§ 2
Подходы к измерению информации

Информация и её свойства являются объектом исследования целого ряда научных дисциплин, таких как:

- теория информации (математическая теория систем передачи информации);
- кибернетика (наука об общих закономерностях процессов управления и передачи информации в машинах, живых организмах и обществе);
- информатика (изучение процессов сбора, преобразования, хранения, защиты, поиска и передачи всех видов информации и средств их автоматизированной обработки);
- семиотика (наука о знаках и знаковых системах);
- теория массовой коммуникации (исследование средств массовой информации и их влияния на общество) и др.

Рассмотрим более детально подходы к определению понятия информации, важные с позиций её измерения:

1) определение К. Шеннона, применяемое в математической теории информации;

2) определение А. Н. Колмогорова, применяемое в отраслях информатики, связанных с использованием компьютеров.

2.1. Содержательный подход к измерению информации

Клод Шеннон, разрабатывая теорию связи, предложил характеризовать информативность сообщения содержащейся в нём полезной информацией, т. е. той частью сообщения, которая снимает полностью или уменьшает существующую до её получения неопределённость какой-либо ситуации.

Клод Элвуд Шеннон (1916–2001) — американский инженер и математик. Является основателем теории информации, нашедшей применение в современных высокотехнологических системах связи. В 1948 году предложил использовать слово «бит» для обозначения наименьшей единицы информации.

Информация — это снятая неопределённость. Величина неопределённости некоторого события — это количество возможных результатов (исходов) данного события.
Сообщение, уменьшающее неопределённость знания в 2 раза, несёт 1 бит информации.
Такой подход к измерению информации называют содержательным.

Пример 1. Допустим, вы подбрасываете монету, загадывая, что выпадет: «орёл» или «решка». Перед подбрасыванием монеты неопределённость знания о результате равна двум. Действительно, есть всего два возможных результата этого события (бросания монеты). Эти результаты мы считаем равновероятными, т. к. ни один из них не имеет преимущества перед другим.

После того как конкретный исход стал известен (например, подброшенная монета упала «орлом» вверх), неопределённость уменьшилась в 2 раза. Таким образом, сообщение о том, что подброшенная монета упала «орлом» вверх, несёт в себе 1 бит информации.

Пример 2. Предположим, в книжном шкафу восемь полок. Книга может быть поставлена на любую из них. Сколько бит

информации несёт сообщение о том, что книга поставлена на третью полку?

Ответ на этот вопрос можно получить, если дополнить исходное сообщение ещё несколькими сообщениями так, чтобы каждое из них уменьшало неопределённость знания в 2 раза.

Итак, количество возможных результатов (исходов) события, состоящего в том, что книга поставлена в шкаф, равно восьми: 1, 2, 3, 4, 5, 6, 7 и 8.

Сообщение «Книга поставлена на полку не выше четвёртой» уменьшает неопределённость знания о результате в два раза. Действительно, после такого сообщения остаётся всего четыре варианта: 1, 2, 3 и 4. Получен один бит информации.

Сообщение «Книга поставлена на полку выше второй» уменьшает неопределённость знания о результате в два раза: после этого сообщения остаётся всего два варианта: 3 и 4. Получен ещё один (второй) бит информации.

Сообщение «Книга поставлена на третью полку» также уменьшает неопределённость знания о результате в два раза. Получен третий бит информации.

Итак, мы построили цепочку сообщений, каждое из которых уменьшало неопределённость знания о результате в два раза, т. е. несло 1 бит информации. Всего было набрано 3 бита информации. Именно столько информации и содержится в сообщении «Книга поставлена на третью полку».

Подумайте, сколько информации содержится в сообщении о том, что книга поставлена на пятую полку. Обоснуйте свой ответ, построив соответствующую цепочку сообщений.

Метод поиска, на каждом шаге которого отбрасывается половина вариантов, называется методом половинного деления. Этот метод широко используется в компьютерных науках.

Пример 3. О результатах футбольного матча между клубами «Спартак» и «Динамо» известно, что больше трёх мячей никто не забил. Всего возможных вариантов счёта матча — 16:

0 : 0	0 : 1	0 : 2	0 : 3
1 : 0	1 : 1	1 : 2	1 : 3
2 : 0	2 : 1	2 : 2	2 : 3
3 : 0	3 : 1	3 : 2	3 : 3

Здесь первая цифра в каждой паре соответствует количеству мячей, забитых командой «Спартак», вторая — командой «Динамо».

Будем считать все варианты равновероятными и отгадывать счёт, задавая вопросы, на которые можно ответить только «да» или «нет». Вопросы будем формулировать так, чтобы количество возможных вариантов счёта каждый раз уменьшалось вдвое. Это позволит нам:

1) обойтись минимальным количеством вопросов;
2) подсчитать, сколько бит информации содержит сообщение о счёте матча.

Вопрос 1. «Спартак» забил больше одного мяча? Предположим, получен ответ «Нет». Такой ответ позволяет не рассматривать варианты, расположенные в нижней части таблицы, т. е. сокращает количество возможных исходов в 2 раза:

0 : 0	0 : 1	0 : 2	0 : 3
1 : 0	1 : 1	1 : 2	1 : 3
~~2 : 0~~	~~2 : 1~~	~~2 : 2~~	~~2 : 3~~
~~3 : 0~~	~~3 : 1~~	~~3 : 2~~	~~3 : 3~~

Вопрос 2. «Спартак» забил один мяч? Предположим, получен ответ «Да». Такой ответ позволяет не рассматривать варианты, расположенные в верхней строке таблицы, т. е. сокращает количество возможных исходов ещё в 2 раза:

~~0 : 0~~	~~0 : 1~~	~~0 : 2~~	~~0 : 3~~
1 : 0	1 : 1	1 : 2	1 : 3

Вопрос 3. «Спартак» пропустил больше одного мяча? Предположим, получен ответ «Нет». Можно отбросить ещё два варианта:

1 : 0	1 : 1	~~1 : 2~~	~~1 : 3~~

Вопрос 4. «Спартак» пропустил один мяч? Предположим, получен ответ «Да». Получаем единственный вариант:

~~1 : 0~~	1 : 1

Итак, нам удалось выяснить счёт матча, задав четыре вопроса, ответ на каждый из которых уменьшал неопределённость результата в два раза, т. е. нёс 1 бит информации. Сообщение о счёте матча несёт четыре бита информации.

Выясните, какому счёту матча будут соответствовать следующие цепочки ответов на поставленные выше вопросы:

1) Да – Да – Да – Да;
2) Нет – Нет – Нет – Нет;
3) Да – Нет – Да – Нет.

Попробуйте придумать такие вопросы, чтобы цепочка ответов Нет – Да – Нет – Да приводила к счёту 2 : 3.

Вычислять количество информации, содержащееся в сообщении о том, что имел место один из множества равновероятных результатов некоторого события, с помощью метода половинного деления возможно, но затруднительно. Гораздо проще воспользоваться следующей закономерностью.

Количество информации i, содержащееся в сообщении об одном из N равновероятных результатов некоторого события, определяется из формулы:

$$2^i = N.$$

При N, равном целой степени двойки (2, 4, 8, 16, 32 и т. д.), это уравнение легко решается в уме. Решать такие уравнения при других N вы научитесь чуть позже, в курсе математики 11 класса.

Пример 4. Петя и Вася заинтересовались игрой «Крестики-нолики» на поле $n \times n$. Количество информации, полученное вторым игроком после первого хода первого игрока, составляет 6 бит. Требуется выяснить размеры поля, на котором играют Петя и Вася.

Дано:

$i = 6$	$2^i = N$	$2^6 = 64$
n — ?	$n \times n = N$	$64 = 8 \times 8$

Ответ: 8×8.

2.2. Алфавитный подход к измерению информации

Определение количества информации на основе уменьшения неопределённости наших знаний рассматривает информацию с точки зрения её содержания, понятности и новизны для чело-

века. С этой точки зрения в примере о подбрасывании монеты одинаковое количество информации содержит и зрительный образ упавшей монеты, и короткое сообщение «Орёл», и длинная фраза «В результате подбрасывания монета упала так, что на её видимой части изображён орёл».

Однако при хранении и передаче информации с помощью технических устройств целесообразно отвлечься от её содержания и рассматривать информацию как последовательность символов (букв, цифр, кодов цвета точек изображения и т. д.) некоторого алфавита.

Информация — последовательность символов (букв, цифр, кодов цвета точек изображения и т. д.) некоторого алфавита.

Минимальная мощность алфавита (количество входящих в него символов), пригодного для кодирования информации, равна 2. Такой алфавит называется двоичным. Один символ двоичного алфавита несёт 1 бит информации.

Согласно Колмогорову, количество информации, содержащейся в последовательности символов, определяется минимально возможным количеством двоичных знаков, необходимых для кодирования этой последовательности, безотносительно к содержанию представленного ею сообщения. Данный подход к определению количества информации называют алфавитным.

Андрей Николаевич Колмогоров (1903–1987) — один из крупнейших математиков XX века. Им получены основополагающие результаты в математической логике, теории сложности алгоритмов, теории информации, теории множеств и ряде других областей математики и её приложений.

Информационным объёмом сообщения называется количество двоичных символов, которое используется для кодирования этого сообщения. В двоичном коде один двоичный разряд несёт 1 бит информации.

В отличие от определения количества информации по Колмогорову в определении информационного объёма не требуется,

чтобы число двоичных символов было минимально возможным. При оптимальном кодировании понятия количества информации и информационного объёма совпадают.

Из курса информатики основной школы вы знаете, что двоичные коды бывают равномерные и неравномерные. Равномерные коды в кодовых комбинациях содержат одинаковое число символов, неравномерные — разное.

Первый равномерный двоичный код был изобретён французом Жаном Морисом Бодо в 1870 году. В коде Бодо используются сигналы двух видов, имеющие одинаковую длительность и абсолютную величину, но разную полярность. Длина кодов всех символов алфавита равна пяти (рис. 1.7).

Таблица букв							
A	..o..	B	.o..o	K	oo o..	S	o...o
É	..oo.	C	.o.o.o	L	oo oo.	T	o.o.o
E	...o.	D	.o ooo	M	oo .o.	V	o. ooo
I	...oo	F	.o .oo	N	oo .oo	W	o..oo
O	.. ooo	G	.o .o.	P	oo ooo	X	o..o.
U	..o.o	H	.o oo.	Q	oo o.o	Z	o.oo.
Y	....o	J	.o o..	R	oo ..o	—	o.o..

Рис. 1.7. Фрагмент кодовой таблицы кода Бодо

Всего с помощью кода Бодо можно составить $2^5 = 32$ комбинации.

Пример 5. Слово WORD, закодированное с помощью кода Бодо, будет выглядеть так:

o	.	.	o	o	.	.	o	o	o	o	o	.	.	o	.	o	o	o	o

Информационный объём такого сообщения равен 20 битам; таково количество двоичных символов, которое используется для кодирования этого сообщения.

Пример 6. Для двоичного представления текстов в компьютере чаще всего используется равномерный восьмиразрядный код. С его помощью можно закодировать алфавит из 256 символов ($2^8 = 256$). Фрагмент кодовой таблицы ASCII представлен на рисунке 1.8.

65	A	01000001	78	N	01001110
66	B	01000010	79	O	01001111
67	C	01000011	80	P	01010000
68	D	01000100	81	Q	01010001
69	E	01000101	82	R	01010010
70	F	01000110	83	S	01010011
71	G	01000111	84	T	01010100
72	H	01001000	85	U	01010101
73	I	01001001	86	V	01010110
74	J	01001010	87	W	01010111
75	K	01001011	88	X	01011000
76	L	01001100	89	Y	01011001
77	M	01001101	90	Z	01011010

Рис. 1.8. Фрагмент кодовой таблицы ASCII

Слово WORD, закодированное с помощью таблицы ASCII:

0	1	0	1	0	1	1	1	0	1	0	0	1	1	1	1	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0

Информационный объём такого сообщения равен 32 битам.

Из курса информатики основной школы вам известно, что с помощью i-разрядного двоичного кода можно закодировать алфавит, мощность N которого определяется из соотношения:

$$2^i = N.$$

Иными словами, зная мощность используемого алфавита, всегда можно вычислить информационный вес символа — минимально возможное количество бит, требуемое для кодирования символов этого алфавита. При этом информационный вес символа должен быть выражен целым числом.

Соотношение для определения информационного веса символа алфавита можно получить и из следующих соображений.

Пусть N — мощность алфавита, используемого для кодирования сообщений. При этом в каждом конкретном сообщении какие-то символы алфавита будут использоваться чаще, какие-то — реже, а какие-то — не будут использоваться вообще. Не станем обращать на это внимание, наоборот, будем считать, что любой из символов может появиться в сообщении с равной вероятностью. Другими словами, появление в сообщении очередного символа — одно из N равновероятных событий. Но количество информации i, содержащееся в сообщении об одном из N равновероятных результатов некоторого события, определяется из формулы $2^i = N$.

Алгоритм вычисления информационного объёма сообщения выглядит так:

1) определить мощность используемого алфавита N;
2) из соотношения $2^i = N$ определить i — информационный вес символа алфавита в битах (длину двоичного кода символа из используемого алфавита мощности N);
3) вычислить информационный объём сообщения I, умножив информационный вес символа i на количество символов в сообщении K.

При алфавитном подходе информационный объём сообщения I, состоящего из K символов, вычисляется по формуле:

$$I = K \cdot i,$$

где i — информационный вес символа в битах, связанный с мощностью используемого алфавита N соотношением:

$$2^i = N.$$

Пример 7. Для регистрации на некотором сайте пользователю надо придумать пароль, состоящий из 10 символов. В качестве символов можно использовать десятичные цифры и шесть первых букв латинского алфавита, причём буквы используются только заглавные. Пароли кодируются посимвольно. Все символы кодируются одинаковым и минимально возможным количеством бит. Для хранения сведений о каждом пользователе в системе отведено одинаковое и минимально возможное целое число байт.

Необходимо выяснить, какой объём памяти потребуется для хранения 100 паролей.

Дано:

$N = 10 + 6 = 16$	$I_{100} = 100 \cdot I$	$16 = 2^i$, $i = 4$ (бита на символ)
$K = 10$	$I = K \cdot i$	$I = 10 \cdot 4 = 40$ (бит) $= 5$ (байт)
I_{100} — ?	$N = 2^i$	$I_{100} = 100 \cdot 5 = 500$ (байт)

Ответ: 500 байт.

2.3. Единицы измерения информации

Итак, в двоичном коде один двоичный разряд несёт 1 бит информации. 8 бит образуют один байт. Помимо бита и байта, для измерения информации используются более крупные единицы:

1 Кбайт (килобайт) $= 2^{10}$ байт;
1 Мбайт (мегабайт) $= 2^{10}$ Кбайт $= 2^{20}$ байт;
1 Гбайт (гигабайт) $= 2^{10}$ Мбайт $= 2^{20}$ Кбайт $= 2^{30}$ байт;
1 Тбайт (терабайт) $= 2^{10}$ Гбайт $= 2^{20}$ Мбайт $= 2^{30}$ Кбайт $= 2^{40}$ байт;
1 Пбайт (петабайт) $= 2^{10}$ Тбайт $= 2^{20}$ Гбайт $= 2^{30}$ Мбайт $= 2^{40}$ Кбайт $= 2^{50}$ байт.

Исторически сложилось так, что приставки «кило», «мега», «гига», «тера» и др. в информатике трактуются не так, как в математике, где «кило» соответствует 10^3, «мега» — 10^6, «гига» — 10^9, «тера» — 10^{12} и т. д.

Это произошло потому, что $2^{10} = 1024 \approx 1000 = 10^3$. Поэтому 1024 байта и стали называть килобайтом, 2^{10} килобайта стали называть мегабайтом и т. д.

Чтобы избежать путаницы с различным использованием одних и тех же приставок, в 1999 г. Международная электротехническая комиссия ввела новый стандарт наименования двоичных приставок. Согласно этому стандарту, 1 килобайт равняется 1000 байт, а величина 1024 байта получила новое название — 1 кибибайт (Кибайт).

У нас в стране в 2009 году принято «Положение о единицах величин, допускаемых к применению в Российской Федерации». В нём сказано, что наименование и обозначение единицы количества информации «байт» (1 байт = 8 бит) применяются с двоичными приставками «кило», «мега», «гига», которые соответствуют множителям «2^{10}», «2^{20}» и «2^{30}» (1 Кбайт = 1024 байт, 1 Мбайт = 1024 Кбайт, 1 Гбайт = 1024 Мбайт). Данные приставки пишутся с большой буквы.

Рассмотрим ещё несколько примеров решения задач, связанных с определением информационного объёма сообщений.

Пример 8. При регистрации в компьютерной системе каждому пользователю выдаётся пароль длиной в 12 символов, образованный из десятичных цифр и первых шести букв английского алфавита, причём буквы могут использоваться как строчные, так и прописные — соответствующие символы считаются разными. Пароли кодируются посимвольно. Все символы кодируются одинаковым и минимально возможным количеством бит. Для хранения сведений о каждом пользователе в системе отведено одинаковое и минимально возможное целое число байт. Кроме собственно пароля для каждого пользователя в системе хранятся дополнительные сведения, для которых отведено 12 байт. На какое максимальное количество пользователей рассчитана система, если для хранения сведений о пользователях в ней отведено 200 Кбайт?

Прежде всего, выясним мощность алфавита, используемого для записи паролей: $N = 6$ (буквы прописные) + 6 (буквы строчные) + 10 (десятичные цифры) = 22 символа.

Для кодирования одного из 22 символов требуется 5 бит памяти (4 бита позволят закодировать всего $2^4 = 16$ символов, 5 бит позволят закодировать уже $2^5 = 32$ символа); 5 — минимально возможное количество бит для кодирования 22 разных символов алфавита, используемого для записи паролей.

Для хранения всех 12 символов пароля требуется $12 \cdot 5 = 60$ бит. Из условия следует, что пароль должен занимать целое число байт; т. к. 60 не кратно восьми, возьмём ближайшее большее значение, которое кратно восьми: $64 = 8 \cdot 8$. Таким образом, один пароль занимает 8 байт.

Информация о пользователе занимает 20 байт, т. к. содержит не только пароль (8 байт), но и дополнительные сведения (12 байт).

Максимальное количество пользователей («польз.»), информацию о которых можно сохранить в системе, равно 10 240:

$$\frac{200\ \text{Кбайт}}{20\ \text{байт/польз.}} = \frac{200 \cdot 1024\ \text{байт}}{20\ \text{байт/польз.}} = 10240\ \text{польз.}$$

Пример 9. Объём сообщения, состоящего из 8192 символов, равен 16 Кбайт. Какова максимальная мощность алфавита, использованного при передаче сообщения?

Дано:

$I = 16$ Кбайт $= 16 \cdot 2^{13}$ бит $K = 8192$ символа N — ?	$I = K \cdot i$, $i = I/K$ $N = 2^i$	$i = 16 \cdot 2^{13}/8192 = 16$ (бит) $N = 2^{16} = 65\,536$ (символов)

Ответ: максимальная мощность алфавита — 65 536 символов.

САМОЕ ГЛАВНОЕ

Информация (по Шеннону) — это снятая неопределённость. Величина неопределённости некоторого события — это количество возможных результатов (исходов) данного события. Сообщение, уменьшающее неопределённость знания в 2 раза, несёт 1 бит информации. Количество информации i, содержащееся в сообщении об одном из N равновероятных результатов некоторого события, определяется из формулы: $2^i = N$. Такой подход к измерению информации называют содержательным.

Информация (по Колмогорову) — последовательность символов (букв, цифр, кодов цвета точек изображения и т. д.) некоторого алфавита. Информационным объёмом сообщения называется количество двоичных символов, которое используется для кодирования этого сообщения. В двоичном коде один двоичный разряд несёт 1 бит информации. Такой подход к измерению информации называют алфавитным.

При алфавитном подходе информационный объём сообщения I, состоящего из K символов, вычисляется по формуле:

$$I = K \cdot i,$$

где i — информационный вес символа в битах, связанный с мощностью используемого алфавита N соотношением $2^i = N$.

Единицы измерения информации:

1 байт = 8 бит;
1 Кбайт (килобайт) $= 2^{10}$ байт;
1 Мбайт (мегабайт) $= 2^{10}$ Кбайт $= 2^{20}$ байт;
1 Гбайт (гигабайт) $= 2^{10}$ Мбайт $= 2^{20}$ Кбайт $= 2^{30}$ байт;
1 Тбайт (терабайт) $= 2^{10}$ Гбайт $= 2^{20}$ Мбайт $= 2^{30}$ Кбайт $= 2^{40}$ байт;
1 Пбайт (петабайт) $= 2^{10}$ Тбайт $= 2^{20}$ Гбайт $= 2^{30}$ Мбайт $= 2^{40}$ Кбайт $= 2^{50}$ байт.

Вопросы и задания

1. Что такое неопределённость знания о результате какого-либо события? Приведите пример.
2. В чём состоит суть содержательного подхода к определению количества информации? Что такое бит с точки зрения содержательного подхода?
3. Паролем для приложения служит трёхзначное число в шестнадцатеричной системе счисления. Возможные варианты пароля:

189	101	654	FFE	123
A41	880	391	110	125

Ответ на какой вопрос (см. ниже) содержит 1 бит информации?

1) Это число записано в двоичной системе счисления?
2) Это число записано в четверичной системе счисления?
3) Это число может быть записано в восьмеричной системе счисления?
4) Это число может быть записано в десятичной системе счисления?
5) Это число может быть записано в шестнадцатеричной системе счисления?

4. При угадывании целого числа в некотором диапазоне было получено 5 бит информации. Каковы наибольшее и наименьшее числа этого диапазона?
5. Какое максимальное количество вопросов достаточно задать вашему собеседнику, чтобы точно определить день и месяц его рождения?
6. В чём состоит суть алфавитного подхода к измерению информации? Что такое бит с точки зрения алфавитного подхода?
7. Закодируйте фразу «ALL IN GOOD TIME» кодом Бодо и восьмиразрядным компьютерным кодом. Сравните полученные информационные объёмы текста.
8. Какие единицы используются для измерения объёма информации, хранящейся на компьютере?
9. Объём сообщения, содержащего 11 264 символа, равен 11 Кбайт. Определите максимальную мощность алфавита, который мог быть использован для кодирования этого сооб-

щения? Какова минимальная мощность алфавита, использование которого привело к такому же информационному объёму закодированного сообщения?

10. В школе 750 учащихся, коды учащихся записаны в школьной информационной системе с помощью минимального количества бит. Каков информационный объём в байтах сообщения о кодах 180 учащихся начальных классов?
11. В школьной базе данных каждый ученик получил идентификатор, состоящий ровно из 6 символов. В качестве символов используются все заглавные буквы русского алфавита, кроме «Ё», «Ы», «Ъ» и «Ь», а также все десятичные цифры за исключением цифры 0. Каждый такой идентификатор в информационной системе записывается минимально возможным и одинаковым целым количеством байт (при этом используют посимвольное кодирование и все символы кодируются одинаковым и минимально возможным количеством бит). Определите объём памяти, необходимый для хранения в этой системе 180 идентификаторов учащихся начальных классов. Ответ выразите в килобайтах.
12. В ходе телевизионного шоу проводится СМС-голосование: каждый телезритель отдаёт свой голос за одного из 12 артистов-участников шоу, отправляя сообщение с его номером. Голос каждого телезрителя, отданный за того или иного участника, кодируется одинаковым и минимально возможным количеством бит и сохраняется для подведения итогов. За время телевизионного шоу в голосовании приняли участие 163 840 зрителей. Определите объём сохранённой информации о голосовании и выразите его в килобайтах.
13. При регистрации в компьютерной системе каждому пользователю выдаётся пароль, состоящий из 6 символов и содержащий только символы из шестибуквенного набора A, B, C, D, E, F. Для хранения сведений о каждом пользователе отведено одинаковое и минимально возможное целое число байт. При этом используют посимвольное кодирование паролей и все символы кодируются одинаковым и минимально возможным количеством бит. Кроме собственно пароля для каждого пользователя в системе хранятся дополнительные сведения, занимающие 15 байт. Определите объём памяти в байтах, необходимый для хранения сведений о 120 пользователях.