Теория информации

Теорема кодирования

Теория информации

Задача – представление **сообщений** из заданного дискретного множества **последовательностью символов**, принадлежащих заданному алфавиту.

Цель – конструирование последовательностей символов, **минимизирующих среднее число** символов, приходящихся на одно сообщение по ансамблю статистически **независимых** сообщений.

Теория информации

Нижняя граница для средней длины кодового слова

U – ансамбль из M сообщений $u_1, u_2, \ldots, u_k, \ldots, u_M$ с соответствующими вероятностями $P(u_k)$.

D – число символов в алфавите.

n_k – число символов в кодовом слове, соответствующем сообщению u_k.

Среднее число символов на одно сообщение:

$$\overline{n} = \sum_{k=1}^{M} P(u_k) n_k$$

Найдем нижнюю границу для $\overline{n}$.

Теория информации

Нижняя граница для средней длины кодового слова

$$I(X) \equiv M\left[I(x_k)\right] \equiv -\sum_X P(x)\log P(x) \equiv H(X)$$

$H(X) \le \log M$ при $P(u_k) = \dfrac{1}{M}$

$\log D$ – пропускная способность кодового алфавита

$H(Y|X) \le H(Y)$

$$\bar{n}\log D \ge H(U) \qquad \frac{H(U)}{\log D} \le \bar{n}$$

Теория информации

Общие правила конструирования кодовых слов со средней длиной, достаточно близкой к нижней границе по Шеннону

.В **каждой из позиций** кодового слова различные символы алфавита должны использоваться с **равными вероятностями**, с тем чтобы максимизировать среднее количество информации, доставляемое ими.

.**Вероятности** появления символов в каждой позиции кодового слова должны **не зависеть от всех предыдущих символов**.

Теория информации

Оптимальное множество кодовых слов для равновероятных сообщений

Сообщения	$P(u_k)$	1 разбиение	2 разбиение	3 разбиение	Кодовые слова
u_1	0,125	0	0	0	000
u_2	0,125			1	001
u_3	0,125		1	0	010
u_4	0,125			1	011
u_5	0,125	1	0	0	100
u_6	0,125			1	101
u_7	0,125		1	0	110
u_8	0,125			1	111

Теория информации

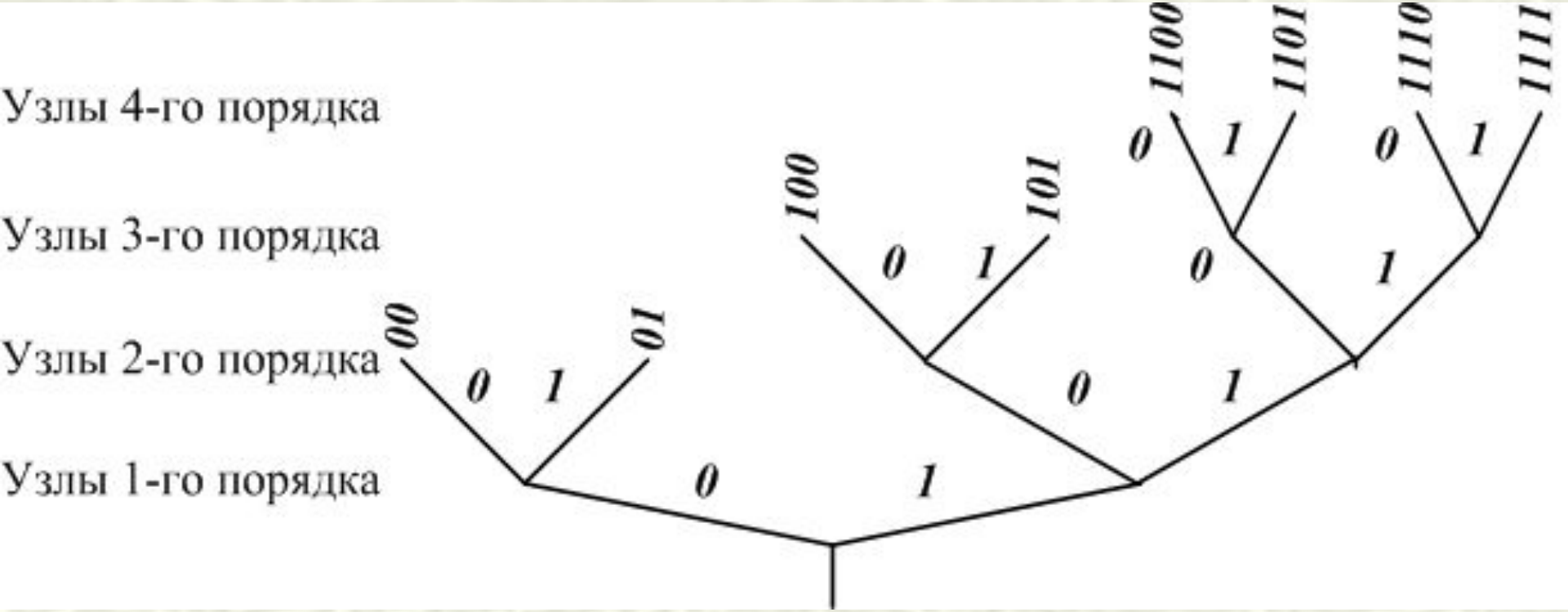
Оптимальное множество кодовых слов $n_k = I(u_k)$

2*2*0,25+2*3*0,125+4*4*0,0625=1+0,75+1=2,75

Сообщения	$P(u_k)$	1 разбиение	2 разбиение	3 разбиение	4 разбиение	Кодовые слова
u_1	0,25	0	0			00
u_2	0,25		1			01
u_3	0,125	1	0	0		100
u_4	0,125			1		101
u_5	0,0625		1	0	0	1100
u_6	0,0625				1	1101
u_7	0,0625			1	0	1110
u_8	0,0625				1	1111

Теория информации

Кодовое дерево для множества кодовых слов



Теория информации

Кодовое дерево для множества кодовых слов

Сообщения могут быть сопоставлены только **концевым** узлам, иначе алфавит становится **троичным** (0, 1, остановка). Кодовое слово - совокупность **указаний** для достижения узла, соответствующего сообщению. Ни одно из кодовых слов не совпало с **началом** (префиксом) какого-либо более длинного кодового слова. Иначе при непрерывной передаче сообщений невозможно однозначно разбить последовательность символов на **последовательные** сообщения.

Неравенство Крафта

Теорема. Неравенство $\sum_{k=1}^{M} D^{-n_k} \le 1$ является необходимым и достаточным условием существования кодовых слов, соответствующих концевым узлам дерева с длинами, равными n_k.

Доказательство. Покажем сначала, что это соотношение **необходимо**.

Из каждого узла дерева исходит не более D ветвей. Отсюда следует, что может быть не более D^n узлов порядка n.

Неравенство Крафта

Наличие концевого узла порядка n_k (не большего, чем n) **исключает** D^{n-n_k} возможных узлов порядка n

$$\sum_{k=1}^{M} D^{n-n_k} \le D^n$$

$$\sum_{k=1}^{M} D^{-n_k} \le 1$$

Необходимость условия теоремы доказана.

Неравенство Крафта

Для доказательства **достаточности** условия необходимо показать, что дерево с концевыми узлами, т. е. множество кодовых слов с заданными длинами, может быть фактически построено.

Пусть $n_k \le n_{k+1}$.

Предположим, что удалось построить дерево, содержащее все заданные концевые узлы порядка, меньшего m, и что дерево должно содержать ω_m концевых узлов порядка m. Согласно условию

$$\sum_{k=1}^{j} D^{-n_k} + \omega_m D^{-m} + \sum_{k=j+\omega_m+1}^{M} D^{-n_k} \le 1$$

Теория информации

Неравенство Крафта

$$\sum_{k=1}^{j} D^{-n_k} + \omega_m D^{-m} + \sum_{k=j+\omega_m+1}^{M} D^{-n_k} \le 1$$

$n_k < m$ $\qquad$ $n_k > m$

j – число концевых узлов порядка, меньшего чем m

$$D^m \sum_{k=1}^{j} D^{-n_k} + \omega_m D^{-m} D^m + D^m \sum_{k=j+\omega_m+1}^{M} D^{-n_k} \le D^m$$

$n_k < m$ $\qquad$ $n_k > m$

$$\omega_m \le D^m - \sum_{k=1}^{j} D^{m-n_k} - \sum_{k=j+\omega_m+1}^{M} D^{m-n_k}$$

Неравенство Крафта

$$\omega_m \le D^m - \sum_{k=1}^{j} D^{m-n_k} - \sum_{k=j+\omega_m+1}^{M} D^{m-n_k}$$

j – число концевых узлов порядка, меньшего чем m

Число доступных узлов порядка m: $D^m - \sum_{k=1}^{j} D^{m-n_k}$

Отсюда следует, что число доступных узлов порядка m не меньше заданного числа концевых узлов того же порядка, а потому все они могут быть включены в дерево. Что и требовалось доказать.

Теория информации

Дерево, содержащее M концевых узлов порядка n_1 , n_2 ,...,n_M может иметь или не иметь еще **добавочные концевые узлы**. Заданное множество концевых узлов называется **полным**, если существует дерево, имеющее эти концевые узлы и не имеющее никаких других концевых узлов (т.е. если соответствующее дерево не может быть дополнено никакими узлами), т. е., другими словами, если заданное множество узлов целиком заполняет дерево.

Теория информации

Теорема. Равенство $\sum_{k=1}^{M} D^{-n_k} = 1$

является необходимым и достаточным условием того, чтобы заданное множество концевых узлов было полным.

Теорема. Равенство $M = \nu(D-1)+1,$

где ν – целое положительное число, является необходимым и достаточным условием существования дерева с полным множеством из M концевых узлов.

Доказательство.

M_i – число имеющихся в дереве свободных узлов, когда дерево построено вплоть до порядка i

m_i – число промежуточных узлов порядка i, когда дерево построено до узлов порядка более высокого, чем i.

$$M_1 = D = (D-1)+1$$

$$M_2 = M_1 - m_1 + m_1 D = M_1 + m_1(D-1)$$

$$M_{i+1} = M_i + m_i(D-1)$$

$$M = 1 + (D-1)\left[1 + \sum_{i=1}^{n_M - 1} m_i\right]$$

Необходимость условия доказана.

Теория информации

Основная теорема кодирования

Теорема. При заданном ансамбле U из M сообщений с энтропией $H(U)$ и алфавитом, состоящим из D символов, возможно так закодировать сообщения ансамбля посредством последовательностей символов, принадлежащих заданному алфавиту, что среднее число символов на сообщение удовлетворяет неравенству

$$\frac{H(U)}{\log D} \le \bar{n} < \frac{H(U)}{\log D} + 1. \qquad (1)$$

Число $\bar{n}$ не может быть сделано меньше, чем нижняя граница в выражении (1).

Теория информации

Основная теорема кодирования

Вывод нижней границы

$$\bar{n} = \sum_{k=1}^{M} P(u_k) n_k$$

$$H(U) - \bar{n} \log D \le 0$$

Пусть $Q_k = D^{-n_k}$ $\quad n_k \log D = -\log Q_k$

$$H(U) - \bar{n} \log D = -\sum_{k=1}^{M} P(u_k) \log P(u_k) - \sum_{k=1}^{M} P(u_k) n_k \log D =$$

$$= -\sum_{k=1}^{M} P(u_k) \log P(u_k) + \sum_{k=1}^{M} P(u_k) \log Q_k =$$

$$= \sum_{k=1}^{M} \left[P(u_k) \log \frac{Q_k}{P(u_k)} \right] \le \sum_{k=1}^{M} P(u_k) \left[\frac{Q_k}{P(u_k)} - 1 \right] \log e =$$

$$= \left[\sum_{k=1}^{M} D^{-n_k} - 1 \right] \log e \le [1 - 1] \log e = 0$$

Теория информации

Вывод верхней границы.

Лемма. Для существования множества кодовых слов со средней длиной

$$\overline{n} = \frac{H(U)}{\log D}$$

необходимо и достаточно, чтобы для каждого сообщения выполнялось условие

$$\frac{I(u_k)}{\log D} \equiv \frac{-\log P(u_k)}{\log D} = a$$

, где a - целое число.

Когда это условие выполнено,

$$n_k = \frac{I(u_k)}{\log D}$$

Теория информации

Вывод верхней границы.

$$\frac{I(u_k)}{\log D} \le n_k^* < \frac{I(u_k)}{\log D} + 1$$

$$\sum_{k=1}^{M} P(u_k)\frac{I(u_k)}{\log D} \le \sum_{k=1}^{M} P(u_k) n_k^* < \sum_{k=1}^{M} P(u_k)\left[\frac{I(u_k)}{\log D} + 1\right]$$

$$\frac{H(U)}{\log D} \le \bar{n} < \frac{H(U)}{\log D} + 1$$

Теория информации

Источники статистически независимых сообщений

Теорема. При любом заданном как угодно малом положительном числе ε можно найти натуральное число ν и соответствующее множество $M\nu$ кодовых слов, такое, что среднее число символов на сообщение удовлетворяет неравенству

$$\overline{n} \le \frac{H(U)}{\log D} + \varepsilon .$$

Обратно, невозможно найти натуральное число ν и соответствующее множество кодовых слов, такое, что

$$\overline{n} < \frac{H(U)}{\log D}$$

Источники статистически независимых сообщений

Доказательство.

$$\frac{\nu H(U)}{\log D} \le \bar{n}_\nu < \frac{\nu H(U)}{\log D} + 1; \quad \frac{H(U)}{\log D} \le \bar{n} < \frac{H(U)}{\log D} + \frac{1}{\nu}; \quad \nu \ge \frac{1}{\varepsilon}$$

Когда каждое сообщение статистически **независимо** от всех предыдущих сообщений, то кодирование последовательности сообщений вместо отдельных сообщений может уменьшить среднее число символов на сообщение **не более** чем на **один** символ.

Теория информации

Метод оптимального кодирования Хафмана

Этот метод всегда приводит к получению **оптимального** множества кодовых слов в том смысле, что **никакое другое множество не имеет меньшего среднего числа символов на сообщение**.

1-й шаг. M сообщений располагаются в порядке убывания вероятностей.

2-й шаг. Пусть m_o – целое число, удовлетворяющее двум требованиям

$2 \le m_o \le D$, $\frac{M - m_o}{D - 1} = a$, где a - целое положительное число.

Теория информации

Метод оптимального кодирования Хафмана

Группируем вместе m_o сообщений, имеющих наименьшие вероятности, и вычисляем общую вероятность такого подмножества сообщений.

3-й шаг. Из первоначального ансамбля образуем вспомогательный ансамбль сообщений, рассматривая подмножество из сообщений, образованных на 2-м шаге, как отдельное сообщение с вероятностью, равной вероятности всего подмножества. Вновь располагаем сообщения этого вспомогательного ансамбля в порядке убывания вероятностей.

Теория информации

Метод оптимального кодирования Хафмана

4-й шаг. Образуем подмножество из D сообщений вспомогательного ансамбля, имеющих наименьшие вероятности, и вычисляем их общую вероятность.

5-й шаг. Из первого вспомогательного ансамбля образуем второй вспомогательный ансамбль, рассматривая подмножество из сообщений, образованное на четвертом шаге, как отдельное сообщение с вероятностью, равной общей вероятности всего подмножества. Располагаем сообщения этого второго вспомогательного ансамбля в порядке убывания вероятностей.

Теория информации

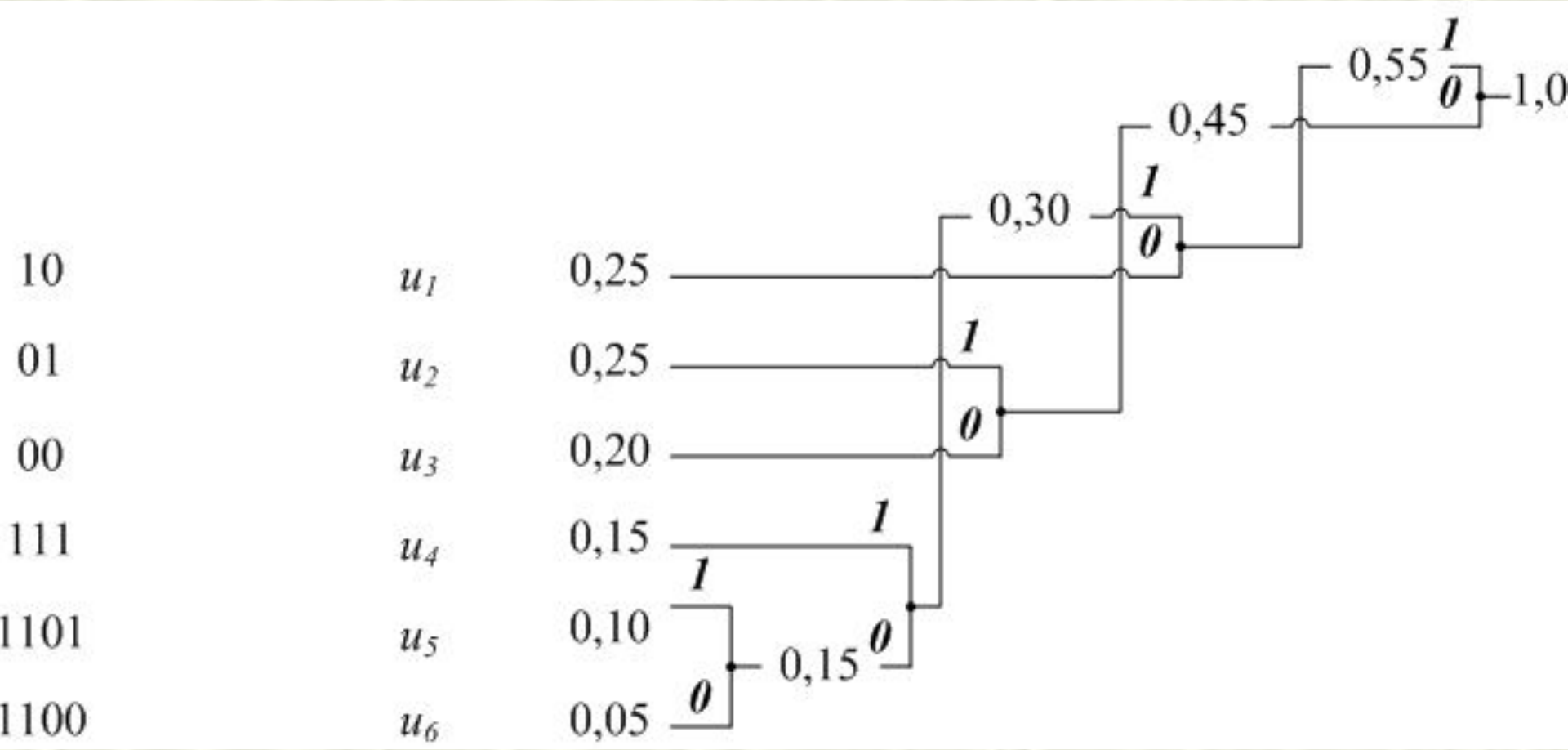
Метод оптимального кодирования Хафмана

6-й шаг. Образуем последовательные вспомогательные ансамбли путем повторения 4-го и 5-го шагов, пока в ансамбле не останется единственное сообщение с вероятностью единица.

7-й шаг. Проводя линии, соединяющие сообщения, образующие последовательные подмножества, получаем дерево, в котором отдельные сообщения являются концевыми узлами. Соответствующие им кодовые слова можно построить, приписывая различные символы из заданного алфавита ветвям, исходящим из каждого промежуточного узла. Только один из промежуточных узлов может иметь меньше чем D ветвей, а именно узел, образованный на 2-м шаге.

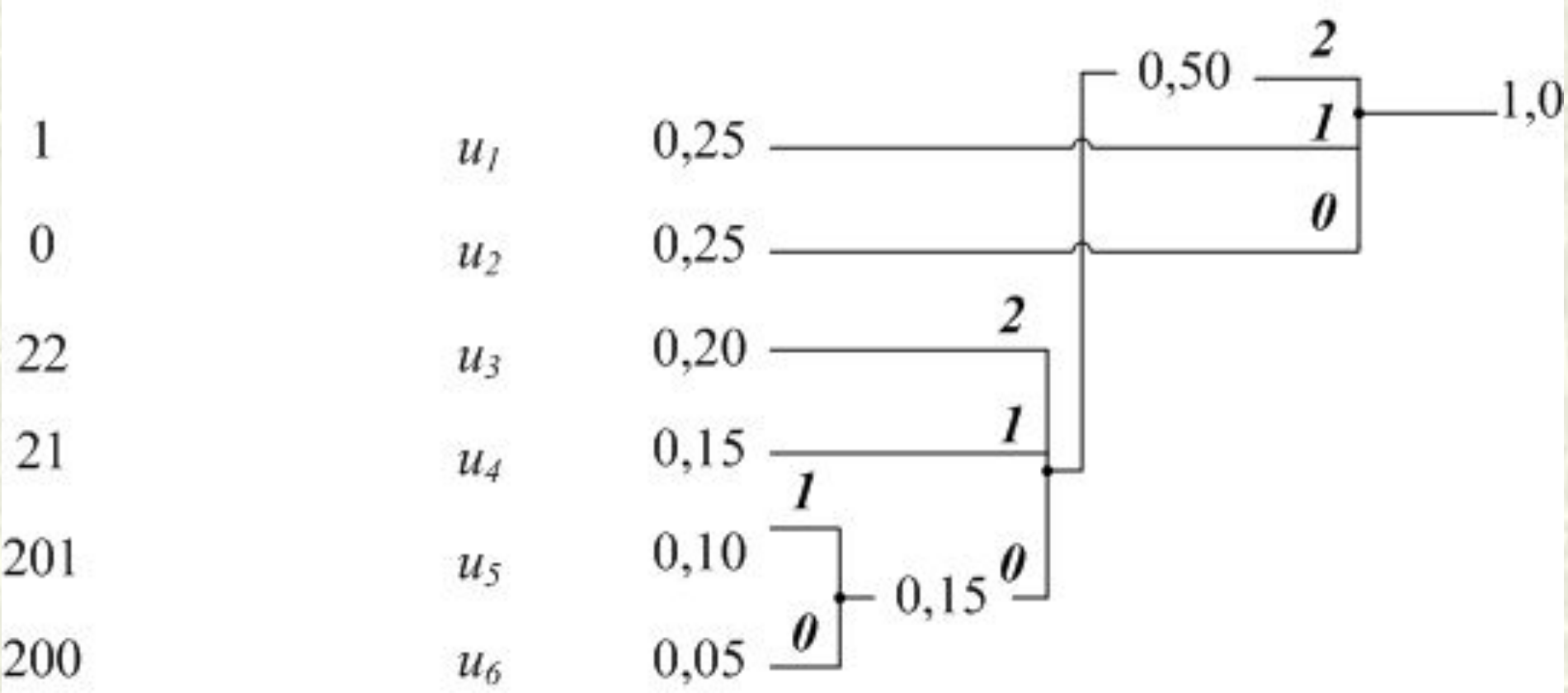
Теория информации

Оптимальное множество двоичных кодовых слов



Теория информации

Оптимальное множество троичных кодовых слов



Теория информации

Метод оптимального кодирования Хафмана

Условие 1. Сообщениям меньшей вероятности должны быть сопоставлены слова большей длины.

Условие 2. Два наименее вероятных сообщения должны соответствовать кодовым словам одной и той же длины n_M (максимальная длина), т. е. узлам одного и того же порядка.

Условие 3. Число $\omega_i(n_M)$ концевых узлов порядка n_M, сопоставленных сообщениям, и число $\omega_i(n_M-1)$ промежуточных узлов порядка n_M-1 должны удовлетворять неравенству

$$D\omega_i(n_M-1)-\omega_i(n_M)<D-1$$

Теория информации

Метод оптимального кодирования Хафмана

Условие 4. Из каждого промежуточного узла порядка меньшего, чем $n_M - 1$, должно исходить D ветвей.

Условие 5. Предположим, что некоторый промежуточный узел преобразован в концевой, т. е. исключены все порождаемые им ветви и узлы. Сопоставим этому узлу составное сообщение, имеющее вероятность, равную сумме вероятностей сообщений, сопоставленных исключенным концевым узлам. Тогда, если первоначальное дерево было оптимальным для исходного ансамбля сообщений, новое дерево должно быть оптимальным для нового ансамбля сообщений.

Теория информации

Метод оптимального кодирования Хафмана

Вспомогательное условие 6. Каждый промежуточный узел порядка, меньшего n_M, должен порождать D ветвей.

Теория информации

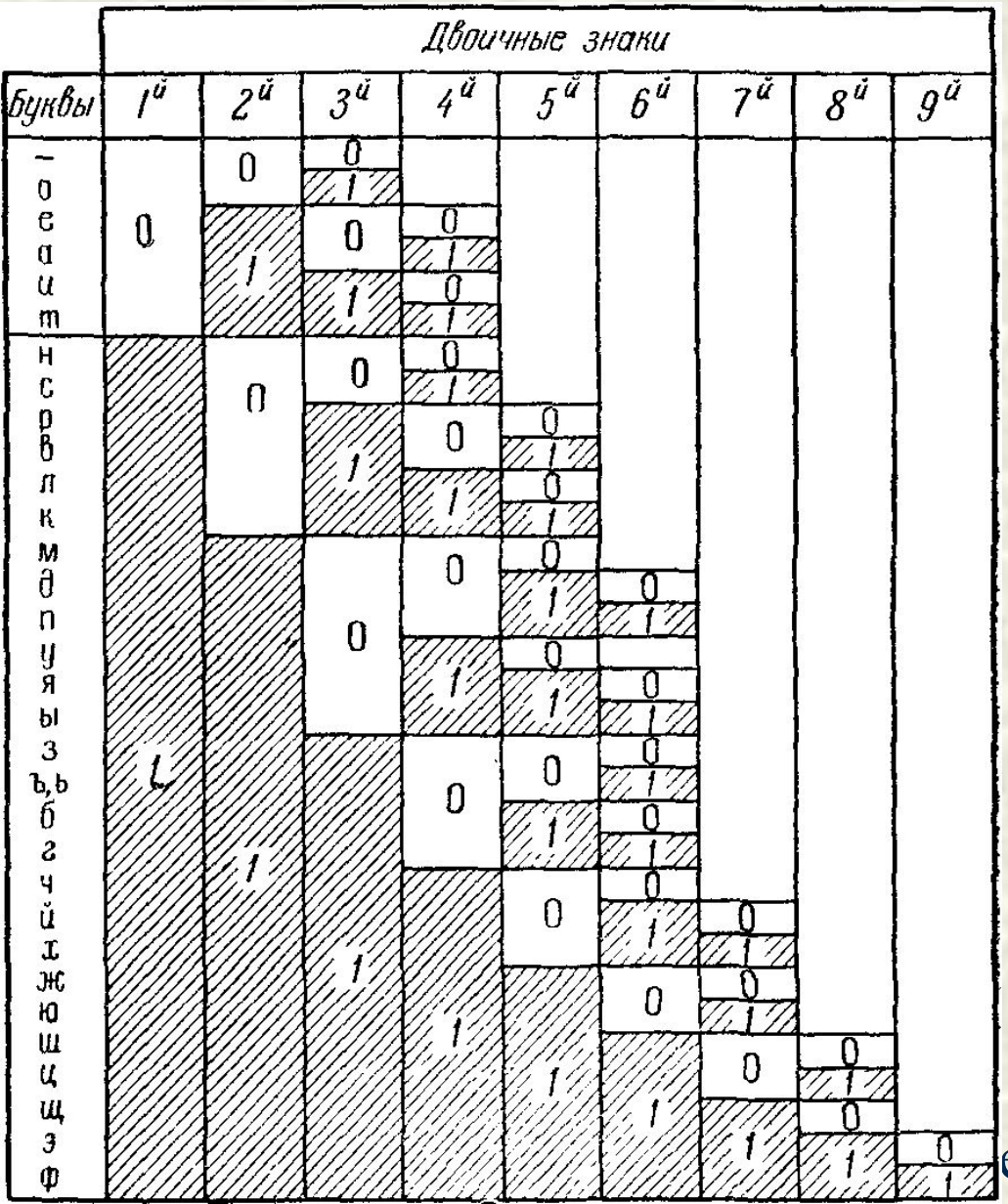
Построение оптимального кода для русского алфавита

При кодировании двоичных номеров букв:

$$\overline{n} = 5; \quad I_{1c} = 0{,}884 \text{бит}$$

Буква	Частота	Буква	Частота	Буква	Частота	Буква	Частота
«—»	0,145	р	0,041	я	0,019	х	0,009
о	0,095	в	0,039	ы	0,016	ж	0,008
е	0,074	л	0,036	з	0,015	ю	0,007
а	0,064	к	0,029	ъ, ь	0,015	ш	0,006
и	0,064	м	0,026	б	0,015	ц	0,004
т	0,056	д	0,026	г	0,014	щ	0,003
н	0,056	п	0,024	ч	0,013	э	0,003
с	0,047	у	0,021	й	0,010	ф	0,002

Теория информации



Построение оптимального кода для русского алфавита

Теория информации

Построение оптимального кода для русского алфавита ($\overline{n} \approx 4{,}42; \quad I_{1c} = 0{,}994$бит)

Буква	Двоичное число	Буква	Двоичное число	Буква	Двоичное число
«—»	000	к	10111	ч	111100
о	001	м	11000	й	1111010
е	0100	д	110010	х	1111011
а	0101	п	110011	ж	1111100
и	0110	у	110100	ю	1111101
т	0111	я	110110	ш	11111100
н	1000	ы	110111	ц	11111101
с	1001	з	111000	щ	11111110
р	10100	ъ, ь	111001	э	111111110
в	10101	б	111010	ф	111111111
л	10110	г	111011		