ЛЕКЦИЯ № 12.

4. Основы теории информации.

Теория информации - математическая дисциплина. Предмет изучения – характеристики и передача информации. В теории информации (ТИ) рассматриваются понятия: объем данных, скорость передачи, пропускная способность канала, источник информации, энтропия источника, эффективное и помехоустойчивое кодирование.

ТИ, созданная математиком Клодом Элвудом Шенноном в 1948 г, первоначально применялась в области связи. Сейчас она применяется и в других областях, например, в вычислительной технике. На рисунке 4.1 показана упрощенная структурная схема системы передачи и приема информации.

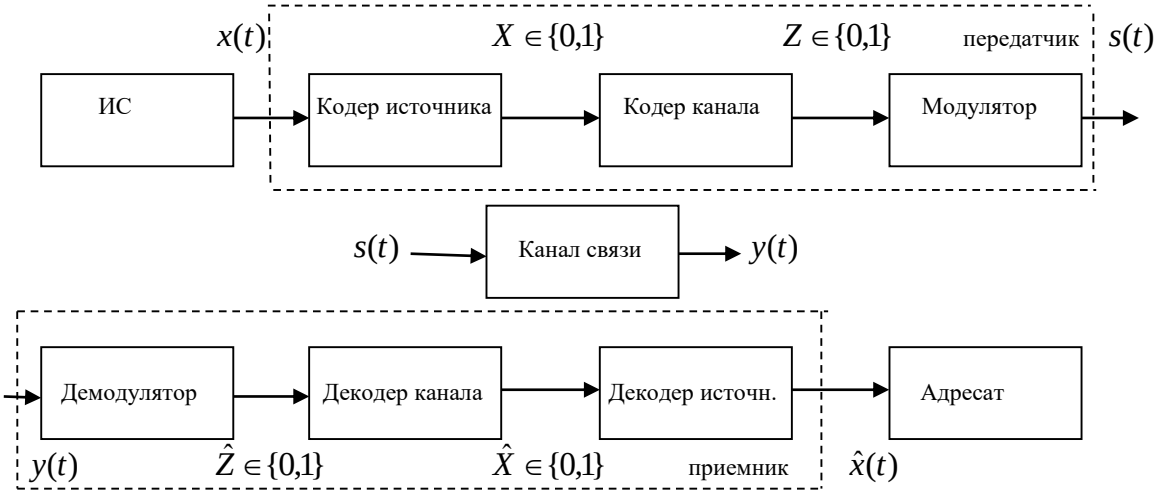
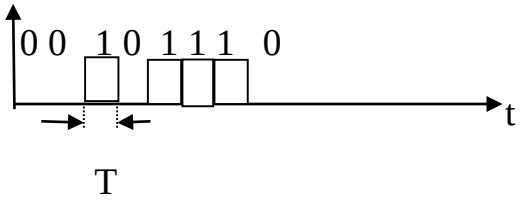


Рисунок 4.1. Обобщенная структурная схема системы передачи и приема сообщений.

1) ИС – источник сообщений. На его выходе – аналоговый $x(t)$ или цифровой сигнал $x_i, i = 1,2,3,....$.

На выходе ДИ информации – дискретные случайные последовательности сообщений (символов), на выходе НИ – непрерывный случайный процесс.

2) Кодер источника – устройство, преобразующее передаваемое сообщение в последовательность двоичных символов $X \in \{0,1\}$. Например, 00101110….. – кодовое слово длины **к** (**к** – количество символов «0» и «1» в кодовом слове).



Символы «0» и «1» называются **битом**. Т – длительность одного бита. Тогда говорят, что двоичные символы следуют со скоростью

$$R = \frac{1}{T} \quad \text{(бит/с)}$$

Кодер источника осуществляет сжатие данных с помощью **эффективного кодирования**. Цель – избавиться от избыточности, которой обладают реальные источники информации, для эффективного использования канала связи при передаче сообщений.

3) Кодер канала – устройство, преобразующее кодовые слова с выхода кодера источника в **помехоустойчивые (корректирующие) коды** Z, которые позволяют обнаруживать и исправлять ошибки в приемнике.

4) Модулятор преобразует последовательность $Z \in \{0,1\}$ в передаваемый по каналу сигнал, соответствующий передаваемому сообщению. Некоторые виды цифровой модуляции рассмотрены в главе 3.

5) Канал связи – техническое устройство или физическая среда распространения сигналов. Например, провода, коаксиальный кабель, волоконно - оптический кабель (ВОК), радиоканал. В канале происходит искажение сигнала из-за помех и шумов. Модели каналов рассмотрены в главе 1.

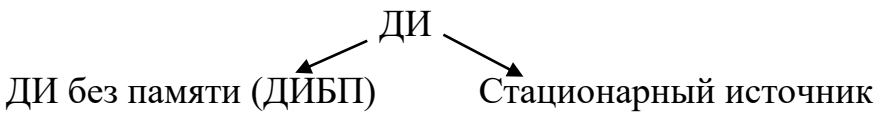
6) Демодулятор преобразует искаженный каналом сигнал в последовательность двоичных символов, т.е. оценивает помехоустойчивый код $\hat{Z}$. Алгоритмы демодуляции (алгоритмы различения сигналов) рассмотрены в главе 2.

7) Декодер канала восстанавливает первоначальную последовательность по полученному помехоустойчивому коду, т.е. оценивает эффективный код $\hat{X}$.

8) Декодер источника – устройство, преобразующее последовательность двоичных символов $\hat{X} \in \{0,1\}$ в сообщение $\hat{x}(t)$ ($\hat{x}_i, i = 1,2,3,....$).

9) Адресат – лицо или устройство, которому предназначено переданное сообщение.

4.1. Дискретный источник информации (ДИ).



Дискретный источник X с алфавитом A из L символов $\{a_1,....,a_L\}$ выдает последовательность букв (символов) $x_i \in \mathrm{A}\,(i=1,2,....)$ выбираемых из этого алфавита. Здесь i - дискретное время. Например, двоичный источник выдает двоичную последовательность 01100010100011110…… . Причем алфавит состоит из $L=2$ символов $A \in \{a_1,a_2\}=\{0,1\}$. Пусть каждый символ алфавита имеет заданную вероятность выбора $p_k = p(a_k) = P\{X=a_k\}, k=1,2,...L$, где $\sum_{k=1}^{L} p_k = 1.$ Рассмотрим две математические модели для ДИ.

1) Если символы выходной последовательности источника статистически независимы, то такой источник называется **источником без памяти (ДИБП)**.

2) Если символы источника взаимозависимы, то можно создать модель на основе статистической стационарности. ДИ называется **стационарным**, если совместные вероятности двух последовательностей длины n $x_1,....,x_n$ и $x_{1+m},...., x_{n+m}$ одинаковы для всех $n \geq 1$ и при всех сдвигах m:

$$p(x_1,.....x_n) = p(x_{1+m},...., x_{n+m}).$$

Т.е. совместные вероятности двух последовательностей инвариантны по отношению к произвольному сдвигу.

4.1.1. Мера информации ДИ.

Рассмотрим две случайные величины X,Y с возможными значениями $X \in \{a_k, k=1,2,..., L\}$ и $Y \in \{b_l, l=1,2,..., M\}$. Пусть мы наблюдаем некоторый выход $Y=b_l$ и желаем количественно определить величину информации, которая содержится в выборке Y относительно события $X=a_k$. Замечание: если X и Y статистически независимы, тогда выбор Y не дает информации о событии X. С другой стороны, если Y однозначно определяется X, то информационное содержание у них одинаковое. **Взаимная информация** определяется как

$$I(a_k,b_l) = \log_2\left(\frac{p(a_k/b_l)}{p(a_k)}\right) \text{ (бит)}, \qquad (4.1)$$

где $p(a_k/b_l) = P\{X=a_k/Y=b_l\}$ - вероятность наступления события $X=a_k$ при условии, что $Y=b_l$.

1) Если X,Y независимы, тогда $p(a_k,b_l) = p(a_k)p(b_l)$, а $p(a_k/b_l) = \frac{p(a_k,b_l)}{p(b_l)} = p(a_k)$

Тогда по формуле (4.1) $I(a_k,b_l) = \log_2(1) = 0$.

2) Если X,Y полностью зависимы, тогда $p(a_k/b_l)=1 \Rightarrow$

$$I(a_k, b_l) = \log_2(\frac{1}{p(a_k)}) = -\log_2(p(a_k)) = I(a_k) \text{ (бит)} \qquad (4.2.)$$

Выражение (4.2.) определяет информацию о X и называется **собственной информацией**. Она является информационной мерой Шеннона.

Свойства собственной информации.

1. Пусть $p(a_k) = 1$, тогда $I(a_k) = 0$, т.е. достоверное событие информации не несет. Собственная информация является меройнеопределенности.

2. Пусть a_k, a_q независимы, тогда $I(a_k, a_q) = -\log_2(p(a_k, a_q)) = -\log_2(p(a_k)p(a_q)) =$

$= -\log_2(p(a_k)) - \log_2(p(a_q)) = I(a_k) + I(a_q),\ k = 1,2,...,L, q = 1,2,....,L.$

3. Если источник выдает за τ_s секунд цифру «0» или «1» ($L = 2$) с равными вероятностями $p(a_k) = 0.5$, то $I(a_k) = -\log_2(0.5) = 1$ бит.

4. Пусть имеется блок a'_k символов источника из n двоичных цифр $a'_k = (10110100\ ...1)_{1\times n}$. Тогда существует 2^n возможных n- битовых блоков, появляющихся с одинаковыми вероятностями $p(a'_k) = 2^{-n}$. Средняя собственная информация такого блока равна $I(a'_k) = -\log_2(p(a'_k)) = -\log_2(2^{-n}) = n$ бит.

Зная взаимную информацию (4.1), связанную с парой событий (a_k, b_l), которые являются возможной реализацией двух случайных величин X, Y, можно получить **среднее значение взаимной информации** следующим образом:

$$I(X,Y) = \sum_{k=1}^{L}\sum_{l=1}^{M} p(a_k, b_l)I(a_k, b_l) = \sum_{k=1}^{L}\sum_{l=1}^{M} p(a_k, b_l)\log_2(\frac{p(a_k, b_l)}{p(a_k)p(b_l)}) = I(Y,X) \qquad (4.3)$$

Аналогично определяем **среднюю собственную информацию** источника:

$$H(X) = \sum_{k=1}^{L} p(a_k)I(a_k) = -\sum_{k=1}^{L} p(a_k)\log_2(p(a_k)) \qquad (4.4)$$

Выражение (4.4) называют **энтропией** ДИ.

Свойства энтропии ДИ.

1. $H(X) \geq 0$, т.е. энтропия – величина неотрицательная.

2. $H(X) = H_{\max}$, если $p(a_k) = p = \frac{1}{L}, k = 1,2,...,L$. Энтропия ДИ максимальна, когда символы на его выходе равновероятны.

$$H_{\max} = -\sum_{k=1}^{L}\frac{1}{L}\log_2(\frac{1}{L}) = \log_2(L) \qquad (4.5)$$

Энтропия является мерой неопределенности источника. Чем больше энтропия, тем больше неопределенность.

3. Энтропия наступления независимых событий $X_1, X_2, ..., X_m$:

$$H(X_1, ..., X_m) = \sum_{i=1}^{m} H(X_i) \qquad (4.6)$$

4. Если сообщения $X_i, i = 1,2,..., m$ зависимы, то вводят понятие условной энтропии:

$$H(X_i / X_{i-1}) = -\sum_{k=1}^{L}\sum_{q=1}^{L} p(a_k, a_q)\log_2(p(a_k / a_q), \qquad (4.7)$$

где $X_i \in \{a_k\}, X_{i-1} \in \{a_q\}$. Формула (4.7) – информационная мера неопределенности, содержащаяся в X_i после наблюдения X_{i-1} . Тогда энтропия совместного наступления событий X_i, X_{i-1} определяется следующим образом:

$$H(X_i, X_{i-1}) = H(X_{i-1}) + H(X_i / X_{i-1}) \qquad (4.8)$$

Формулы (4.7) и (4.8) описывают дискретный марковский источник. Оставшаяся или условная неопределенность всегда меньше исходной (безусловной): $H(X_i / X_{i-1}) \le X(X_i)$.

Вывод: Энтропия ДИ тем больше, чем меньше взаимосвязи между символами, чем более равномерно распределены вероятности появления этих символов и чем больше алфавит источника L.

4.1.2. Производительность, информационная насыщенность и избыточность источника.

Производительность источника – количество средней собственной информации, вырабатываемое в единицу времени:

$$I'(X) = \frac{H(X)}{T_H} \text{ (бит/с)}, \qquad (4.9)$$

где T_H - интервал наблюдений.

Информационная насыщенность определяется как

$$I_H(X) = \frac{H(X)}{H_{\max}} = \frac{I'(X)}{I'_{\max}} . \qquad (4.10)$$

Если $H(X) \to 0$, то и $I_H(X) \to 0$. Если $H(X) \to H_{\max}$, то $I_H(X) \to 1$.

Избыточность источника:

$$r(X) = 1 - I_H(X) = 1 - \frac{H(X)}{H_{\max}} \ . \qquad (4.11)$$

Формула (4.11) показывает недоиспользованность предельных возможностей источника. Чем больше избыточность, тем меньше насыщенность и тем менее эффективно используется канал связи, по которому передается сообщение.

Реальные источники информации обладают большой избыточностью. Поэтому для ее уменьшения прибегают к **эффективному кодированию**.

4.1.3. Эффективное кодирование.

Кодирование ДИБП.

Пусть ДИБП выдает буквы или символы каждые τ_s секунд. Каждый символ выбирается из конечного алфавита $A \in \{a_k\}, k = 1,2,...,L$ с вероятностью $p(a_k)$. Энтропия такого источника определяется по формуле (2.4) и ограничивается сверху значением, вычисляемым по (4.5), т.е. $H(X) \le \log_2(L)$. Как говорилось выше, знак «=» выполняется, если вероятности символов на выходе источника одинаковы и равны $p = \frac{1}{L}$.

1. Кодовые слова фиксированной длины.

Рассмотрим блоковое кодирование, которое состоит в сопоставлении уникального ряда из K двоичных символов, каждому символу источника. Так как существует L возможных символов ДИБП, то число двоичных символов кодера на один символ источника при уникальном кодировании определяется как $K = \begin{cases} \log_2(L), L = 2^Q \\ \lfloor \log_2(L) \rfloor + 1, L \ne 2^Q \end{cases}$, где Q - целое положительное число, $\lfloor \bullet \rfloor$ - наибольшее целое, меньшее, чем $\log_2(L)$. K - **скорость кодирования**. Поскольку $H(X) \le \log_2(L)$, то $K \ge H(X)$. **Эффективность кодирования** определяется отношением $\frac{H(x)}{K}$.

А) Если $L = 2^Q$ и символы источника равновероятны, то $K = H(X)$ и эффективность кодирования равна 1 (100%).

Б) Если $L \ne 2^Q$, но символы источника равновероятны, то K отличается от $H(X)$ самое большее на 1 бит на символ.

В) Если $\log_2(L) >> 1$, то эффективность кодирования высокая.

Г) Если L мало, тогда эффективность кода можно повысить путем кодирования блока из J символов источника за время $J\tau_s$. Для этого надо выбрать L^J уникальных кодовых слов. Используя кодовую

последовательность из K_J двоичных символов, можно образовать 2^{K_J} возможных кодовых слов, причем $K_J \geq J \log_2(L)$. Следовательно, требуется минимальное целое значение для K_J:

$$K_J = \lfloor J \log_2(L) \rfloor + 1 \ .$$

Теперь среднее число символов кода на один символ источника $K = \frac{K_J}{J}$. При эффективность кодирования увеличивается в J раз: $\frac{H(X)}{K} = \frac{H(X)J}{K_J}$. Взяв J достаточно большим, можно эффективность приблизить к 1.

Такие методы кодирования не приводят к искажениям, т.к. кодирование символов источника или блоков символов в кодовые слова выполняется однозначно (уникально). Эти коды называются **бесшумными**.

Теперь рассмотрим ситуацию, когда только часть L^J блоков символов источника кодируется однозначно. Например, $2^{K_J} - 1$ наиболее вероятных J символьных блоков кодируется однозначно. Остальные $L^J - (2^{K_J} - 1)$ блоков длины J представляются одним оставшимся кодовым словом. Такая процедура кодирования вызывает ошибку декодирования каждый раз, когда источник выдает маловероятный блок. Обозначим через p_e вероятность ошибки декодирования. Шеннон в 1948 г. доказал теорему кодирования источника.

Теорема Шеннона кодирования ДИБП. Пусть X - ансамбль символов ДИБП с конечной энтропией $H(X)$. Блоки из J символов источника кодируются в двоичные кодовые слова длины K_J. Тогда для любого $\varepsilon > 0$ p_e можно сделать сколь угодно малой, если выполняется неравенство

$$K = \frac{K_J}{J} \geq H(X) + \varepsilon \qquad (4.12)$$

и J достаточно велико.

2. Кодовые слова переменой длины.

Если символы источника не равновероятны, то более эффективно использовать кодовые слова переменной длины. Пример: код Морзе (19 век). Символам, возникающим более часто, ставятся в соответствие более короткие кодовые слова, а символам, возникающим менее часто, сопоставляются более длинные кодовые слова. Такой метод кодирования, который требует знания вероятностей появления символов источника, называется **энтропийным**.

Рассмотрим пример. Пусть ДИБП имеет алфавит объемом $L=4$, $A=\{a_1,a_2,a_3,a_4\}$. Символы появляются с вероятностями $p(a_1)=\frac{1}{2}, p(a_2)=\frac{1}{4}$, $p(a_3)=p(a_4)=\frac{1}{8}$. Предположим, что они кодируются следующим образом:

код 1: $a_1\to 0,\ a_2\to 01,\ a_3\to 011,\ a_4\to 111$, код 2: $a_1\to 0,\ a_2\to 10,\ a_3\to 110,\ a_4\to 111$

Пусть принимается последовательность 0010010111… . Тогда декодирование кода 1 дает результат: a_1,a_2,a_1,a_2,a_1,a_4 или a_1,a_2,a_1,a_2,a_3. Т.е. имеем не однозначное декодирование. По коду 2: a_1,a_1,a_2,a_1,a_2,a_4. Здесь существует только один вариант декодирования. Ни одно кодовое слово кода 2 не является началом (**префиксом**) другого кодового слова.

В общем, **префиксное условие** кода требует, чтобы для кодового слова длины K $(b_1....b_M b_{M+1}.....b_K)$ не существовало других кодовых слов длины $M<K$ с элементами $(b_1....b_M)$. Это свойство делает кодовые слова однозначно декодируемыми.

Критерий оптимальности однозначно декодируемых кодов переменной длины имеет вид:

$$\overline{K}=\sum_{k=1}^{L} n_k\, p(a_k)=\min\ , \qquad (4.13)$$

где $\overline{K}$ - среднее число бит, приходящихся на один символ источника, n_k - длина k - го кодового слова.

Теорема Шеннона кодирования ДИБП. Пусть X - ансамбль символов ДИБП с конечной энтропией $H(X)$ и выходными символами из алфавита $A=\{a_1,....,a_L\}$ с вероятностями выхода $p(a_k), k=1,2,..,L$. Тогда существует возможность создать код, который удовлетворяет префиксному условию и имеет среднюю длину $\overline{K}$, удовлетворяющую неравенству

$$H(X)\le \overline{K} < H(X)+1 \qquad (4.14)$$

Алгоритм кодирования Фано.

Пример. Рассмотрим ДИБП с объемом алфавита $L=8$. Символы источника имеют вероятности выхода

$$p(a_1)=p(a_2)=\frac{1}{4},\ p(a_3)=p(a_4)=\frac{1}{8},\ p(a_5)=p(a_6)=p(a_7)=p(a_8)=\frac{1}{16}.$$

1) Располагаем сообщения источника в порядке не возрастания их вероятностей.

2) Множество символов разбивается (сверху вниз) на два подмножества так, чтобы суммы вероятностей, входящих в них сообщений, оказались бы равными или минимально отличающимися друг от друга. Сообщениям первого подмножества приписываем «0», а сообщениям второго подмножества – «1» или наоборот.

3) С каждым, из образовавшихся подмножеств, повторяем пункт 2).

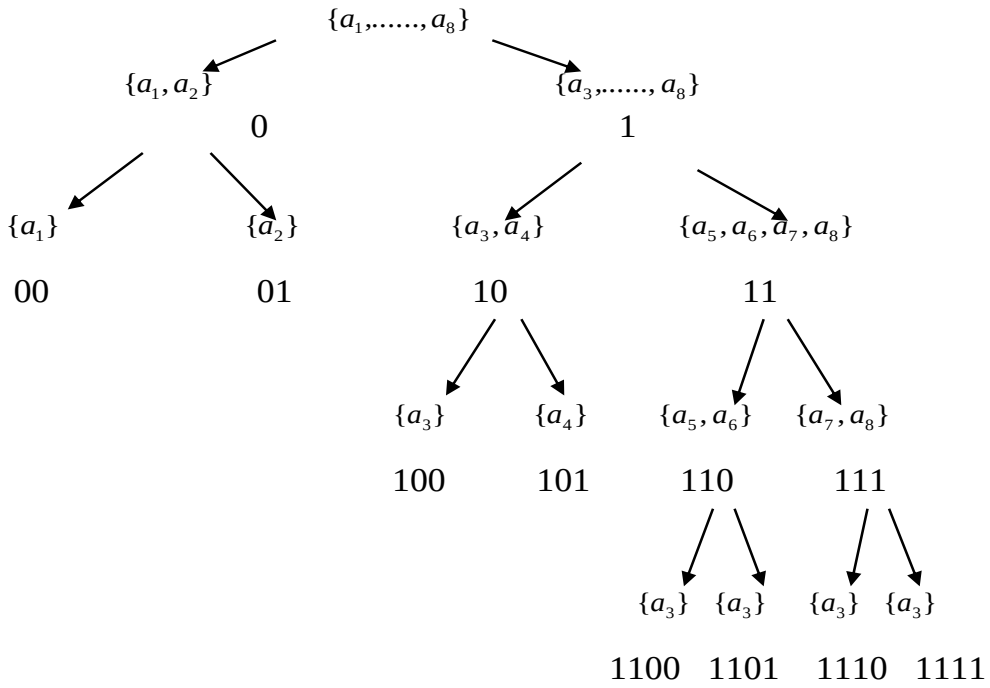


Рисунок 4.2. Кодовое дерево кода Фано.

Символ	Вероятность	Код
a_1	1/4	00
a_2	1/4	01
a_3	1/8	100
a_4	1/8	101
a_5	1/16	1100
a_6	1/16	1101
a_7	1/16	1110
a_8	1/16	1111

Вывод. **Метод кодирования Фано позволяет строить оптимальные префиксные коды в том случае, если вероятности символов источника раны** $p(a_k) = 2^{-c}$ **(для двоичных кодов), где с – положительное целое число.**

Рассмотрим метод кодирования Хаффмена, применение которого к любому произвольному ансамблю символов ДИБП обеспечивает получение оптимального по критерию (4.13) префиксного кода.

Алгоритм кодирования Хаффена.

Критерий оптимальности кодов Хаффмена – минимум средней длины кодового слова (4.13).

Рассмотрим пример. ДИБП выдает символы из алфавита объемом $L = 7$ с вероятностями:
$p(a_1) = 0.2,\ p(a_2) = 0.35,\ p(a_3) = 0.1,\ p(a_4) = 0.3,\ p(a_5) = 0.005,\ p(a_6) = 0.04,\ p(a_7) = 0.005$.

1) Расположить символы источника в порядке убывания (не возрастания) вероятностей.

2) Процесс кодирования начинается с двух наименее вероятных символов a_5, a_7. Эти символы объединяются, причем верхней ветви присваивается «0», нижней «1» или наоборот.

3) Вероятности этих двух ветвей складываются, суммарному узлу присваивается вероятность 0.01.

4) Далее пункты 2), 3) повторяются, пока не исчерпаются символы источника. Вероятность последнего узла равна 1.

Построим кодовое дерево.

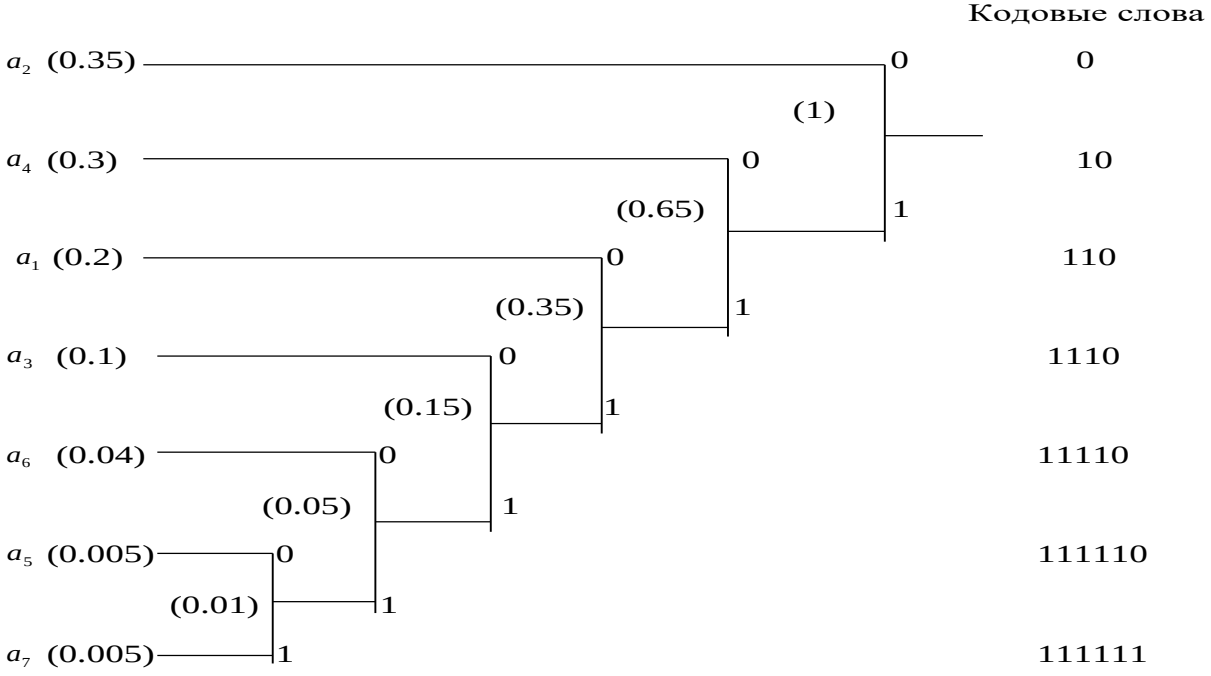


Рисунок 4.3. Кодовое дерево кода Хаффмена.

Кодовые слова записываются по кодовому дереву, проходя по нему, справа налево до кодируемого символа. Полученный код не является единственно возможным.

Энтропия заданного ДИБП $H(X) = 2.11$бит/символ (см. ф-лу (4.4)), средняя длина кодового слова $\overline{K} = 2.21$бит/символ (см. ф-лу (4.13)). Тогда эффективность кода равна $\frac{H(X)}{\overline{K}} = \frac{2.11}{2.21} = 0.95\,(95\%)$.

Как уже отмечалось выше, предложенное кодовое дерево не является единственным. Возможен, например, следующий вариант:

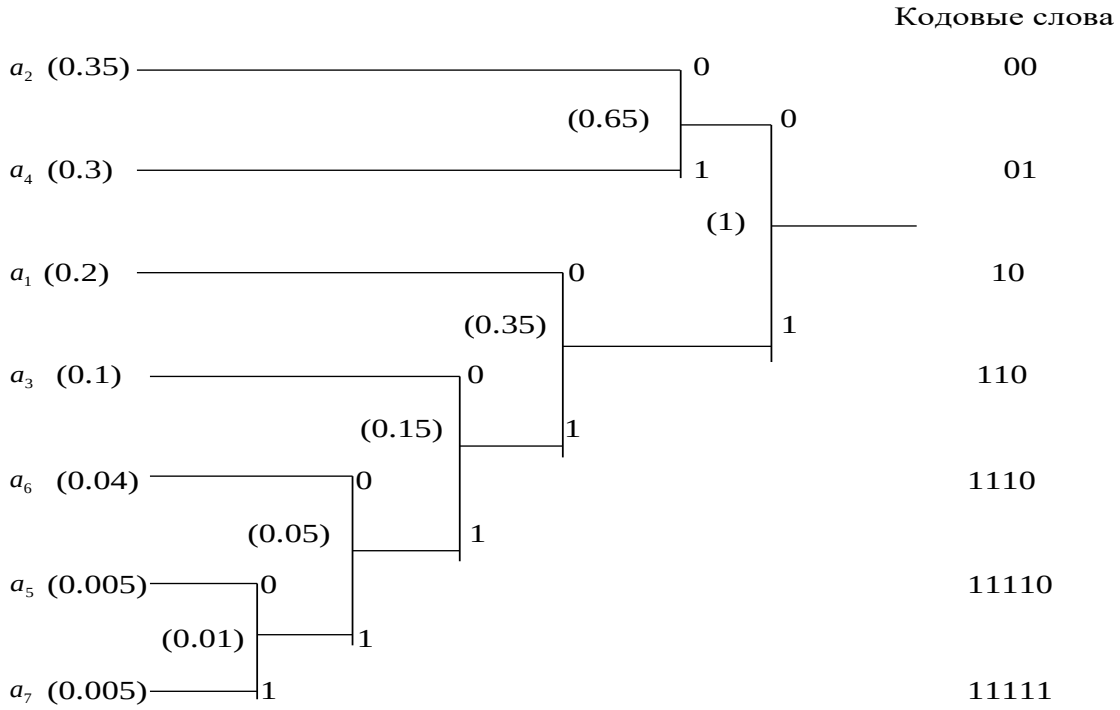


Рисунок 4.4. Альтернативное кодовое дерево кода Хффмена.

Для такой схемы средняя длина кодовой комбинации тоже равна

$\overline{K} = 2.21$бит/символ, поэтому ее эффективность тоже составляет 95%.

Рассмотренный пример показывает посимвольное кодирование. Более эффективно – кодирование блоков из J символов источника одновременно.

В этом случае неравенство (4.14) можно переписать в следующем виде:

$$JH(X) \le \overline{K}_J < JH(X) + 1,$$

где $\overline{K}_J$ - среднее число бит в J символьном блоке. Далее разделим это неравенство на J:

$$H(X) \le \overline{K} < H(X) + \frac{1}{J} \qquad (4.15)$$

Здесь $\overline{K} = \frac{\overline{K}_J}{J}$ - среднее число бит, приходящееся на один символ источника.

Следовательно, $\overline{K}$ можно сделать как угодно близким к $H(X)$, выбирая J достаточно большим, т.е. $\overline{K} \to H(X)$, при $J \to \infty$.

Пример. ДИБП выдает символы из алфавита объемом $L = 3$ с вероятностями $p(a_1) = 0.45, p(a_2) = 0.35, p(a_3) = 0.2$.

Сначала рассмотрим посимвольное кодирование.

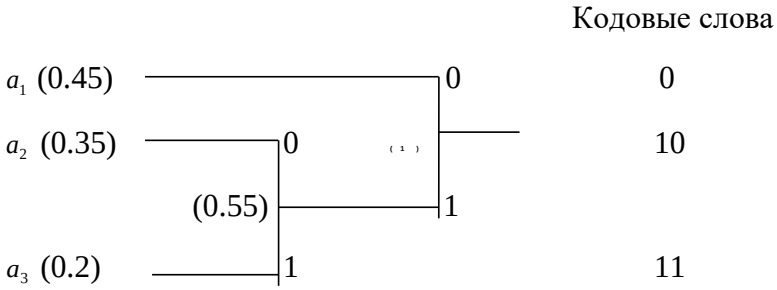


Рисунок 4.5 Кодовое дерево Хаффмена для посимвольного кодирования.

Энтропия источника $H(X) = 1.513$ бит/символ, средняя длина кодовой комбинации $\overline{K} = 1.55$ бит/символ. Эффективность такой схемы кодирования равна $\frac{H(X)}{\overline{K}} = \frac{1.513}{1.55} = 0.976$ (97,6%) .

Если символы закодировать парами, то получим кодовое дерево:

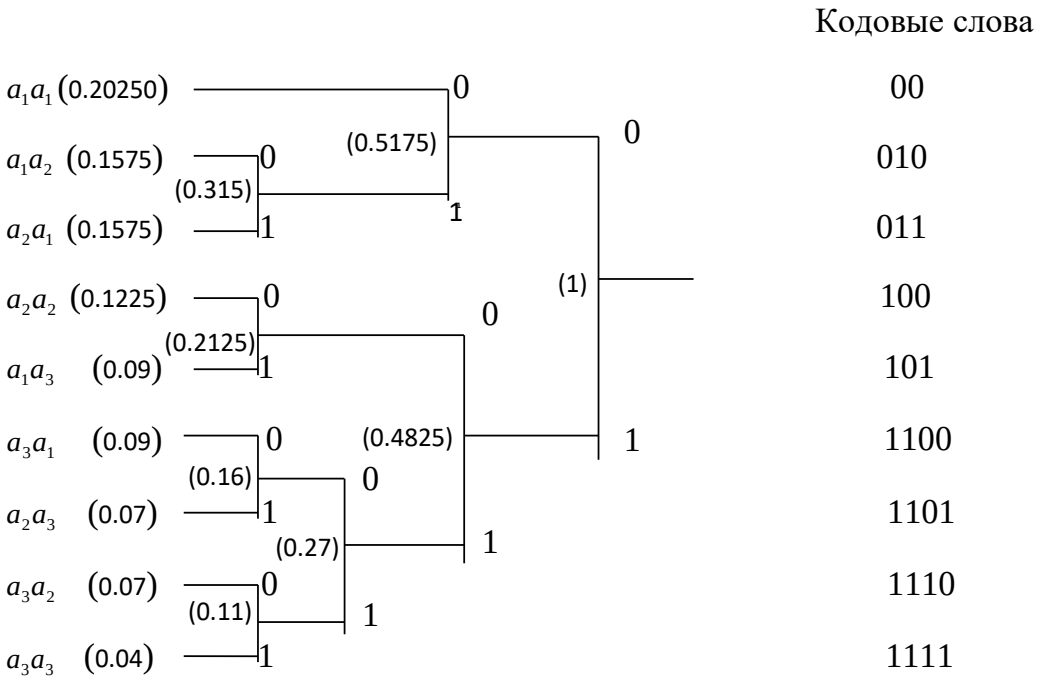


Рисунок 4.6. Кодовое дерево Хаффмена для кодирования пар символов.

Энтропия $H_2(X) = 3.0255$ бит/пара символов, средняя длина кодовой комбинации $\overline{K}_2 = 3.0675$ бит/пара символов ($\overline{K} = \frac{3.0675}{2} = 1.53375$ бит/символ), тогда эффективность схемы кодирования $\frac{3.0255}{3.0675} = \frac{1.513}{1.53375} = 0.986$ (98.6%) увеличилась по сравнению с посимвольным кодированием.