§3 Измерение информации. Объемный подход

Вопрос об измерении количества информации является очень важным как для науки, так и для практики. В самом деле, информация является предметом нашей деятельности: мы ее храним, передаем, принимаем, обрабатываем. Поэтому важно договориться о способе ее измерения, позволяющем, например, ответить на вопросы: достаточно ли места на носителе, чтобы разместить нужную нам информацию, или сколько времени потребуется, чтобы передать ее по имеющемуся каналу связи. Величина, которая нас в этих ситуациях интересует, называется **объемом информации**. В таком случае говорят об **объемном подходе** к измерению информации.

Как измерить объем информации

Объем информации не связан с ее содержанием.
Когда говорят об объеме информации, то имеют в виду размер текста в том алфавите, с помощью которого эта информация представлена.

Объем текста в печатном издании — книге, журнале, газете — обычно измеряют в страницах. В таком случае мы считаем, что, например, книга в 500 страниц содержит больше информации, чем книга в 250 страниц.

А как вы думаете, всегда ли книга в 500 страниц содержит в два раза больше информации, чем книга в 250 страниц? Конечно, нет! Ведь разные книги могут иметь разные форматы листов. Формат листа — это его стандартный размер. Существуют различные форматы печатного листа. Кроме того, разными бывают величина шрифта, длина строки, межстрочное расстояние. Очень часто детские книги печатаются крупным шрифтом с большими интервалами между строками, да еще и с большим количеством картинок. Зачастую содержание такой книги, состоящей из нескольких десятков страниц, можно перепечатать мелким шрифтом на 2–3 страницах. Но поскольку текст в обоих случаях один и тот же, то и количество информации должно быть одинаковым. Значит, измерение информации в страницах не является универсальным.

Количество страниц в печатном издании определяет расход бумаги, а не объем информации. Разумнее измерять объем информации, заключенный в тексте, количеством знаков этого текста. Знаки — это прежде всего буквы из алфавита того языка, на котором написана книга. Но в текст входят также и знаки препинания, скобки, цифры. В тексте могут использоваться буквы других алфавитов, например английского и греческого. Пробел между словами — тоже знак. Например, на странице формата А4 (21 см × 30 см) при размере шрифта (кегле), равном 12 пунктам (стандартным единицам), и одинарном интервале между строками помещается текст объемом примерно 4000 знаков.

Определением объема информации в знаках часто пользуются в издательской практике. Например, журналисту может быть дано ограничение на объем статьи в 40 000 знаков.

Объемный способ измерения информации называют еще **алфавитным подходом.**

Объем информации в электронном сообщении

Выше мы говорили о телеграфном коде Бодо. В нем каждая буква алфавита кодируется пятизначным двоичным кодом. В русском алфавите 32 буквы (не считая буквы ё). Из базового курса информатики вы знаете, что если с помощью i-разрядного двоичного кода можно закодировать алфавит, состоящий из N символов, то эти величины связаны между собой по формуле:

$$2^i = N.$$

Поскольку $2^5 = 32$, то все русские буквы можно закодировать всевозможными пятиразрядными двоичными кодами от 00000 до 11111. Рус-

ский телеграфный аппарат содержит 32 клавиши с буквами. Буква «ё» пропускается, вместо нее имеется более нужная клавиша «пробел». Знаки препинания передаются словами: «зпт», «тчк». Таким образом, телеграфный аппарат при вводе переводит русский текст в двоичный код, количество символов в котором в 5 раз больше, чем в исходном тексте.

Весь английский алфавит, состоящий из 26 букв, также можно закодировать пятиразрядным двоичным кодом. В отличие от русского алфавита, остается еще 6 свободных кодов, которые можно использовать для пробела и пяти знаков препинания.

Из базового курса информатики вам известно, что в компьютерах используется двоичное кодирование информации. Для двоичного представления текстов в компьютере чаще всего используется равномерный восьмиразрядный код. С его помощью можно закодировать алфавит из 256 символов, поскольку $256 = 2^8$. В стандартную кодовую таблицу (например, ASCII) помещаются все необходимые символы: английские и русские прописные и строчные буквы, цифры, знаки препинания, знаки арифметических операций, всевозможные скобки и пр.

В двоичном коде один двоичный разряд несет одну единицу информации, которая называется 1 бит.

При двоичном кодировании объем информации, выраженный в битах, равен длине двоичного кода, в котором информация представлена.

Более крупной единицей измерения информации является байт: 1 байт = 8 битов.

Информационный объем текста в памяти компьютера измеряется в байтах. Он равен количеству знаков в записи текста.

Одна страница текста на листе формата А4 кегля 12 с одинарным интервалом между строками (см. пример выше) в компьютерном представлении будет иметь объем примерно 4000 байтов, так как на ней помещается примерно 4000 знаков.

Помимо бита и байта, для измерения информации используются и более крупные единицы:

1 Кб (килобайт) = 2^{10} байт = 1024 байта;
1 Мб (мегабайт) = 2^{10} Кб = 1024 Кб;
1 Гб (гигабайт) = 2^{10} Мб = 1024 Мб.

Объем той же страницы текста будет равным приблизительно 3,9 Кб. А книга из 500 таких страниц займет в памяти компьютера примерно 1,9 Мб.

Система основных понятий

<table>
<tr><th colspan="5">Измерение информации — объемный подход</th></tr>
<tr><td>На бумажных носителях</td><td colspan="4">На цифровых носителях и в технических системах передачи информации</td></tr>
<tr><td rowspan="4">Объем текста измеряется в знаках</td><td colspan="4">Объем информации равен длине двоичного кода
Основная единица:
1 бит — один разряд двоичного кода</td></tr>
<tr><td colspan="2">Длина кода символа (i бит) кодируемого алфавита мощностью N символов:
$2^i = N$</td><td colspan="2">Информационный объем текста (I), содержащего K символов:
$I = K \cdot i$</td></tr>
<tr><td colspan="4">Производные единицы</td></tr>
<tr><td>Байт
1 байт = 8 бит</td><td>Килобайт (Кб)
1 Кб = 1024 байт</td><td>Мегабайт (Мб)
1 Мб = 1024 Кб</td><td>Гигабайт (Гб)
1 Гб = 1024 Мб</td></tr>
</table>

Вопросы и задания

1. Есть ли связь между объемным подходом к измерению информации и содержанием информации?
2. В чем измеряется объем письменного или печатного текста?
3. Оцените объем одной страницы данного учебника в количестве знаков.
4. Что такое бит с позиции объемного подхода к измерению информации?
5. Какой информационный вес имеет каждая буква русского алфавита?
6. Чем удобнее английский алфавит по сравнению с русским для передачи сообщений с помощью телеграфного кода Бодо?
7. Какие единицы используются для измерения объема информации на компьютерных носителях?
8. Возьмите страницу текста из данного учебника и подсчитайте получаемые информационные объемы текста при кодировании его кодом Морзе, кодом Бодо и восьмиразрядным компьютерным кодом.
9. Результат ответа на задание 3 пересчитайте в килобайтах и мегабайтах.

§4 Измерение информации. Содержательный подход

В предыдущем параграфе рассмотрен объемный подход к измерению информации. Он используется для определения количества информации, заключенного в тексте, записанном с помощью некоторого алфавита. При этом содержательная сторона текста в учет не берется. Совершенно бессмысленное сочетание символов с данной позиции имеет ненулевой информационный объем.

Неопределенность знания и количество информации

Сейчас мы обсудим другой подход к измерению информации, который называют **содержательным подходом**. В этом случае количество информации связывается с содержанием (смыслом) полученного человеком сообщения. Вспомним, что с «человеческой» точки зрения информация — это знания, которые мы получаем из внешнего мира. Количество информации, заключенное в сообщении, должно быть тем больше, чем больше оно пополняет наши знания.

Как же с этой точки зрения определяется единица измерения информации? Вы уже знаете, что эта единица называется битом. Проблема измерения информации исследована в *теории информации*, основатель которой — Клод Шеннон. В теории информации для бита дается следующее определение:

> Сообщение, уменьшающее неопределенность знания в два раза, несет 1 бит информации.

В этом определении есть понятия, которые требуют пояснения. Что такое неопределенность знания? Поясним на примерах. Допустим, вы бросаете монету, загадывая, что выпадет: орел или решка. Есть всего два возможных результата бросания монеты. Причем ни один из этих результатов не имеет преимущества перед другим. В таком случае говорят, что они *равновероятны**.

В случае с монетой перед ее подбрасыванием неопределенность знания о результате равна двум. Игральный же кубик с шестью гранями может с равной вероятностью упасть на любую из них. Значит, неопределенность знания о результате бросания кубика равна шести. Еще пример: спортсмены-лыжники перед забегом путем жеребьевки определяют свои порядковые номера на старте. Допустим, что имеется 100 участников соревнований, тогда неопределенность знания спортсмена о своем номере до жеребьевки равна 100.

* Более строгое определение равновероятности: если увеличивать количество бросаний монеты (100, 1000, 10000 и т. д.), то число выпадений орла и число выпадений решки будут все ближе к половине количества бросаний монеты.

Следовательно, можно сказать так:

Неопределенность знания о результате некоторого события (бросание монеты или игрального кубика, вытаскивание жребия и др.) — это количество возможных результатов.

Вернемся к примеру с монетой. После того как вы бросили монету и посмотрели на нее, вы получили зрительное сообщение, что выпал, например, орел. Определился один из двух возможных результатов. Неопределенность знания уменьшилась в два раза: было два варианта, остался один. Значит, *узнав результат бросания монеты, вы получили 1 бит информации.*

Сообщение об одном из двух равновероятных результатов некоторого события несет 1 бит информации.

Это утверждение — частный вывод из определения, данного выше.

А теперь такая задача: студент на экзамене может получить одну из четырех оценок: 5— «отлично», 4 — «хорошо», 3 — «удовлетворительно», 2 — «неудовлетворительно». Представьте себе, что ваш товарищ пошел сдавать экзамен. Причем учится он очень неровно и может с одинаковой вероятностью получить любую оценку от «2» до «5». Вы волнуетесь за него, ждете результата экзамена. Наконец, он пришел и на ваш вопрос: «Ну, что получил?» — ответил: «Четверку!».

Вопрос: сколько битов информации содержится в его ответе?

Если сразу сложно ответить на этот вопрос, то давайте подойдем к ответу постепенно. Будем отгадывать оценку, задавая вопросы, на которые можно ответить только «да» или «нет».

Вопросы будем ставить так, чтобы каждый ответ уменьшал количество возможных результатов в два раза и, следовательно, приносил 1 бит информации.

Первый вопрос:

— Оценка выше «тройки»?

— Да.

После этого ответа число вариантов уменьшилось в два раза. Остались только «4» и «5». Получен 1 бит информации.

Второй вопрос:

— Ты получил «пятерку»?

— Нет.

Выбран один вариант из двух оставшихся: оценка — «четверка». Получен еще 1 бит информации. В сумме имеем 2 бита.

Сообщение об одном из четырех равновероятных результатов некоторого события несет 2 бита информации.

Разберем еще одну частную задачу, а потом получим общее правило.

На книжном стеллаже восемь полок. Книга может быть поставлена на любую из них. Сколько информации содержит сообщение о том, где находится книга?

Будем действовать таким же способом, как в предыдущей задаче. Метод поиска, на каждом шаге которого отбрасывается половина вариантов,

называется *методом половинного деления*. Применим метод половинного деления к задаче со стеллажом.

Задаем вопросы:

— Книга лежит выше четвертой полки?

— Да.

— Книга лежит выше шестой полки?

— Нет.

— Книга — на шестой полке?

— Нет.

— Ну теперь все ясно! Книга лежит на пятой полке!

Каждый ответ уменьшал неопределенность в два раза. Всего было задано три вопроса. Значит, набрано 3 бита информации. И если бы сразу было сказано, что книга лежит на пятой полке, то этим сообщением были бы переданы те же 3 бита информации.

Заметим, что поиск значения методом половинного деления наиболее рационален. Таким способом всегда можно угадать любой из восьми вариантов за три вопроса. Если бы, например, поиск производился последовательным перебором: «Книга на первой полке?» — «Нет». — «На второй полке?» — «Нет» и т. д., то про пятую полку мы бы узнали после пяти вопросов, а про восьмую — после восьми.

Главная формула информатики

А сейчас попробуем получить формулу, по которой вычисляется количество информации, содержащейся в сообщении о том, что имел место один из множества равновероятных результатов некоторого события.

Обозначим буквой N количество возможных результатов события, или, как мы это еще называли, — неопределенность знания. Буквой i будем обозначать количество информации в сообщении об одном из N результатов.

В примере с монетой: $N = 2$, $i = 1$ бит.
В примере с оценками: $N = 4$, $i = 2$ бита.
В примере со стеллажом: $N = 8$, $i = 3$ бита.

Нетрудно заметить, что связь между этими величинами выражается следующей формулой:

$$2^i = N.$$

Действительно: $2^1 = 2$; $2^2 = 4$; $2^3 = 8$.

С полученной формулой вы уже знакомы из базового курса информатики, и еще не однажды мы с ней встретимся. Значение этой формулы столь велико, что мы назвали ее **главной формулой информатики**. Если величина N известна, а i неизвестно, то данная формула становится уравнением для определения i. В математике оно называется *показательным уравнением*.

Пусть на стеллаже не 8, а 16 полок. Чтобы ответить на вопрос, сколько информации содержится в сообщении о месте нахождения книги, нужно решить уравнение:

$$2^i = 16.$$

Поскольку $16 = 2^4$, то $i = 4$ бита.

Количество информации (i), содержащееся в сообщении об одном из N равновероятных результатов некоторого событий, определяется из решения показательного уравнения: $2^i = N$.

Если значение N равно целой степени двойки (4, 8, 16, 32, 64 и т. д.), то показательное уравнение легко решить в уме, поскольку i будет целым числом. А чему, например, равно количество информации в сообщении о результате бросания игральной кости, у которой имеется шесть граней и, следовательно, $N = 6$? Можно догадаться, что решение уравнения

$$2^i = 6$$

будет дробным числом, лежащим между 2 и 3, поскольку $2^2 = 4 < 6$, а $2^3 = 8 > 6$. А как точнее узнать это число?

Пока ваших математических знаний недостаточно для того, чтобы решить это уравнение. Вы научитесь этому в 11-м классе в курсе математики. А сейчас сообщим, что результатом решения уравнения для $N = 6$ будет значение $i = 2{,}58496$ бита с точностью до пяти знаков после запятой.

Система основных понятий

<table>
<tr><th colspan="2">Измерение информации — содержательный подход</th></tr>
<tr><td colspan="2">Измеряется количество информации в сообщении о результате некоторого события</td></tr>
<tr><td colspan="2">Равновероятные результаты:
никакой результат не имеет преимущества перед другими</td></tr>
<tr><td>Неопределенность знания —
число возможных результатов (вариантов сообщения) — N</td><td>Количество информации
в сообщении об одном результате события — i битов</td></tr>
<tr><td colspan="2">Главная формула информатики: $2^i = N$</td></tr>
<tr><td colspan="2">Частный случай: два равновероятных результата события</td></tr>
<tr><td>$N = 2$</td><td>$i = 1$ бит</td></tr>
<tr><td colspan="2">1 бит — количество информации в сообщении об одном из двух равновероятных результатов некоторого события</td></tr>
</table>

Вопросы и задания

1. Что такое неопределенность знания о результате какого-либо события? Приведите примеры, когда неопределенность знания можно выразить количественно.
2. Как определяется единица измерения количества информации?
3. В каких случаях и по какой формуле можно вычислить количество информации, содержащейся в сообщении, используя содержательный подход?
4. Сколько битов информации несет сообщение о том, что из колоды в 32 карты достали «даму пик»?
5. Проводятся две лотереи: «4 из 32» и «5 из 64». Сообщение о результатах какой из лотерей несет больше информации и во сколько раз?