Методические указания
к практическим занятиям по дисциплине
«ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ»
для студентов, обучающихся по направлению 080500.62 –
«Бизнес-информатика»

Томск – 2012

Министерство образования и науки

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ (ТУСУР)

Кафедра автоматизации обработки информации (АОИ)

УТВЕРЖДАЮ
Зав. кафедрой АОИ
профессор
_____ Ю.П. Ехлаков

«__»_____ 2012 г.

Методические указания
к практическим занятиям по дисциплине
«Теоретические основы информатики»
для студентов, обучающихся по направлению 080500.62 –
«Бизнес-информатика»

Разработчик:
профессор
_____М.Т. Решетников

«__»_____ 2012 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Настоящие методические указания направлены на повышение эффективности практических занятий по дисциплине «Теоретические основы информатики».

Сами по себе практические занятия призваны обеспечить закрепление полученных теоретических знаний по теоретическим основам информатики, выработать необходимые навыки вычисления количественных характеристик систем передачи информации, таких как собственная информация, энтропия, скорость передачи информации, пропускная способность канала связи и т.д.

При этом возникают определенные проблемы, связанные с недостаточным знанием предшествующих дисциплин (в частности, теории вероятностей), неоднозначностью или неправильностью толкования отдельных терминов и положений теоретической части курса, неумением учитывать объективные ограничения реальных практических ситуаций.

Определенная часть этих проблем, возникших у автора в процессе преподавания курса «Теория информации» студентам специальности «Автоматизированные системы обработки информации и управлении», стала содержанием настоящих методических указаний.

Методическое пособие соответствует рабочей программе по дисциплине «Теоретические основы информатики», которая, в свою оче-

редь, составлена в соответствии с учебным планом направления 080500 «Бизнес-информатика».

Пособие может быть полезно для преподавателей и студентов, преподающих и осваивающих эту дисциплину в рамках других инженерных специальностей, где такая дисциплина (и практические занятия по ней) предусмотрены учебными планами.

Методические указания ни в коей мере не заменяют имеющиеся пособия по практике и, в частности, предполагают активное использование студентами сборников задач, приведенных в списке литературы.

1. ФУНДАМЕНТАЛЬНЫЕ ПОЛОЖЕНИЯ ТЕОРИИ ИНФОРМАЦИИ. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Задачи этого раздела предназначены для закрепления определения, понятий и свойств основных элементов систем передачи информации, таких как источник информации, приемник, передатчик, алфавит, канал связи, линия связи и др.

Для их решения необходимо лишь точное знание определений, а также обычный здравый смысл.

Пример 1.1. Что будет алфавитом для трехсекционного уличного светофора?

Ответ: Набор из трех цветов секций светофора.

Как видно, ответ достаточно очевидный, однако полная и точная его формулировка рождается не сразу. Задачи этого раздела имеет смысл решать на практических занятиях в режиме свободного обсуждения, когда на основе отдельных мнений и различных формулировок постепенно появляется полный ответ.

При самостоятельном решении таких задач не следует приводить первый пришедший в голову ответ: необходимо продумать все возможные варианты функционирования предложенной в задаче системы и свести их воедино. Так, первой реакцией на вопрос приведенного примера является ответ: «набор трех цветов», после некоторого обсуждения возникают дополнительные варианты сигналов светофора, например, то, что красный и желтый цвета могут гореть одновременно (а желтый и зеленый — нет), или то, что желтый цвет может мигать. Каждый из этих сигналов несет свою информацию.

Тем не менее, обращаясь к аналогу обычного алфавита русского языка, делаем вывод о том, что алфавит светофора — это именно три цвета, а их сочетания, режим включения — это сообщения, составленные из «букв» алфавита.

Пример 1.2. Сигнальщики двух кораблей в открытом море передают друг другу сообщения с помощью прожекторного телеграфа. Что

в данном случае будет являться каналом связи, линией связи? Какие помехи могут быть в таком канале связи?

Ответ: В соответствии с определением, каналом связи является совокупность «передатчик – линия связи – приемник». Передатчик в данном случае — прожектор, приемник — органы зрения сигнальщика. Линия связи, по определению — любая физическая среда, обеспечивающая передачу сигнала от передатчика к приемнику. В примере это воздушная среда, а возможные помехи в канале связи: атмосферные явления (туман, дождь, снег, гроза), высокие волны между кораблями, технические неполадки прожекторов.

В процессе решения задач этого раздела формируется понимание сложности любых систем передачи информации, разнообразия условий, в которых они функционируют. Важно, что при этом возникают конкретные образы абстрактных понятий «знак», «сигнал», «источник информации», «алфавит», «сообщение», что в дальнейшем облегчает понимание других разделов теории информации.

2. КОЛИЧЕСТВЕННАЯ МЕРА ИНФОРМАЦИИ

Строго говоря, единственной формулой, на которую опирается решение задач по вычислению количества собственной информации, является формула, содержащаяся в определении собственной информации:

$$I(x_i) = -\log p(x_i), \quad i = 1,2,...,k \tag{2.1}$$

где x_i — сообщение, входящее в ансамбль сообщений из k элементов;

$p(x_i)$ — вероятность сообщения x_i;

$I(x_i)$ — количество собственной информации (или собственная информация) в сообщении x_i.

Однако, несмотря на простоту формулы (2.1), следует иметь в виду несколько принципиальных соображений, связанных с ее применением.

1. Важным является понятие ансамбля сообщений

$$\{X, p(x)\} = \begin{Bmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{Bmatrix},$$

где x_i, $i = 1,2,...,k$ — сообщения;

p_i, $i = 1,2,...,k$ — их вероятности,

причем

$$\sum_{i=1}^{k} p(x_i) = 1 .$$

Ансамбль сообщений, таким образом, является аналогом полного пространства состояний в теории вероятностей. Как правило, в задачах не заданы ни ансамбль, ни вероятность сообщения, — их необходимо найти, прежде чем применять формулу (2.1). Именно поэтому задачи на вычисление собственной информации (как и энтропии – см. в разделе 3) — это, прежде всего, обычные вероятностные задачи.

Пример 2.1. Солдат стреляет по мишени из винтовки. Вероятность осечки при каждом выстреле $p_1 = 0{,}2$. Вероятность попадания в мишень при одном выстреле $p_2 = 0{,}7$. Солдат производит два выстрела. Сколько информации содержится в сообщении, что мишень поражена?

Решение. Найдем сначала вероятность непопадания в мишень после двух выстрелов. При одном выстреле либо произошла осечка (вероятность p_1), либо солдат не попал в мишень (вероятность $(1-p_1)\cdot(1-p_2)$). Общая вероятность непопадания после первого выстрела есть ($p_1+(1-p_1)\cdot(1-p_2)$). Вероятность непопадания за два выстрела равна $1-(p_1+(1-p_1)\cdot(1-p_2))^2$, так как ансамбль в данном случае состоит из двух возможных событий: мишень поражена и мишень не поражена. Подставляя в последнюю формулу числовые значения из условия задачи, получим

$$p = 1-(0{,}2+0{,}8\cdot 0{,}3)^2 = 1-0{,}44^2 = 0{,}8064 .$$

Далее применяем формулу (2.1)

$$I = -\log p = -\log_2 0{,}8064 \approx 0.31 \text{ (бит).}$$

Ответ: примерно 0,31 бит.

2. В определении количества собственной информации (2.1) основание логарифма роли не играет, однако в практических расчетах всегда нужно учитывать, какой логарифм используется, поскольку этим определяется единица измерения информации. Напомним, что если в (2.1) логарифм двоичный, полученная информация измеряется в битах, если десятичный — в дитах, если натуральный — в натах. При решении более сложных задач, чем вычисление одного единственного значения количества информации $I(x_i)$, недопустимо смешение единиц измерения, когда, например, для одного события информация вычисляется в битах, для другого - в натах.

3. Для самопроверки правильности решения задач на вычисление количества собственной информации следует помнить одно из основ-

ных свойств информации: информация тем больше, чем меньше вероятность события.

Полезно помнить основные формулы и теоремы теории вероятностей, позволяющие в каждом конкретном случае получить необходимую вероятность для последующего вычисления собственной информации события. Это, в первую очередь, формула полной вероятности

$$p(a) = p(b_1) \cdot p(a/b_1) + p(b_2) \cdot p(a/b_2) + \ldots + p(b_k) \cdot p(a/b_k),$$

если событие a может наступить только при условии одного из несовместных событий $b_1, b_2, \ldots, b_k$, образующих полную группу. Часто используется также формула Байеса:

$$p(b_i/a) = \frac{p(b_i) \cdot p(a/b_i)}{p(b_1) \cdot p(a/b_1) + p(b_2) \cdot p(a/b_2) + \ldots + p(b_k) \cdot p(a/b_k)},$$

(условия те же, что и в формуле полной вероятности).

Пример 2.2. Имеется два набора деталей. Вероятность того, что деталь первого набора стандартна, равна 0,8, а второго — 0,9. Сколько информации содержится в сообщении о том, что взятая наудачу деталь (из наугад взятого набора) — стандартна?

Решение. Обозначим через a событие «извлеченная деталь стандартна». Деталь может быть извлечена либо из первого набора (событие b_1), либо из второго (событие b_2). Вероятность того, что деталь вынута из первого набора $p(b_1) = \frac{1}{2}$, аналогично $p(b_2) = \frac{1}{2}$.

Условная вероятность того, что из первого набора будет извлечена стандартная деталь, $p(a/b_1) = 0{,}8$, для второго набора аналогичная вероятность $p(a/b_2) = 0{,}9$.

Применяя формулу полной вероятности, получим

$$p(a) = p(b_1) \cdot p(a/b_1) + p(b_2) \cdot p(a/b_2) = 0{,}5 \cdot 0{,}8 + 0{,}5 \cdot 0{,}9 = 0{,}85.$$

Информацию вычислим по формуле (2.1):

$$I(a) = -\log p(a) = -\log 0{,}85 \approx 0{,}23 \text{ (бит).}$$

3. ЭНТРОПИЯ ДИСКРЕТНЫХ СООБЩЕНИЙ

В этом разделе, так же как и в предыдущем, используется одна основная формула, входящая в определение энтропии ансамбля $\{X, p(x)\}$:

$$H(X) = M\{I(x)\} = -\sum_{x \in X} p(x) \cdot \log p(x), \qquad (3.1)$$

Или, в дискретном случае

$$H(X) = -\sum_{i=1}^{k} p(x_i) \cdot \log p(x_i), \qquad (3.2)$$

где $H(X)$ — обозначение энтропии ансамбля X, а остальные обозначения аналогичны обозначениям в (2.1).

Здесь следует обратить внимание на то, что понятие энтропии относится к ансамблю сообщений ***в целом*** (а не к отдельным сообщениям), поскольку энтропия, по определению, есть среднее количество собственной информации в сообщениях ансамбля.

Ответы к задачам типа «энтропия этого сообщения равна 5 бит» или «собственная информация полученного ансамбля равна 2,3 бит» свидетельствуют о принципиальном непонимании математической сути понятий «количество собственной информации» и «энтропия». В этом случае полезно обращение к известному примеру «средней температуры по больнице».

Для решения задач на вычисление энтропии, как и в предыдущем разделе, необходимо, в первую очередь, найти сам ансамбль сообщений и вероятности всех его элементов. После этого механическое применений формулы (3.2) трудности не представляет.

Пример 3.1. Рассчитать энтропию (в битах) ансамбля, связанного с получением случайных чисел при бросании двух тетраэдров.

Решение. Тетраэдр — правильная треугольная пирамида с четырьмя гранями, на которые нанесены числа от 1 до 4. При бросании двух тетраэдров все возможные сочетания выпавших очков и суммы этих очков при каждом сочетании можно свести в следующую таблицу:

2-й тетраэдр / 1-й тетраэдр	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

В ансамбль возможных реализаций включаем все полученные семь сумм выпавших очков с вероятностями получения этих сумм:

$$\{X, p(x)\} = \begin{Bmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \frac{1}{16} & \frac{2}{16} & \frac{3}{16} & \frac{4}{16} & \frac{3}{16} & \frac{2}{16} & \frac{1}{16} \end{Bmatrix}.$$

Теперь применяем формулу (3.2):

$$H(X) = -\sum_{i=1}^{k} p(x_i) \cdot \log p(x_i) =$$

$$= -\left(\frac{1}{16} \cdot \log \frac{1}{16} \cdot 2 + \frac{2}{16} \cdot \log \frac{2}{16} \cdot 2 + \frac{3}{16} \cdot \log \frac{3}{16} \cdot 2 + \frac{4}{16} \cdot \log \frac{4}{16} \right) \approx 2{,}6$$

(все логарифмы двоичные, так как по условию задачи решение нужно получить в битах).

Ответ: примерно 2,6 бит.

Из примера видно, что даже в такой простой задаче вычисления достаточно трудоемки, но не настолько, чтобы использовать сложную вычислительную технику. Рекомендуется либо использовать при решении задач микрокалькулятор, либо обращаться к таблицам логарифмов или значений $-p \log p$, имеющимся практически в каждом учебном пособии по теории информации (Приложения 1,2). В отдельных задачах, где главная трудность — получение ансамбля сообщений и их вероятностей, возможно закончить решение на стадии формирования ансамбля.

Энтропия трактуется как степень или мера неопределенности в эксперименте с получением сообщений ансамбля. Поэтому задачи с формулировкой «найти неопределенность исхода эксперимента» сводятся к вычислению энтропии ансамбля, полученного в ходе эксперимента.

Пример 3.2. В каком соотношении находятся энтропии двух ансамблей

$$\{X, p(x)\} = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 \\ 0{,}25 & 0{,}5 & 0{,}125 & 0{,}125 \end{Bmatrix}$$

и

$$\{Y, p(y)\} = \begin{Bmatrix} y_1 & y_2 & y_3 & y_4 \\ 0{,}5 & 0{,}25 & 0{,}125 & 0{,}125 \end{Bmatrix} ?$$

Решение. Непосредственным вычислением энтропий ансамблей X и Y получаем, что их энтропии одинаковы и равны 1,75 бит. Это легко видеть и по структуре ансамблей, отличающихся перестановкой вероятностей первых двух элементов.

Ответ: неопределенности ансамблей одинаковы.

В некоторых случаях решению задач на вычисление собственной информации и энтропии помогает знание свойств энтропии, в частности, свойства энтропии достигать максимального значения, равного

$\log k$, если все события ансамбля из k элементов равновероятны и, следовательно, вероятность каждого из них равна $\frac{1}{k}$.

Пример 3.3. При угадывании целого числа в диапазоне от 1 до N было получено 7 бит информации. Чему равно N?

Решение: Поскольку предполагается, что вероятности загаданных чисел одинаковы (и равны $\frac{1}{N}$), информация, содержащаяся в сообщении об угадывании любого из этих чисел, равна $-\log\frac{1}{N}$, а по условию задачи равна 7 бит. Это значит, что логарифм в формуле двоичный и нужно решать уравнение $-\log_2 N = 7$, что дает

Ответ: $N = 2^7 = 128$.

С этим свойством энтропии связано и понятие избыточности информации (в ансамбле, в алфавите и т.п.).

Пример 3.4. Источник сообщений вырабатывает ансамбль символов $\{X, p(x)\} = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 \\ 0{,}25 & 0{,}5 & 0{,}125 & 0{,}125 \end{Bmatrix}$. Символы в последовательности статистически независимы. Вычислить энтропию источника и определить избыточность алфавита.

Решение. Энтропия достаточно просто вычисляется по формуле (3.2):

$H(X) = -0{,}5\log 0{,}5 - 0{,}25\log 0{,}25 - 2\cdot 0{,}125\log 0{,}125 = 1{,}75$ бит.

Максимального значения энтропия достигает в случае равновероятных символов, и это максимальное значение равно логарифму количества символов в ансамбле. В нашем случае

$H_{\max} = \log 4 = 2$.

Вычислим избыточность:

$\gamma = \frac{H}{H_{\max}} = \frac{1{,}75}{2} = 0{,}875$.

Ответ: энтропия источника 1,75 бит; избыточность 0,875.

4. ЭНТРОПИЯ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Поскольку абсолютной энтропии непрерывных случайных величин не существует, пользуются понятием дифференциальной энтропии, определяемой для случайной величины с плотностью распределе-

ния вероятностей $f(x)$ относительно некоторой другой (эталонной) случайной величины с известным, например, равномерным распределением:

$$H_{\varepsilon}(x) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx \qquad (4.1)$$

Значок ε при записи дифференциальной энтропии обычно опускают, но при этом всегда нужно помнить, что (4.1) — это относительная энтропия, свойства которой не полностью совпадают со свойствами энтропии дискретных случайных величин. В частности, $H_{\varepsilon}(x)$ может принимать отрицательные значения.

Пример 4.1. [2] Вычислить дифференциальную энтропию случайной величины, заданной распределением

$$f(x) = \begin{cases} 0, x < 0, \\ x^2, 0 \le x \le 1, \\ 1, x > 1. \end{cases}$$

Решение. Подставим функцию $f(x)$ в формулу (4.1), разбив ее на три части:

$$H(x) = -\int_{-\infty}^{0} 0\log 0dx - \int_{0}^{1} x^2 \log x^2 dx - \int_{1}^{\infty} 1\log 1dx\,.$$

Легко убедиться, что первое и третье слагаемое обращаются в нуль, а второе слагаемое вычисляется по табличному интегралу:

$$\int x^p \ln x dx = x^{p+1}\left[\frac{\ln x}{p+1} - \frac{1}{(p+1)^2}\right].$$

С учетом табличного интеграла имеет смысл вычислять энтропию в натах, тогда

$$H(x) = -\int_{0}^{1} x^2 \ln x^2 dx = -2\int_{0}^{1} x^2 \ln x dx = -2\cdot x^3\left[\frac{\ln x}{3} - \frac{1}{9}\right]\Bigg|_0^1 = 0{,}22 \text{ нат.}$$

Ответ: 0,22 нат.

Интерес представляют так называемые экстремальные распределения, т.е. распределения непрерывных случайных величин, обладающие максимальной энтропией.

Пример 4.2. Известно, что область возможных значений случайной величины x ограничена интервалом $[a,b]$: $a \le x \le b$. Найти распределение, обладающее наибольшей энтропией.

Решение. Вариационная задача: найти функцию $p(x)$, обеспечивающую максимум функционала

$$H_\varepsilon(x) = -\int_a^b p(x)\log p(x)dx$$

при дополнительном условии

$$\int_a^b p(x)dx = 1. \qquad (*)$$

Для этого необходимо максимизировать функцию

$$\int_a^b F(x,p)dx = \int_a^b [-p(x)\log p(x) + \lambda p(x)]dx,$$

что приводит к уравнению:

$$\frac{\partial F(x,p)}{\partial p} = [-1 - \log p + \lambda] = 0,$$

откуда $p = \exp(\lambda - 1)$.

Неизвестную константу λ найдем из условия (*):

$$\int_a^b \exp(\lambda - 1)dx = \exp(\lambda - 1)\int_a^b dx = [\exp(\lambda - 1)]\cdot(b-a) = 1.$$

Отсюда:

$$p(x) = \begin{cases} 0, x < a, \\ \dfrac{1}{b-a}, a \le x \le b, \\ 0, x > b. \end{cases}$$

Ответ: при заданных ограничениях распределением, обладающим наибольшей энтропией, является равномерное распределение.

5. СКОРОСТЬ ПЕРЕДАЧИ ИНФОРМАЦИИ И ПРОПУСКНАЯ СПОСОБНОСТЬ КАНАЛОВ СВЯЗИ

Напомним основные формулы, необходимые для решения задач этого раздела.

Скорость передачи информации

$$I(A,B) = H(A) - H(A/B) \qquad (5.1)$$

где A — сигнал на входе источника;

B — сигнал, поступающий к потребителю;

$H(A)$ — энтропия источника (скорость создания информации);

$H(A/B)$ — скорость потери информации в канале (ненадежность канала).

Пропускная способность канала

$$C = \max[H(A) - H(A/B)] \tag{5.2}$$

где максимум берется по всем возможным кодам.

При этом в канале без шумов

$$C = \log m \tag{5.3}$$

где m — объем алфавита передаваемых сообщений.

Пример 5.1. В информационном канале используется алфавит, содержащий 8 символов. Длительности всех символов одинаковы и равны $\tau = 2$ мкс. Определить пропускную способность канала при отсутствии шумов.

Решение. Подставляем $m = 8$ в формулу (5.3) и получаем

$$C = \frac{\log 8}{\tau} = \frac{4}{2 \cdot 10^{-6}} = 2 \cdot 10^6 \text{ бит/сек.}$$

Ответ: $2 \cdot 10^6$ бит/сек.

При вычислении скорости передачи информации необходимо знать свойства канала связи. Они однозначно определяются матрицей переходных вероятностей, или канальной матрицей $P(b_j / a_i)$, задающей вероятности приема сигнала b_j при условии отправки сигнала a_i. Заметим, что канальная матрица может задаваться и в виде $P(a_i / b_j)$, когда известны условные вероятности отправки сообщения a_i при наличии принятого сообщения b_j.

Имеет смысл привести формулы, связывающие между собой энтропии источника сообщений A, приемника B и условные энтропии $P(A/B)$ и $P(B/A)$.

а) неопределенность отправленных элементов (энтропия источника):

$$H(A) = -\sum_{i=1}^{m} p(a_i) \log p(a_i) ;$$

б) неопределенность полученных элементов (энтропия приемника):

$$H(B) = -\sum_{j=1}^{m} p(b_j) \log p(b_j);$$

в) неопределенность получения элементов при зафиксированном отправляемом элементе a_i (частная энтропия принятых элементов):

$$H(B/a_i) = -\sum_{j=1}^{m} p(b_j/a_i) \log p(b_j/a_i);$$

г) полная энтропия принятых элементов:

$$H(B/A) = \sum_{i=1}^{m} H(B/a_i) \log p(a_i);$$

д) неопределенность отправки элементов при зафиксированном полученном элементе b_j (частная энтропия отправляемых элементов):

$$H(A/b_j) = -\sum_{i=1}^{m} p(a_i/b_j) \log p(a_i/b_j);$$

е) полная энтропия отправляемых элементов:

$$H(A/B) = \sum_{j=1}^{m} H(A/b_j) \log p(b_j).$$

Скорость передачи информации

$I(A,B) = H(A) - H(A/B) = H(B) - H(B/A)$.

Этих формул достаточно для решения задач на вычисление скорости передачи информации и ненадежности передачи в каналах с шумами. Для простейшей самопроверки следует помнить, что величины $I(A,B), H(A), H(B), H(A/B), H(B/A)$ — неотрицательные.

Приведем пример типичной задачи этого типа.

Пример 5.2. Двоичный симметричный канал без памяти задан канальной матрицей

$$P(B/A) = \begin{Vmatrix} 1-p & p \\ p & 1-p \end{Vmatrix}$$

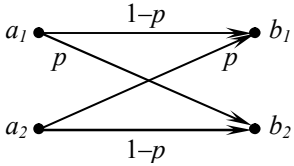
и вероятностями элементов на входе: $p(a_1) = p(a_2) = \frac{1}{2}$.

Построить граф канала, найти скорость создания информации $H(A)$, скорость передачи информации.

Решение. Сначала находим скорость создания информации как энтропию источника:

$$H(A) = -\sum_{i=1}^{2} p(a_i)\log p(a_i) = -p(a_1)\log p(a_1) - p(a_2)\log p(a_2) = 1.$$

Строим граф канала:



Граф должен быть взвешенным, т.е. на дугах, соединяющих вершины, должны быть написаны значения переходных вероятностей.

Далее возможны два пути решения: либо вычислить скорость передачи информации $I(A,B)$, а ненадежность найти как разность $H(A)$ и $I(A,B)$, либо, наоборот, вычислить ненадежность $H(A/B)$, а скорость найти по формуле (5.1).

Найдем ненадежность канала:

$$H(A/B) = -\sum_{i=1}^{2}\sum_{j=1}^{2} p(b_j)\cdot p(a_i/b_j)\log p(a_i/b_j) =$$

$$= -\sum_{j=1}^{2} p(b_j)\sum_{i=1}^{2} p(a_i/b_j)\log p(a_i/b_j) = -\,p(b_1)[p(a_1/b_1)\cdot\log p(a_1/b_1) +$$

$$+\,p(a_2/b_1)\log p(a_2/b_1)] - p(b_2)[p(a_1/b_2)\cdot\log p(a_1/b_2) + p(a_2/b_2)\cdot$$

$$\cdot\log p(a_2/b_2)] = -p(b_1)[(1-p)\log(1-p) + p\log p] - p(b_2)[p\log p +$$

$$+(1-p)\log(1-p)] = -[p(b_1) + p(b_2)]\cdot[p\log p + (1-p)\log(1-p)].$$

Так как $p(b_1) + p(b_2) = 1$, окончательно получаем:

$$H(A/B) = -p\log p - (1-p)\log(1-p).$$

Теперь осталось найти скорость передачи информации.

$$I(A,B) = H(A) - H(A/B) = 1 - [-p\log p - (1-p)\log(1-p)] =$$

$$= 1 + p\log p + (1-p)\log(1-p).$$

Ответ: Скорость создания информации $H(A) = 1$, скорость передачи информации $I(A,B) = 1 + p\log p + (1-p)\log(1-p)$, ненадежность канала $H(A/B) = -p\log p - (1-p)\log(1-p)$.

Необходимо обратить внимание на то, что при вычислении $H(A/B)$ использовались вероятности $p(a_i/b_j)$ из матрицы, приведенной в условии задачи. Это возможно в простом случае симметричного канала, иначе вероятности $p(a_i/b_j)$ приходится вычислять по

заданным значениям $p(a_i)$ и $p(b_j / a_i)$. Для этого полезно помнить известные формулы теории вероятностей:

Если заданы две дискретных последовательности случайных величин $a_i, i = 1,2,...,m$ и $b_j, j = 1,2,...,s$, то:

- $p(a_i) = \sum_{j=1}^{s} p(a_i, b_j)$;
- $p(b_j) = \sum_{i=1}^{m} p(a_i, b_j)$;
- $p(a_i / b_j) = \dfrac{p(a_i, b_j)}{p(b_j)}$;
- $p(b_j / a_i) = \dfrac{p(a_i, b_j)}{p(a_i)}$;
- $p(a_i, b_j) = p(a_i) \cdot p(b_j / a_i) = p(b_j) p(a_i / b_j)$.

6. КОДИРОВАНИЕ В КАНАЛАХ БЕЗ ШУМОВ. КОДЫ ШЕННОНА — ФЭНО И ХАФФМАНА

В каналах без шумов сигналы передаются без искажений, поэтому единственная проблема, которую необходимо решать при передаче — экономичность и, следовательно, увеличение скорости передачи информации.

Основной результат, определяющий свойства оптимального кода, сформулирован в фундаментальной теореме Шеннона о кодировании в каналах без шумов. Согласно этой теореме длина кодового слова L не может быть меньше, чем отношение $\dfrac{H(U)}{\log m}$, где $H(U)$ — энтропия кодируемого множества сигналов U, m — количество символов в алфавите.

Не менее важна вторая часть теоремы, которая гласит, что если вероятности сигналов не являются целочисленными отрицательными степенями числа m, то точное достижение границы $L = \dfrac{H(U)}{\log m}$ невозможно; но при кодировании достаточно длинными блоками к этой границе можно сколь угодно приблизиться.

Из этого следует, во-первых, что код может быть оптимальным и средняя длина кодового слова для двоичного кодирования совпадает с энтропией источника, если вероятности сигналов являются целочис-

ленными отрицательными степенями двойки: $\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\ldots$. Во-вторых, эффективность кодирования повышается, если кодировать не сами сигналы, а блоки, получаемые путем группировки исходных сигналов по 2, 3 и т.д.

Рассмотрим основные положения теоремы Шеннона на нескольких примерах двоичного кодирования с помощью кодов Шеннона – Фэно и Хаффмана.

Код Шеннона Фэно.

Процедура построения кода направлена на максимальное удовлетворение требований равновероятности и независимости символов в кодовых словах.

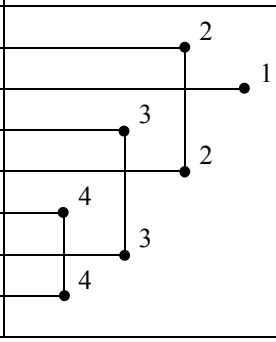
Кодируемые сигналы располагаются в порядке убывания их вероятностей. Множество кодируемых сигналов разбивается на две группы с примерно одинаковыми суммарными вероятностями.

Если кодируемый сигнал принадлежит к первой группе, первый символ кодового слова для него — 0, если ко второй — 1.

Затем каждая группа разбивается на две по возможности равновероятные подгруппы и формируется вторые символы кодовых слов. Такое разбиение на подгруппы продолжается до тех пор, пока в подгруппе не останется один сигнал.

Пример 6.1. Имеются 8 сигналов $u_1,u_2,\ldots,u_8$ с вероятностями, приведенными в таблице 6.1. Требуется закодировать их кодом Шеннона – Фэно.

Таблица 6.1

Кодируемые сигналы	Вероятности	Разбиения	Кодовые слова
u_1	0,25		0 0
u_2	0,25		0 1
u_3	0,125		1 0 0
u_4	0,125		1 0 1
u_5	0,0625		1 1 0 0
u_6	0,0625		1 1 0 1
u_7	0,0625		1 1 1 0
u_8	0,0625		1 1 1 1

Определим эффективность кода. Для этого вычислим энтропию $H(U)$ и среднюю длину L кодовых слов.

$$H(U) = -\sum_{i=1}^{8} p(u_i)\log p(u_i) = 2{,}75 \text{ (бит)}.$$

$$L = \sum_{k=1}^{8} n_k\, p(u_k) = 2\cdot 2\cdot 0{,}25 + 3\cdot 2\cdot 0{,}125 + 4\cdot 4\cdot 0{,}0625 = 2{,}75\,.$$

Поскольку $L = H(U)$, код в данном случае оптимален. Это стало возможным благодаря тому, что вероятности $p(u_i)$ равны целочисленным отрицательным степеням двойки.

При решении аналогичных задач целесообразно проверять правильность кодирования, имея в виду следующее:

а) код должен получиться префиксным, т.е. ни одно кодовое слово нельзя получить из другого, более короткого, путем добавления дополнительных символов;

б) сигналам с большей вероятностью соответствуют более короткие кодовые слова;

в) средняя длина кодового слова при одинаковых вероятностях сигналов всегда будет находиться в промежутке $(s, s+1)$, если количество сигналов находится в промежутке $(2^s, 2^{s+1})$.

Пример 6.2. Ваш собеседник задумал два целых положительных числа, не превосходящих 3. Вы пытаетесь отгадать сумму заданных чисел, задавая собеседнику вопросы, на которые он отвечает лишь

«да» или «нет». Сколько в среднем нужно задать вопросов, чтобы отгадать сумму и каково максимальное число вопросов, необходимых для отгадывания при правильной стратегии?

Решение. Сначала сформируем ансамбль возможных значений сумм загаданных чисел.

Первое число / Второе число	1	2	3
1	2	3	4
2	3	4	5
3	4	5	6

Если исходить из того, что загаданные числа равновероятны, искомый ансамбль имеет вид:

$$\{U, p(u)\} = \begin{Bmatrix} 2 & 3 & 4 & 5 & 6 \\ \frac{1}{9} & \frac{2}{9} & \frac{3}{9} & \frac{2}{9} & \frac{1}{9} \end{Bmatrix}.$$

Таким образом, отгадать нужно одно из чисел 2, 3, 4, 5, 6. Для формирования стратегии воспользуемся кодом Шеннона – Фэно, но в данном случае в таблице для кодирования числа расположим в порядке их следования, а не в порядке убывания вероятностей:

Кодируемые числа	Вероятности	Разбиения	Кодовые слова
2	$\frac{1}{9}$	2	0 0
3	$\frac{2}{9}$	1	0 1
4	$\frac{3}{9}$	2	1 0
5	$\frac{2}{9}$	3	1 1 0
6	$\frac{1}{9}$		1 1 1

Найдем среднюю длину кодового слова:

$$L = 2 \cdot \frac{1}{9} + 2 \cdot \frac{2}{9} + 2 \cdot \frac{3}{9} + 3 \cdot \frac{2}{9} + 3 \cdot \frac{1}{9} \approx 2{,}33.$$

Префиксность кода легко проверить, значение L лежит в промежутке (2, 3), поскольку количество кодируемых чисел (5) находится в промежутке $(2^2, 2^3)$.

В данном примере число 2,33 соответствует среднему количеству вопросов, которое придется задать, чтобы отгадать сумму заданных чисел. Максимальное количество вопросов, очевидно, будет равно ближайшему большему целому числа 2,33, т.е. трем.

При этом стратегия отгадывания определяется таблицей кодирования: сначала задается вопрос: «Сумма меньше 4?», при положительном ответе нужно спросить «Сумма меньше 3?», что сразу дает ответ. При отрицательном ответе на подгруппы разбивается группа из трех чисел 4, 5, 6. Если загаданная сумма — 4, на втором вопросе она будет отгадана, в противном случае придется задать еще один вопрос для различения чисел 5 и 6.

Ответ: среднее число вопросов равно 2,33; сумму задуманных чисел можно угадать максимум за 3 вопроса.

При вычислении средней длины кодового слова не следует забывать формулу, по которой она вычисляется:

$$L = \sum_{k=1}^{m} n_k \, p(u_k),$$

где n_k — индивидуальная длина кодового слова для сигнала u_k;

$p(u_k)$ — вероятность сигнала u_k.

Частая ошибка — вычисление L путем суммирования индивидуальных длин кодовых слов n_k и последующего деления на число сигналов m.

Пример 6.3. Сообщение на выходе источника без памяти состоит из букв, принимающих значение A и B с вероятностями 0,8 и 0,2 соответственно. Произвести кодирование по методу Шеннона – Фэно отдельных букв, двух- и трехбуквенных блоков. Сравнить коды по их эффективности.

Решение.

1. Кодируем исходный алфавит.

Кодируемые буквы	Вероятности	разбиения	Кодовые слова
A	0,8	1	0
B	0,2		1

Вычислим энтропию источника:

$$H(A, B) = -0{,}8 \log 0{,}8 - 0{,}2 \log 0{,}2 \approx 0{,}7219 \text{ (бит).}$$

Средняя длина кодового слова

$$L_1 = 1 \cdot 0{,}8 + 1 \cdot 0{,}2 = 1 .$$

Эффективность кода

$$\gamma_1 = \frac{0{,}7219}{1} = 0{,}7219 .$$

2. Формируем блоки из двух букв: *AA, AB, BA, BB*. Вероятность каждого блока находится путем перемножения вероятностей букв, входящих в блок.

Кодируем двухбуквенные блоки:

Кодируемые блоки	Вероятности	Разбиения	Кодовые слова
AA	0,64	1	0
AB	0,16	2	1 0
BA	0,16	3	1 1 0
BB	0,04		1 1 1

Средняя длина кодового слова

$$L = 1 \cdot 0{,}64 + 2 \cdot 0{,}16 + 3 \cdot 0{,}16 + 3 \cdot 0{,}04 = 1{,}56 .$$

Однако следует иметь в виду, что мы считаем среднюю длину кодового слова на одну букву исходного алфавита, поэтому полученную длину L нужно разделить на количество букв в блоке, т.е. на 2:

$$L_2 = \frac{L}{2} = 0{,}78 .$$

Эффективность кода

$$\gamma_2 = \frac{H}{L_2} = \frac{0{,}7219}{0{,}78} \approx 0{,}925 .$$

Типичная ошибка: вычисляется энтропия ансамбля блоков и эффективность считается по отношению к этой энтропии. Это неверно, так как нас интересует эффективность кода по отношению к исходному ансамблю сигналов.

3. Формируем блоки из трех букв: *AAA, AAB, ABA, ABB, BAA, BAB, BBB*.

Кодируем полученные блоки:

Кодируемые блоки	Вероятности	Разбиения	Кодовые слова
AAA	0,512	1	0
AAB	0,128	3	1 0 0
ABA	0,128	2	1 0 1
BAA	0,128	3	1 1 0
ABB	0,032	5	1 1 1 0 0
BAB	0,032	4	1 1 1 0 1
BBA	0,032	5	1 1 1 1 0
BBB	0,008		1 1 1 1 1

Средняя длина кодового слова:

$$L = \frac{1}{3}\left(1 \cdot 0{,}512 + 3 \cdot 3 \cdot 0{,}128 + 5 \cdot 3 \cdot 0{,}032 + 5 \cdot 0{,}008\right) \approx 0{,}728 .$$

Эффективность кода:

$$\gamma_3 = \frac{H}{L_3} = \frac{0{,}7219}{0{,}728} \approx 0{,}992 .$$

Ответ: Эффективность кода возрастает при блочном кодировании с 0,72 (кодирование исходного алфавита) до 0,925 при двухбуквенном и до 0,992 при трехбуквенном кодировании.

Аналогичные результаты получаются при кодировании по методу Хаффмана. Код Хаффмана является оптимальным в том смысле, что он имеет среднюю длину кодового слова меньшую, (или, по крайней мере, равную) чем любой другой код.

Приведем здесь алгоритм кодирования кодом Хаффмана, изложенный в [1].

Код Хаффмана

Кодирование по методу Хаффмана осуществляется следующим образом:

1. s букв алфавита располагаем в порядке убывания их вероятностей.

2. Выбираем целое число m_0 такое, что

$$2 \le m_0 \le m \text{ и } \frac{s - m_0}{m - 1} = a ,$$

где a — целое положительное число.

При кодировании в двоичном канале $m_0 = m = 2$.

Производим первое сжатие алфавита, т.е. группируем вместе m_0 букв источника, имеющих наименьшие вероятности, и обозначаем новой буквой. Вычисляем общую вероятность такого сгруппированного подмножества букв.

3. Буквы нового алфавита, полученного в результате первого сжатия, снова располагаем в порядке убывания вероятностей.

4. Производим второе сжатие этого алфавита, т.е. снова группируем вместе m букв, имеющих наименьшие вероятности, и обозначаем новой буквой. Вычисляем общую вероятность сгруппированного подмножества букв.

5. Буквы нового алфавита, полученного на 4-м шаге, располагаем в порядке убывания вероятностей.

6. Осуществляем последовательные сжатия алфавита путем повторения операций 4 и 5, пока в новом алфавите не останется единственная буква.

7. Проводим линии, которые соединяют буквы, образующие последовательные вспомогательные алфавиты. Получается дерево, в котором отдельные сообщения являются концевыми узлами. Соответствующие им кодовые слова получаем, если приписать различные буквы m-ичного алфавита ветвям, исходящим из каждого промежуточного узла.

ЗАКЛЮЧЕНИЕ

Настоящие методические указания не исчерпывают всего многообразия задач по основам информатики, однако заостряют внимание на основополагающих подходах к практическому применению алгоритмов и методов решения задач, без которых невозможно дальнейшее, более глубокое, освоение прикладных аспектов информатики.

Автор будет благодарен всем пользователям этого пособия за предложения по дополнению перечня разделов, конкретных примеров и типичных трудностей, встречающихся при решении задач.

ЛИТЕРАТУРА

1. Акулиничев Ю.П., Дроздова В.И. Сборник задач по теории информации. Томск. Изд-во ТГУ, 1976 г., 146 с.

2. Орлов В.А., Филиппов Л.И. Теория информации в упражнениях и задачах. М., «Высшая школа», 1976 г., 136 с.

ПРИЛОЖЕНИЯ

Приложение 1

Таблица значений вспомогательной функции $\eta(p) = -p\log_2 p$

p	$h(p)$	p	$h(p)$	p	$h(p)$	p	$h(p)$
0,00	0,0000						
0,01	0,0664	0,26	0,5053	0,51	0,4954	0,76	0,3009
0,02	0,1129	0,27	0,5100	0,52	0,4906	0,77	0,2903
0,03	0,1518	0,28	0,5142	0,53	0,4854	0,78	0,2796
0,04	0,1858	0,29	0,5179	0,54	0,4800	0,79	0,2687
0,05	0,2161	0,30	0,5211	0,55	0,4744	0,80	0,2575
0,06	0,2435	0,31	0,5238	0,56	0,4684	0,81	0,2462
0,07	0,2686	0,32	0,5260	0,57	0,4623	0,82	0,2348
0,08	0,2915	0,33	0,5278	0,58	0,4558	0,83	0,2231
0,09	0,3127	0,34	0,5292	0,59	0,4491	0,84	0,2113
0,10	0,3322	0,35	0,5301	0,60	0,4422	0,85	0,1993
0,11	0,3503	0,36	0,5306	0,61	0,4350	0,86	0,1871
0,12	0,3671	0,37	0,5307	0,62	0,4276	0,87	0,1748
0,13	0,3826	0,38	0,5305	0,63	0,4199	0,88	0,1623
0,14	0,3971	0,39	0,5298	0,64	0,4121	0,89	0,1496
0,15	0,4105	0,40	0,5288	0,65	0,4040	0,90	0,1368
0,16	0,4230	0,41	0,5274	0,66	0,3956	0,91	0,1238
0,17	0,4346	0,42	0,5256	0,67	0,3871	0,92	0,1107
0,18	0,4453	0,43	0,5236	0,68	0,3783	0,93	0,0974
0,19	0,4552	0,44	0,5211	0,69	0,3694	0,94	0,0839
0,20	0,4644	0,45	0,5184	0,70	0,3602	0,95	0,0703
0,21	0,4728	0,46	0,5153	0,71	0,3508	0,96	0,0565
0,22	0,4806	0,47	0,5120	0,72	0,3412	0,97	0,0426
0,23	0,4877	0,48	0,5083	0,73	0,3314	0,98	0,0286
0,24	0,4941	0,49	0,5043	0,74	0,3215	0,99	0,0144
0,25	0,5000	0,50	0,5000	0,75	0,3113	1,00	0,0000

Приложение 2

Значения двоичных логарифмов целых чисел от 1 до 100

n	$\log_2 n$	n	$\log_2 n$	n	$\log_2 n$	n	$\log_2 n$
1	0,00000	26	4,70044	51	5,67243	76	6,24793
2	1,00000	27	4,75489	52	5,70044	77	6,26679
3	1,58496	28	4,80735	53	5,72792	78	6,28540
4	2,00000	29	4,85798	54	5,75489	79	6,30378
5	2,32193	30	4,90689	55	5,78136	80	6,32193
6	2,58496	31	4,95420	56	5,80735	81	6,33985
7	2,80735	32	5,00000	57	5,83289	82	6,35755
8	3,00000	33	5,04439	58	5,85798	83	6,37504
9	3,16993	34	5,08746	59	5,88264	84	6,39232
10	3,32193	35	5,12928	60	5,90689	85	6,40939
11	3,45943	36	5,16993	61	5,93074	86	6,42626
12	3,58496	37	5,20945	62	5,95420	87	6,44294
13	3,70044	38	5,24793	63	5,97728	88	6,45943
14	3,80735	39	5,28540	64	6,00000	89	6,47573
15	3,90689	40	5,32193	65	6,02237	90	6,49185
16	4,00000	41	5,35755	66	6,04439	91	6,50779
17	4,08746	42	5,39232	67	6,06609	92	6,52356
18	4,16993	43	5,42626	68	6,08746	93	6,53916
19	4,24793	44	5,45943	69	6,10852	94	6,55459
20	4,32193	45	5,49185	70	6,12928	95	6,56986
21	4,39232	46	5,52356	71	6,14975	96	6,58496
22	4,45943	47	5,55459	72	6,16993	97	6,59991
23	4,52356	48	5,58496	73	6,18982	98	6,61471
24	4,58496	49	5,61471	74	6,20945	99	6,62936
25	4,64386	50	5,64386	75	6,22882	100	6,64386

$\log_2 10^k = 3{,}32193k.$