Лекция 4. ОСНОВЫ ЭКОНОМНОГО КОДИРОВАНИЯ

4.1 Принципы экономного кодирования. Цель сжатия данных и типы систем сжатия

4.2. Префиксные коды

4.3. Код Хаффмена

4.4. Код Шеннона-Фано

4.1 Принципы экономного кодирования. Цель сжатия данных и типы систем сжатия

Целью сжатия данных является компактное представление данных для более экономного их сохранения или повышения скорости передачи информации. Такое кодирование называют **экономным, безызбыточным**, или **эффективным кодированием**, а также **сжатием данных**.

Таблица 4.1 Способ представления кодов

Буква λ_i	Число λ_i	Код с основанием 10	Код с основанием 4	Код с основанием 2
А	0	0	00	000
Б	1	1	01	001
В	2	2	02	010
Г	3	3	03	011
Д	4	4	10	100
Е	5	5	11	101
Ж	6	6	12	110
З	7	7	13	111

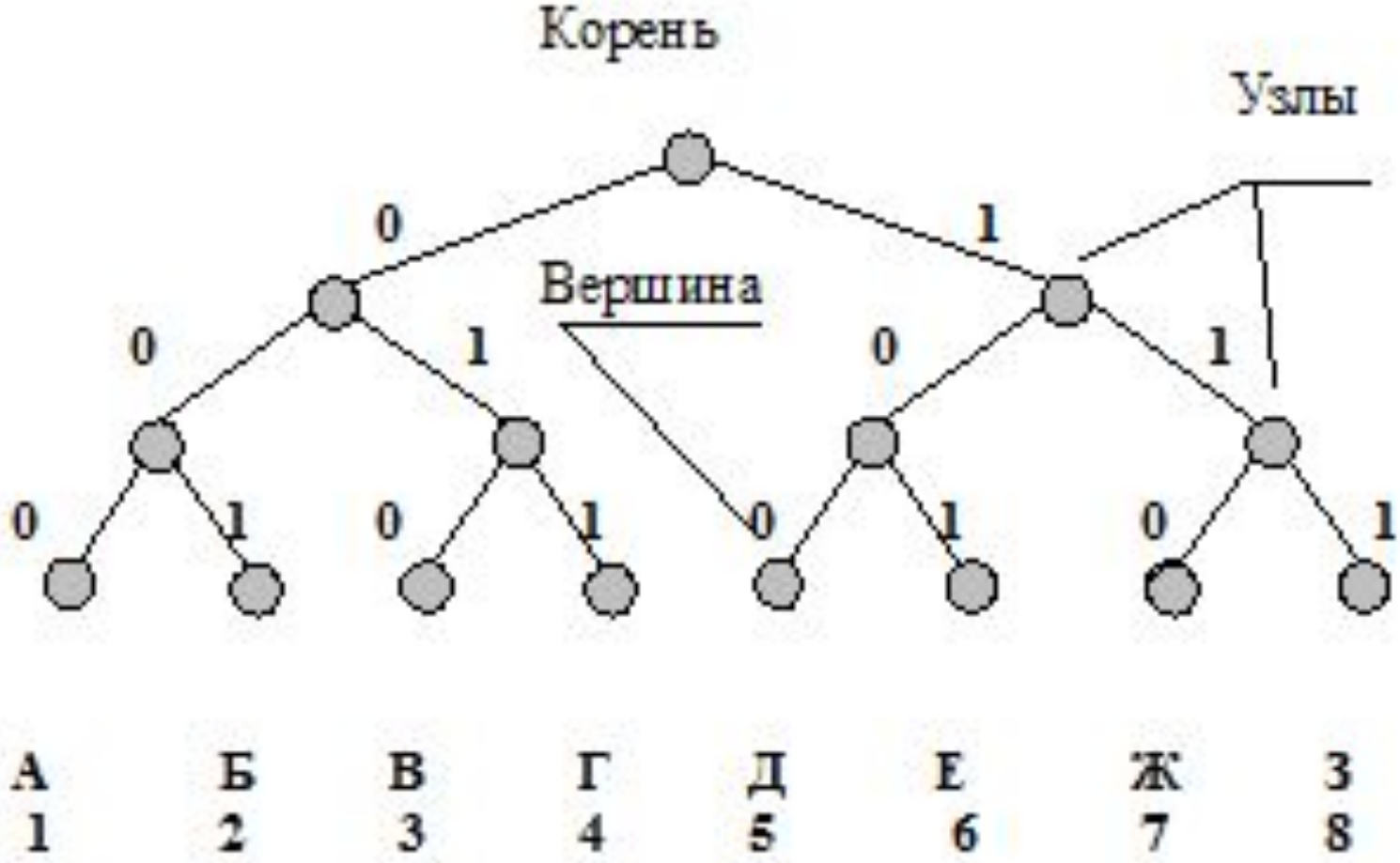

Рис. 4.1. Кодовое дерево для равномерного кодирования

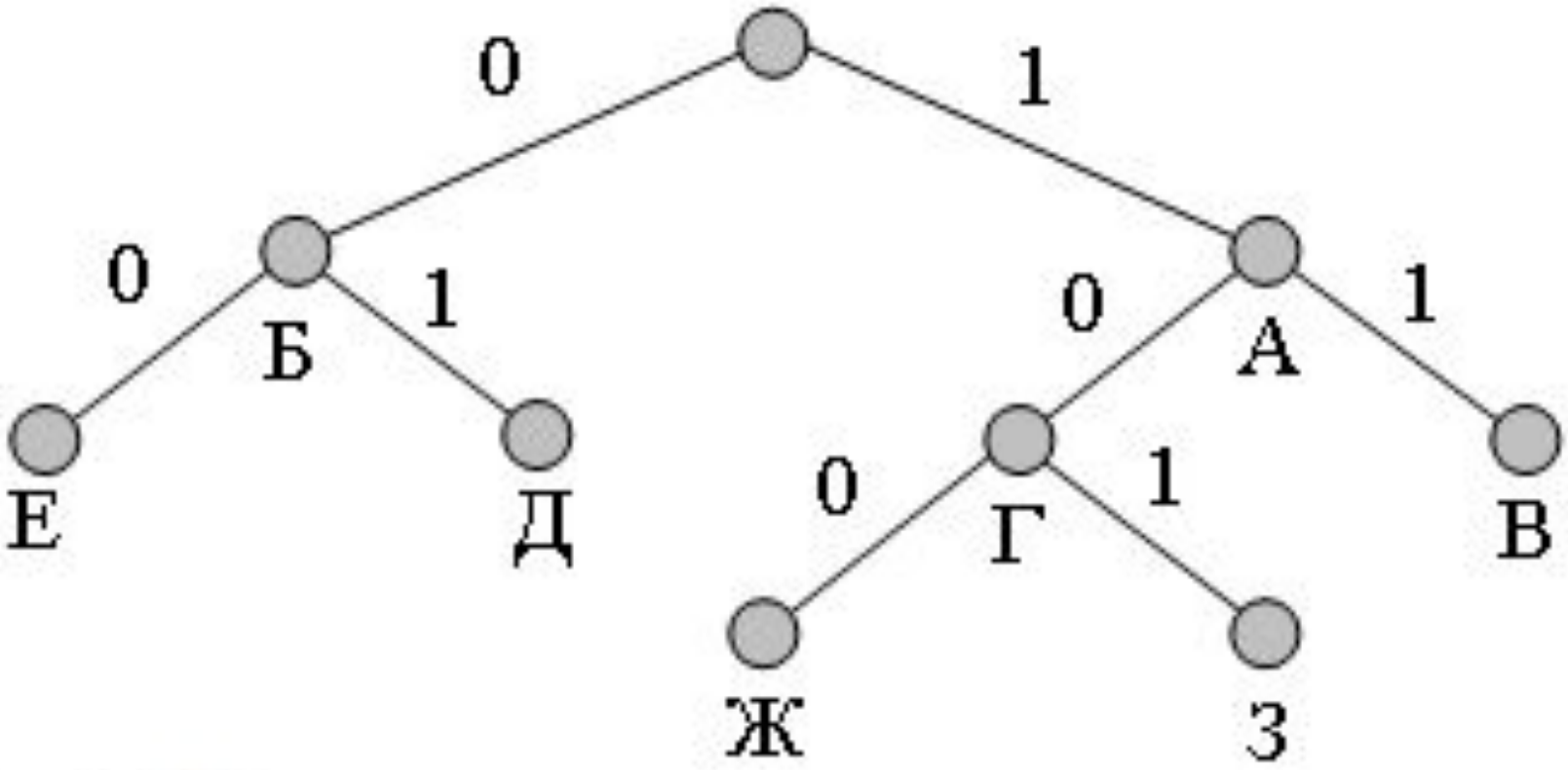

Рис. 4.2. Кодовое дерево для неравномерного кодирования
(**приводимый,** или **непрефиксный** код)

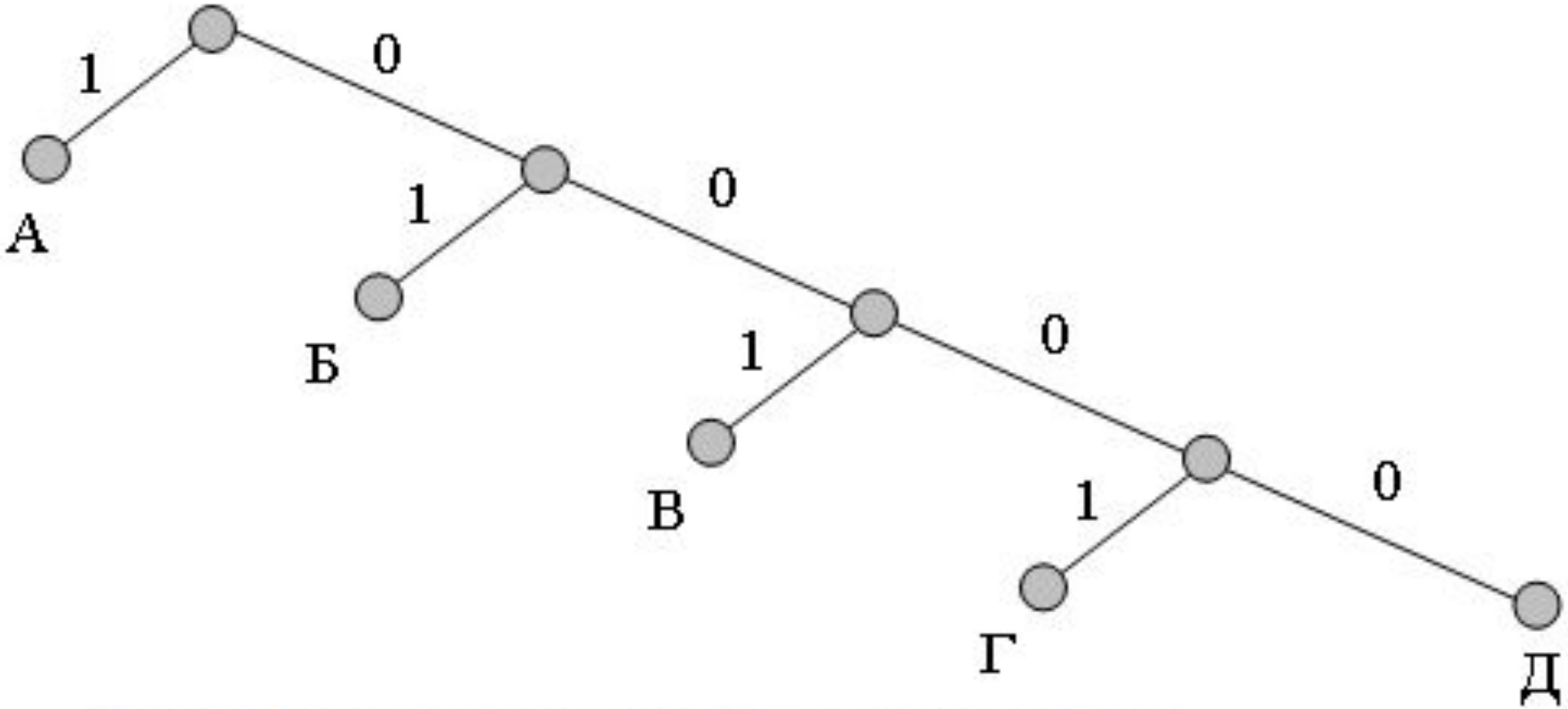

Рис. 4.3. Кодовое дерево для префиксного кодирования

Пример кодирования сообщений λ_i из алфавита объемом $N_\lambda = 8$ с помощью равномерного двоичного кода.

1) Источник сообщения выдает текст с алфавитом от *A* до *3* и вероятностью букв $p(\lambda_i) = 1/8$.

Основные информационные характеристики:

- энтропия источника

$$H(\lambda) = -\sum_{i=1}^{N_\lambda} p_i \log_2 p_i = -8 \cdot \frac{1}{8} \log_2 \frac{1}{8} = 3;$$

- максимальная энтропия $H_{\max}(\lambda) = \log_2 N_\lambda = 3;$

- избыточность источника $r_u = 1 - \frac{H(\lambda)}{H_{\max}(\lambda)} = 0;$

- число бит в коде $m = 3;$

- избыточность кода $r_k = 1 - \frac{H(\lambda)}{m} = 0.$

2) Источник сообщения выдает текст с алфавитом от *А* до *З* и вероятностью букв

Таблица 4.2. Вероятности появления в тексте букв

А	Б	В	Г	Д	Е	Ж	З
Ра=0.6	*Рб*=0.2	*Рв*=0.1	*Рг*=0.04	*Рд*=0.025	*Ре*=0.015	*Рж*=0.01	*Рз*=0.01

Энтропия источника $H(\lambda)=-\sum_{i=1}^{N_\lambda} p_i \log p_i$, $H(\lambda)=1,781.$

Число бит на одну букву при использовании равномерного трехразрядного кода

$$\overline{m}=3.$$

Избыточность кода

$$r_k=1-\frac{H(\lambda)}{\overline{m}}=1-\frac{1.781}{3}\approx 0.41.$$

В связи с тем, что при кодировании неравновероятных сообщений равномерные коды обладают большой избыточностью, было предложено использовать неравномерные коды, длительность кодовых комбинаций которых была бы согласована с вероятностью выпадения различных букв.

Такое кодирование называется **статистическим**.

Неравномерный код при статистическом кодировании выбирают так, чтобы **более вероятные буквы передавались с помощью более коротких комбинаций кода, менее вероятные - с помощью более длинных**. В результате уменьшается средняя длина кодовой группы в сравнении со случаем равномерного кодирования.

Существуют два типа систем сжатия данных:

- системы сжатия без потерь информации (неразрушающее сжатие);
- системы сжатия с потерями информации (разрушающее сжатие).

Характеристики систем сжатия без потерь информации

- **Коэффициент сжатия**:

$$\rho = \frac{n \log_2 N}{k},$$

где n - длина последовательности данных из алфавита (число символов в сжимаемом тексте), N - число состояний источника (количество символов в алфавите), k - размер сжатых данных в битах.

Скорость сжатия:

$$R = \frac{k}{n}.$$

Характеристики систем сжатия с потерями информации

Скорость сжатия

$$R = \frac{k}{n}.$$

Величина искажений

$$D = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^* - x_i\right)^2,$$

где X^* - сжатые данные и X - исходные.

4.2. Префиксные коды

Простейшими кодами, на основе которых может выполняться сжатие данных, являются коды без памяти. В коде без памяти каждый символ заменяется кодовой комбинацией из префиксного множества двоичных последовательностей.

Префиксным множеством двоичных последовательностей S называется конечное множество двоичных последовательностей, таких, что ни одна последовательность в этом множестве не является префиксом, или началом, никакой другой последовательности в S.

Множество $S1$ = {00, 01, 100, 110, 1010, 1011 } является префиксным множеством.

Множество $S2$ = { 00, 001, 1110 } не является префиксным множеством.

Если $S = \{w_1, w_2, \ldots, w_N\}$ - префиксное множество, то можно определить некоторый вектор $\nu(S) = (m_1, m_2, \ldots, m_N)$, состоящий из чисел, являющихся длинами соответствующих префиксных последовательностей (m_i - длина w_i).

Вектор $\nu(S) = (m_1, m_2, \ldots, m_N)$, состоящий из неуменьшающихся положительных целых чисел, называется **вектором Крафта**. Для него выполняется **неравенство Крафта**

$$2^{-m_1} + 2^{-m_2} + \ldots + 2^{-m_N} = \sum_{i=1}^{N} 2^{-m_i} \leq 1.$$

Примеры простейших префиксных множеств и соответствующие им векторы Крафта:

$S1 = \{0, 10, 11\}$ и $v(S1) = (1, 2, 2)$;

$S2 = \{0, 10, 110, 111\}$ и $v(S2) = (1, 2, 3, 3)$;

$S3 = \{0, 10, 110, 1110, 1111\}$ и $v(S3) = (1, 2, 3, 4, 4)$.

Средняя длина двоичной кодовой последовательности на выходе кодера составит

$$\overline{m} = m_1 p_1 + m_2 p_2 + \ldots + m_k p_N = \sum_{i=1}^{N} m_i p_i .$$

Прямая теорема неравномерного кодирования

Для ансамбля $X = \{x, P(x)\}$ с энтропией $H(X)$ существует побуквенный неравномерный префиксный код со средней длиной кодовых слов

$\overline{m} \le H(X) + 1.$

Обратная теорема неравномерного кодирования

Для любого однозначно декодируемого кода дискретного источника $X = \{x, P(x)\}$ с энтропией $H(X)$ средняя длина кодовых слов $\overline{m}$ соответствует неравенству

$\overline{m} \ge H(X).$

Вопрос: при каких условиях возможно равенство?

$$\sum_{i=1}^{N} p(x_i) m_i = -\sum_{i=1}^{N} p(x_i) \log p(x_i).$$

Для этого при каждом x должно выполняться соотношение

$$p(x_i) = 2^{-m_i}.$$

Или $m_i = -\log p(x_i)$.

Для этого кода неравенство Крафта преобразуется в равенство

$$\sum_{i=1}^{N} 2^{-m_i} = 1$$

4.3. Код Хаффмена

Алгоритм Хаффмена:

1. Все символы алфавита выписываются в порядке не возрастания вероятности их появления в тексте.

2. Последовательно объединяются 2 символа с наименьшими вероятностями в один составной символ. Каждому символу из составного приписывается: одному «0», а второму «1». Вероятность составного символа считается равной сумме вероятностей символов, входящих в составной.

3. Процесс продолжается до тех пор, пока все символы не будут объеденены в 1 символ, т.е. построено дерево.

4. Прослеживается путь от корня дерева к вершинам и записываются при этом всречающиеся «0» и «1», т.е. составляются кодовые комбинации.

4.4. Код Шеннона-Фано

Алгоритм Шеннона-Фано:

1. Символы алфавита записываются в порядке не возрастающих вероятностей.
2. Затем они разделяются на две части так, чтобы суммы вероятностей символов, входящих в каждую из таких частей, были примерно одинаковыми. Всем символам первой части приписывается ноль, а символам второй части — единица.

3. Затем каждая из этих частей (если она содержит более одного символа) делится в свою очередь на две, по возможности равновероятные части и к ним применяется то же самое правило кодирования.
4. Этот процесс повторяется до тех пор, пока в каждой из полученных частей не останется по одному символу.