. Пусть $\mathcal{M} = \{a_1, a_2, \ldots, a_m\} \subseteq E^k$. Существуют ли n и слово $a \in E^n$ такие, что множество фрагментов длины k в слове a совпадает с множеством $\mathcal{M}$?

Любое слово a, являющееся решением сформулированной проблемы, называется *реализацией* множества $\mathcal{M}$. Слово a — как бы упаковка для множества $\mathcal{M}$, так как представляет собой экономную запись этого множества.

Примеры. 1. Пусть $m = 1$. Тогда множество $\mathscr{M}$ является реализуемым лишь в случаях $a_1 = 0^k$, $a_1 = 1^k$ и a_1 — любое слово при $k = n$.

2. При $m = 2$ реализуемы лишь следующие множества:

$$\mathscr{M}_1 = \{1^k, 1^{k-1}0\},\ \mathscr{M}_2 = \{0^k, 0^{k-1}1\},\ \mathscr{M}_3 = \{10^{k-1}, 0^k\},\ \mathscr{M}_4 = \{01^{k-1}, 1^k\}.$$

Это утверждение легко обосновать, поскольку любая реализация двухэлементного множества не может иметь более двух серий.

Естественно решать задачу реализуемости с наиболее плотной упаковкой, т. е. искать слово минимальной длины, являющееся реализацией данного множества.

В случае, когда фрагментами слова являются его подслова, образованные последовательными буквами слова, задача реализуемости исследовалась в несколько более общей постановке в работе [23], в которой ставился вопрос о наименьшем числе слов, в совокупности реализующих заданное множество. Эффективные решения соответствующих задач получены и для обобщений — задач покрытия графов путями [24].

В общем виде решение задачи реализуемости может быть представлено как нахождение $(0,1)$- решений некоторого нелинейного уравнения.

Пусть $M = \{a_1, a_2, \ldots, a_m\} \subseteq E^k$ и $a_r = \alpha_1^r \alpha_2^r \ldots \alpha_k^r$, $1 \leqslant t \leqslant m$.

Утверждение. *Слово a является реализацией множества M тогда и только тогда, когда a есть решение уравнения*

$$\sum_{r=1}^{m} \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} x_{i_1}^{\alpha_1^r} x_{i_2}^{\alpha_2^r} \ldots x_{i_k}^{\alpha_k^r} = \binom{n}{k}. \tag{6}$$

Пример. Пусть $m = 1$ и $M = \{a_1 = \alpha_1\alpha_2 \ldots \alpha_k\}$. Тогда уравнение (6) имеет следующий вид:

$$\sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} x_{i_1}^{\alpha_1} x_{i_2}^{\alpha_2} \ldots x_{i_k}^{\alpha_k} = \binom{n}{k}.$$

Так как число слагаемых в сумме равно $\binom{n}{k}$, то все слагаемые равны единице. Пользуясь этим фактом и определением функции x^σ, легко вывести, что $\alpha_1 = \alpha_2 = \cdots = \alpha_k$.

Пример. Пусть $M = \{(10), (01)\}$. Тогда уравнение (6) переходит в следующее:

$$\sum_{r=1}^{2} \sum_{1 \leqslant i_1 < i_2 \leqslant n} x_{i_1}^{\alpha_1^r} x_{i_2}^{\alpha_2^r} = \sum_{1 \leqslant i_1 < i_2 \leqslant n} x_{i_1}(1 - x_{i_2}) + \sum_{1 \leqslant i_1 < i_2 \leqslant n} (1 - x_{i_1}) x_{i_2} = \binom{n}{2}.$$

Пользуясь этим фактом и леммой 4 из [6] или непосредственно вычисляя, получаем

$$\sum_{1\leqslant i_1<i_2\leqslant n} x_{i_1} = \sum_{i=1}^{n}(n-i)x_i, \qquad \sum_{1\leqslant i_1<i_2\leqslant n} x_{i_2} = \sum_{i=1}^{n} i x_i.$$

После простых преобразований приходим к уравнению

$$2y^2 - 2y(n+1) + n(n-1) = 0,$$

где $y = \|x\| \leqslant n$, которое не имеет решений при $n \geqslant 4$. Отсюда легко следует, что множество M не является реализуемым. Этот иллюстративный пример свидетельствует о принципиальных возможностях использования алгебраического подхода.

Задача восстановления слов по их фрагментам. Эта задача в приводимой формулировке, по-видимому, впервые была поставлена в работе [12] . Ниже мы приводим общую постановку, результаты и их содержательную интерпретацию. По своему характеру рассматриваемая задача наиболее близка к широко известной в теории графов проблеме Улама [25].

Заметим, что близкой к рассматриваемой задаче является задача о восстановлении слова по подмножеству его подслов, исследованная в работах [26, 27], в которых найдены условия единственности восстановления в терминах преобразований слов и алгоритм полиномиальной сложности, восстанавливающий слово по множеству перекрывающихся подслов фиксированной длины. Отмечается, что задача исследуется в связи с вопросами восстановления строения молекул полимеров по их перекрывающимся фрагментам.

Вернемся к формулировке исходной задачи. Зафиксируем некоторое множество подпоследовательностей последовательности $1, 2, \ldots, n$ и будем рассматривать только такие фрагменты слова a, которые порождаются этим множеством подпоследовательностей. Каждую подпоследовательность удобно задавать характеристическим набором $(v_1, v_2, \ldots, v_n)$ таким, что $v_i = 1$, если i входит в подпоследовательность, и $v_i = 0$ в противном случае.

Пример. Пусть $n = 4$. Тогда подпоследовательности $(2, 4)$ и $(1, 3, 4)$ задаются двоичными наборами (0101) и (1011).

Заданное множество характеристических наборов будем обозначать через $V = \{v_1, v_2, \ldots, v_N\}$, а процедуру порождения фрагмента e набором v будем описывать в виде операции фрагментирования $\langle a, v\rangle = e$.

Пример. Пусть $a = 1010$, $v = 0011$. Тогда $\langle a, v\rangle = 10$.

Множество V будем называть *характеристическим* для множества фрагментов.

Обозначим через $V^0(x)$ множество фрагментов слова x, порожденных всевозможными наборами из V, а через $V^1(x)$ — соответствующее мультимножество, содержащее каждый фрагмент слова x столько раз, сколько этот фрагмент получается в результате фрагментирования слова x множеством V.

Определение. Слова $a, b \in E^n$ называются V^i-*эквивалентными*, если $V^i(a) = V^i(b)$, $i = 0, 1$.

V^0-эквивалентность была определена в [5], а V^1-эквивалентность — в [10]. V^i-эквивалентность слов a и b будем обозначать следующим образом: $a \overset{V^i}{\sim} b$.

Ясно, что V^0-эквивалентность влечет V^1-эквивалентность, но не наоборот.

Основная проблема в изучении V^i-эквивалентности состоит в описании классов эквивалентности и нахождении эффективных алгоритмов установления V^i-эквивалентности.

Для описания классов эквивалентности потребуется несколько понятий.

Рассмотрим уравнение

$$\langle x, v \rangle = e.$$

Множество решений этого уравнения представляет собой $(n - \|v\|)$-мерный подкуб в E^n.

Далее, для произвольной системы множеств $\{C_{ij}\}$, $1 \leqslant i, j \leqslant N$, определим некоторое новое множество, называемое *теоретико-множественным перманентом*, которое определяется следующим образом:

$$\operatorname{per} A = \operatorname{per} \|C_{ij}\| = \bigcup_{\{i_1, \ldots, i_N\}} (C_{1i_1} \cap C_{2i_2} \cap \cdots \cap C_{Ni_N}).$$

Теоретико-множественный перманент напоминает обычный перманент матрицы, в котором операции умножения и сложения заменены на операции пересечения и объединения.

Обозначим через V_a^0 класс V^0-эквивалентности, содержащий слово a.

Для характеристического множества $V = \{v_1, v_2, \ldots, v_N\}$ и слова $a \in E^n$ образуем множество фрагментов $V^0(a) = \{e_i\}$. Далее рассмотрим уравнение вида

$$\langle x, v_i \rangle = e_j$$

и множество решений этого уравнения обозначим через A_{ij}.

Теорема 3 [5]. *При любом a справедливо равенство*

$$V_a^0 = \operatorname{per} \|A_{ij}\|.$$

Процедура описания классов эквивалентностей при втором определении несколько более громоздка, хотя принципиально ничего не меняется.

1. По характеристическому множеству V находится множество $V^1(a)$. Пусть $V^1(a) = \{e_1, e_2, \ldots, e_m\}$.

2. Из множества $V^1(a)$ образуются все выборки (мультимножества) объема N, каждая из которых содержит множество $V^1(a)$. Пусть такими выборками являются $Q_1, Q_2, \ldots, Q_t$, где $t = \binom{N-1}{m-1}$.

3. Каждой выборке Q_i поставим в соответствие матрицу A_i, определяемую следующим образом. Пусть $Q_i = \{e_1^i, e_2^i, \ldots, e_N^i\}$. Рассмотрим уравнение

$$\langle x, v_r \rangle = e_s^i, \quad 1 \leqslant s, r \leqslant N.$$

Множество решений этого уравнения обозначим через A_{sr}^i и положим $A_i = \|A_{sr}^i\|$.

Теорема 4 [10]. *При любом a справедливо равенство*

$$V_a^1 = \bigcup_{i=1}^{t} \operatorname{per} A_i.$$

Теоремы 3 и 4 хотя и дают дескриптивное описание классов V^i-эквивалентности, но малопригодны для практического использования. Однако в случае $i = 0$ имеется другой, более удобный алгебраический способ описания классов эквивалентностей, к изложению которого мы и перейдем.

Пусть $V = E_n^2$, т. е. в качестве фрагментов слова $a = \alpha_1\alpha_2\ldots\alpha_n$ рассматриваются все пары (α_i, α_j). В этом случае справедливо следующее

Утверждение. *Слова $a = \alpha_1\alpha_2\ldots\alpha_n$ и $b = \beta_1\beta_2\ldots\beta_n$ являются E_n^2-эквивалентными тогда и только тогда, когда выполняются условия*

$$\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \beta_i, \quad \sum_{i=1}^{n} i\alpha_i = \sum_{i=1}^{n} i\beta_i. \tag{7}$$

Таким образом, в случае $V = E_n^2$ каждый класс V_a^0 описывается уравнениями

$$\sum_{i=1}^{n} x_i = p, \quad \sum_{i=1}^{n} ix_i = q,$$

где

$$p = \sum_{i=1}^{n} \alpha_i, \quad q = \sum_{i=1}^{n} i\alpha_i.$$

В общем случае алгебраическое решение проблемы восстановления слов по их фрагментам выглядит следующим образом.

Вернемся к определению характеристического множества как некоторой совокупности подпоследовательностей последовательности $1, 2, \ldots, n$. Это множество будем обозначать через V^*.

Пример. Пусть $V = E_n^2$. Тогда $V^* = \{(i,j) \mid i < j;\ 1 \leqslant i, j \leqslant n\}$.

Множество V^* разобьем на подмножества $V_1^*, V_2^*, \ldots, V_n^*$, включая в подмножество V_i^* все подпоследовательности из V^* длины i. Очевидно, имеет место разбиение $V^* = \bigcup\limits_{i=1}^{n} V_i^*$. Каждому подмножеству V_i^* поставим в соответствие матрицу A_i, строками которой являются фрагменты слова $x = x_1 x_2 \ldots x_n$, соответствующие характеристическому множеству V_i^*.

Пример. Пусть $V^* = \{(1,2,3),(2,3,4),(3,4),(4,5)\}$. Тогда $V_1^* = \varnothing$, $V_2^* = \{(3,4),(4,5)\}$, $V_3^* = \{(1,2,3),(2,3,4)\}$, $V_4^* = \varnothing$, а

$$A_2 = \left\| \begin{matrix} x_3 & x_4 \\ x_4 & x_5 \end{matrix} \right\|, \quad A_3 = \left\| \begin{matrix} x_1 & x_2 & x_3 \\ x_2 & x_3 & x_4 \end{matrix} \right\|.$$

На множестве матриц $A = \|r_{ij}\|$ размером $m \times n$ введем функцию $\Theta(A)$:

$$\Theta(A) = \sum_{j=1}^{m} r_{j_1} r_{j_2} \ldots r_{j_n}.$$

Через $A_q(j_1, j_2, \ldots, j_s)$ обозначим подматрицу матрицы A_q, состоящую из столбцов с номерами $j_1, j_2, \ldots, j_s$.

Пример. Пусть, как и выше, $A_3 = \left\| \begin{matrix} x_1 & x_2 & x_3 \\ x_2 & x_3 & x_4 \end{matrix} \right\|$. Тогда

$$A_3(1,2) = \left\| \begin{matrix} x_1 & x_2 \\ x_2 & x_3 \end{matrix} \right\|, \quad \Theta(A_3(1,2)) = x_1 x_2 + x_2 x_3.$$

Через $\lambda_{i_1 i_2 \ldots i_s}^{j_1 j_2 \ldots j_s}$ обозначим число строк в матрице $A_q(j_1, j_2, \ldots, j_s)$, совпадающих со строкой $(x_{i_1} \ldots x_{i_s})$.

Пример. Пусть $A = \left\| \begin{matrix} x_1 & x_2 & x_3 \\ x_3 & x_2 & x_4 \\ x_1 & x_2 & x_4 \end{matrix} \right\|$. Тогда $\lambda_{12}^{12} = 2$; $\lambda_{13}^{12} = 0$; $\lambda_{23}^{12} = 0$; $\lambda_{32}^{12} = 1$.

В ряде случаев параметры $\lambda_{i_1 i_2 \ldots i_s}^{j_1 j_2 \ldots j_s}$, играющие ключевую роль в проблеме V^0-эквивалентности, можно выписать в явном виде.

Лемма 6 [6]. *Если $V = E_n^k$, т. е. характеристическое множество представляет собой все подпоследовательности длины k последовательности $1, 2, \ldots, n$, то справедливы соотношения*

$$\lambda_{i_1 i_2 \ldots i_s}^{j_1 j_2 \ldots j_s} = \binom{i_1 - 1}{j_1 - 1} \binom{i_2 - i_1 - 1}{j_2 - j_1 - 1} \ldots \binom{n - i_s}{k - j_s},$$

где $1 \leqslant i_1 < i_2 < \cdots < i_s \leqslant n$.

Перейдем к формулировке алгебраического критерия эквивалентности.

1. Характеристическое множество V^* разобьем на подмножества V_i^* и получим представление $V^* = \bigcup\limits_{i=1}^{n} V_i^*$. Для каждого элемента разбиения V_q^* строим матрицу A_q. Пусть мощности непустых подмножеств в $\{V_i^*\}$ есть $q_1, q_2, \ldots, q_m$.

2. Для фиксированного q рассмотрим всевозможные подматрицы матрицы A_q вида $A_q(j_1, j_2, \ldots, j_s)$ $(s = 1, 2, \ldots, q)$ и для каждой такой матрицы строим определенный выше многочлен

$$\Theta[A_q(j_1, j_2, \ldots, j_s] = \sum_{1 \leqslant i_1 < i_2 < \cdots < i_s \leqslant q} \lambda_{i_1 i_2 \ldots i_s}^{j_1 j_2 \ldots j_s} x_{i_1} x_{i_2} \ldots x_{i_s}.$$

Теорема 5 [6]. *Слова* $a = \alpha_1 \alpha_2 \ldots \alpha_n$ *и* $b = \beta_1 \beta_2 \ldots \beta_n$ *являются* V^0*-эквивалентными тогда и только тогда, когда выполняются соотношения*

$$\sum_{1 \leqslant i_1 < i_2 < \cdots < i_s \leqslant q_i} \lambda_{i_1 i_2 \ldots i_s}^{j_1 j_2 \ldots j_s} \alpha_{i_1} \alpha_{i_2} \ldots \alpha_{i_s} = \sum_{1 \leqslant i_1 < i_2 < \cdots < i_s \leqslant q_i} \lambda_{i_1 i_2 \ldots i_s}^{j_1 j_2 \ldots j_s} \beta_{i_1} \beta_{i_2} \ldots \beta_{i_s},$$

$$1 \leqslant s \leqslant q_i, \quad 1 \leqslant i \leqslant m.$$

В этих равенствах s пробегает значения от 1 до q_i и для любых s и q_i выписывается $\binom{q_i}{s}$ уравнений. Таким образом, общее число уравнений в системе равно $\sum\limits_{i=1}^{m} 2^{q_i}$.

Пример. Пусть $n = 4$, $V = E_n^2$. Тогда

$$V^* = \{(1,2), (1,3), (1,4), (2,3), (3,4), (3,4)\}.$$

В данном случае длина всех фрагментов равна двум, $V^* = V_2^*$ и

$$A_2 = \begin{Vmatrix} x_1 & x_2 \\ x_1 & x_3 \\ x_1 & x_4 \\ x_2 & x_3 \\ x_2 & x_4 \\ x_3 & x_4 \end{Vmatrix}.$$

Матрица A_2 содержит две подматрицы $A_2(1)$ и $A_2(2)$, где

$$A_2(1) = \begin{Vmatrix} x_1 \\ x_1 \\ x_1 \\ x_2 \\ x_2 \\ x_3 \end{Vmatrix}, \qquad A_2(2) = \begin{Vmatrix} x_2 \\ x_3 \\ x_4 \\ x_3 \\ x_4 \\ x_4 \end{Vmatrix}.$$

Далее имеем

$$\Theta[A_2(1)] = 3x_1+2x_2+x_3,\ \Theta[A_2(2)] = x_2+2x_3+3x_4,\ \Theta[A_2] = \sigma_2(x_1x_2x_3x_4).$$

Пусть теперь $\alpha = 1001$, $\beta = 0110$. Тогда

$$\begin{aligned} \Theta_\alpha[A_2(1)] &= 3, & \Theta_\beta[A_2(1)] &= 3,\\ \Theta_\alpha[A_2(2)] &= 3, & \Theta_\beta[A_2(2)] &= 3,\\ \Theta_\alpha[A_2] &= 1, & \Theta_\beta[A_2] &= 1. \end{aligned}$$

Из последних равенств вытекает, что $\alpha \overset{V_0}{\sim} \beta$. Путем аналогичных вычислений или с использованием леммы 6 можно получить общие условия E_n^2-эквивалентности (7).

Полнота фрагментарных множеств. Особый интерес представляют множества V, классы V^i-эквивалентностей которых содержат по одному элементу. В этом случае каждое слово $a \in E^n$ однозначно восстанавливается по своим фрагментам.

Определение. Характеристическое множество V называется *i-полным*, если каждый класс V^i-эквивалентности содержит одно слово $(i = 0, 1)$.

Пример. В [12] показано, что множество $V = E_n^k$ является 0-полным при $k > n/2$. Это был, по-видимому, первый нетривиальный пример 0-полного множества. Это же множество является 1-полным при $k \geqslant \lfloor n/2 \rfloor + 1$ (см. [10]).

Из теоремы 5 непосредственно вытекает

Теорема 6. *Характеристическое множество V^* является 0-полным тогда и только тогда, когда система уравнений*

$$\sum_{1 \leqslant i_1 < i_2 < \cdots < i_s \leqslant q} \lambda^{j_1 j_2 \cdots j_s}_{i_1 i_2 \ldots i_s} x_{i_1} x_{i_2} \ldots x_{i_s} = \lambda(i_1, j_2, \ldots, j_s) \qquad (8)$$

имеет не более одного решения в $(0,1)$-числах при любой правой части.

Отметим, что при $V = E_n^k$ и $k > n/2$ рассмотрение только линейной части системы уравнений (8) позволяет построить простой алгоритм мажоритарного типа для восстановления слова по фрагментам длины k и тем самым конструктивно доказать 0-полноту E_n^k при $k > n/2$.

Представляется весьма правдоподобным, что имеющаяся верхняя оценка для минимального числа k, при котором множество E_n^k является 0-полным $(k < n/2 + 1)$, сильно завышена и истинное значение этой величины с точностью до порядка равно $\log_2 n$. Наилучшая к настоящему времени нижняя оценка имеет следующий вид.

Теорема 7. *Пусть $l(n)$ — минимальное значение k, при котором множество E_n^k является 0-полным. Тогда справедливо неравенство*

$$l(n) > c \log_2 n,$$

где $c = (\log_2(1+\sqrt{5}\,) - 1)^{-1}$.

Эта оценка получена в [28]. Более слабая оценка

$$l(n) \geqslant \log_2 n / \log_2 \log_2 n$$

содержится в [5].

В монографии [4] со ссылкой на работу [13] введено и изучено понятие эквивалентности без ограничений на длину слов и приведен ряд глубоких результатов, которые можно трактовать как решение проблемы эквивалентности. Там же получен результат, решающий проблему восстановления слов в случае, когда характеристическое множество представляет собой шар фиксированного радиуса. В наших терминах этот результат можно сформулировать следующим образом.

Теорема 8. *Множество E_n^k является 1-полным тогда и только тогда, когда выполняется неравенство $k \geqslant \lfloor n/2 \rfloor + 1$.*

Некоторые указанные выше результаты получены также в работах [8, 10].

Отметим, что свойство полноты не является монотонным по включению. Другими словами, существуют примеры полных характеристических множество, добавление к которым новых характеристических наборов приводит к потере свойства полноты. Это обстоятельство можно объяснить следующим образом. Добавляя «новый» характеристический набор в некоторое полное характеристическое множество, мы, с одной стороны, увеличиваем количество информации о неизвестном слове, а с другой — добавляем и неопределенность, так как теперь становится более неясным, от «действия» каких «элементов» получены фрагменты. От соотношения «информации» и «неопределенности», возникающей в этом случае, зависит полнота характеристического множества.

Зная некоторое множество фрагментов неизвестного слова, мы обладаем определенной информацией об этом слове. Как оценить количество или ценность этой информации?

Итак, исходными данными являются некоторое множество S слов, вообще говоря, разной длины в алфавите $\{0, 1\}$. Требуется оценить, насколько полно множество S характеризует неизвестное слово. Один из способов такой характеризации, основанной на рассмотренных выше понятиях, состоит в следующем.

1. Образуем «информационное пространство» $V_n(S)$, т. е. находим множество всех слов длины n, каждое из которых в качестве фрагментов содержит все слова из S.

2. Вводим меру неопределенности на множестве S:

$$\mu(S) = \log_2 |V_n(S)|/n.$$

Когда задано характеристическое множество $V = \{v_1, v_2, \ldots, v_N\}$, процесс построения меры таков. Пусть $S = \{a_1, a_2, \ldots, a_N\}$ — множество фрагментов и $V = \{v_1, v_2, \ldots, v_N\}$ — характеристическое множество. Согласно описанной выше процедуре выписываем уравнения

$$\langle x, v_i \rangle = a_j, \quad 1 \leqslant j \leqslant N.$$

Если A_{ij} — множество решений j-го уравнения, то информационное пространство для множества S представляет собой $\operatorname{per} \|A_{ij}\|$, и аналогично тому, как это сделано выше, определяем меру

$$\mu(S) = \log_2 |\operatorname{per} \|A_{ij}\| \,|/n.$$

С помощью введенной меры можно измерять также и «ценность» информации, заключенной в множестве S. Роль такого измерителя связана с функцией $\pi(S) = 1 - \mu(S)$. В частности, если множество S однозначно определяет слово, то содержащаяся в нем информация имеет «абсолютную» ценность и $\pi(S) = 1$.

Замечание. В настоящей работе мы упомянули лишь часть аспектов и результатов, связанных с данной проблемой. В частности, рассмотрели только двоичный алфавит, хотя в ряде случаев многие из упомянутых выше результатов непосредственно можно перенести на произвольный алфавит. Многие задачи и результаты, на наш взгляд, безусловно заслуживают бóльшего внимания. Это относится к проблеме экономного представления конечного множества слов, структурной проблеме полноты, алгоритмическим проблемам восстановления и т. д.

Автор выражает благодарность А. А. Евдокимову за замечания и дополнения к статье, способствовавшие более полному изложению результатов по рассматриваемой тематике, а также за указание литературы, в частности работ [4,13], которые были автору неизвестны.

ЛИТЕРАТУРА

1. **Шеннон К.** Работы по теории информации и кибернетике. М.: Изд-во иностр. лит., 1963.
2. **Журавлев Ю. И.** Об алгебраическом подходе к решению задач распознавания и классификации // Проблемы кибернетики. М.: Наука, 1978. Вып. 33. С. 5–68.
3. **Горелик А. П., Гуревич И. Б., Скрипкин В. А.** Современное состояние проблемы распознавания. М.: Радио и связь, 1985.

4. **Lothaire M.** Combinatorics on words // Encyclopedia of Mathematics and its Applications Reading. MIT: Addison-Wisley Publ. Co., 1983.
5. **Зенкин А. И., Леонтьев В. К.** Об одной неклассической задаче распознавания // Журн. вычисл. математики и мат. физики. 1984. Т. 24, № 6. С. 925–931.
6. **Леонтьев В. К., Сметанин Ю. Г.** О восстановлении векторов по набору их фрагментов // Докл. АН СССР. 1988. Т. 302, № 6. С. 1319–1322.
7. **Левенштейн В. И.** О совершенных кодах в метрике выпадений и вставок // Дискретная математика. 1991. Т. 3, вып. 1. С. 3–20.
8. **Manvel B., Meyerowitz A., Schwenk A., Smith K., Stockmeyr P.** Reconstruction of sequences // Discrete Math. 1991. V. 94, N 3. C. 209–219.
9. **Евдокимов А. А., Ню В.** Длина надпоследовательности для множества двоичных слов с заданным числом единиц // Методы дискретного анализа в теории графов и сложности: Сб. науч. тр. Новосибирск: Ин-т математики СО РАН, 1992. Вып. 52. С. 49–58.
10. **Леонтьев В. К.** Распознавание двоичных слов по их фрагментам // Докл. РАН. 1993. Т. 330, № 4. С. 434–436.
11. **Левенштейн В. И.** Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. АН СССР. 1965. Т. 163, № 4. С. 707–710.
12. **Калашник В. В.** Восстановление слова по его фрагментам // Вычисл. математика и вычисл. техника. Харьков, 1973. Вып. 4. С. 56–57.
13. **Simon I.** Piecewise testable events // Automata Theory and Formal Languages. Berlin etc.: Springer-Verl., 1975. P. 214–222. (Lecture Notes in Comput. Sci.; V. 33).
14. **Левенштейн В. И.** Элементы теории кодирования // Дискретная математика и математические вопросы кибернетики. М.: Наука, 1974. С. 207–305.
15. **Burosch G., Frankl U., Röhl S.** Über Ordnungen von Binäworten // Rostock Math. Kolloq. 1990. N 39. P. 53–64.
16. **Engel K., Gronan H.-D.** Sperner theory in partially ordered sets. Leipzig: BSB B. G. Teubner Verlagsgesellschaft, 1985.
17. **Гэри М., Джонсон Д.** Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982.
18. **Maier D.** The complexity of some problems on subsequences and supersequences // J. Assoc. Comput. Mach. 1978. V. 25, N 2. P. 322–336.
19. **Алиев Ш. М.** Алгоритмические вопросы построения оптимальных кодов для многих приемников: Дис. ... канд. физ.-мат. наук. М., 1993.
20. **Рейнгольд Э., Нивергельт Ю., Део Н.** Комбинаторные алгоритмы. Теория и практика. М.: Мир, 1980.
21. **Холл М.** Комбинаторика. М.: Мир, 1970.

22. Евдокимов А. А. О максимальной длине цепи в единичном n-мерном кубе // Мат. заметки. 1969. Т. 6, № 3. С. 309–319.

23. Ню В. Покрытие множества слов цепями: Дис. ... канд. физ.-мат. наук. Новосибирск, 1990.

24. Ню В., Евдокимов А. А. Покрытие графов маршрутами // Методы дискретного анализа в оптимизации управляющих систем: Сб. науч. тр. Новосибирск: Ин-т математики СО АН СССР, 1983. Вып. 40. С. 72–86.

25. Татт У. Теория графов. М.: Мир, 1988.

26. Сметанич Я. С. О восстановлении слов // Докл. АН СССР. 1971. Т. 201, № 4. С. 798–800.

27. Сметанич Я. С. О восстановлении слов // Тр. Мат. ин-та им. В. А. Стеклова. М.: Наука, 1973. Т. 83. С. 183–202.

27. Татт У. Теория графов. М.: Мир, 1988.

28. Сметанин Ю. Г. О некоторых задачах восстановления слов по фрагментам: Дис. ... канд. физ.-мат. наук. М., 1986.

Адрес автора:

Россия,
117967 Москва, ГСП–1,
ул. Вавилова, 40,
Вычислительный центр РАН

Статья поступила

20 мая 1994 г.