Кодирование информации

План

1 Кодирование и декодирование данных

Для того чтобы хранить, обрабатывать, передавать информацию, ее необходимо как-то зафиксировать. Например, записать с помощью символов какого-либо языка.

Кодирование – это процесс перевода информации с одного языка на другой (запись в другой системе символов, в другом алфавите). Т.е. слово «кодирование» понимается не в узком смысле – кодирование как способ сделать сообщение непонятным для всех, кто не владеет ключом кода, а в широком – как представление информации в виде сообщения на каком-либо языке.

Обычно кодированием называют перевод информации с естественного языка на формальный, например, в двоичный код, а **декодированием** – обратный процесс, т.е. процесс восстановления информационного сообщения из некоторой последовательности кодов.

Один символ исходного сообщения может заменяться одним символом нового кода или несколькими символами, а может быть и наоборот – несколько символов исходного сообщения заменяются одним символом в новом коде (китайские иероглифы обозначают целые слова и понятия). Иногда при кодировании и декодировании происходит искажение сообщения. Например, известно, что перевод художественных текстов на другой язык и затем обратный перевод могут изменить их до неузнаваемости.

Кодирование может быть равномерное и неравномерное. При равномерном кодировании – все символы кодируются кодами равной длины. При неравномерном кодировании – разные символы могут кодироваться кодами разной длины, что затрудняет декодирование.

2 Кодирование символьных данных

В современных компьютерах все виды информации представлены в двоичном коде. Поэтому каждому используемому символу как-то сопоставляется цепочка нулей и единиц. Например, составляется таблица «символ – код»:

1) определяется, сколько символов нужно использовать (N);

2) определяется нужное количество k двоичных разрядов так, чтобы с их помощью можно было закодировать не менее N разных последовательностей (то есть $2^k \geq N$);

3) составляется таблица, в которой каждому символу сопоставляется целое число в интервале от 0 до 2^{k-1} (код символа);

4) коды символов переводят в двоичную систему счисления.

Итак, присвоение символу конкретного двоичного кода – это вопрос соглашения, который фиксируется в кодовой таблице. **Таблицы кодировки** – стандарт, ставящий в соответствие каждому символу алфавита свой порядковый номер. При декодировании коды преобразуются в соответствующие символы.

Стандарт ASCII

Международным стандартом является 7-битная кодировка ASCII (англ. American Standard Code for Information Interchange – американский стандартный код для обмена информацией, 1964 г.), в которую входят $2^7 = 128$ символов с кодами от 0 до 127:

• служебные (управляющие) символы с кодами от 0 до 31;

• цифры от «**0**» до «**9**» с кодами от 48 до 57;

• латинские буквы: заглавные, от «**A**» до «**Z**» (с кодами от 65 до 90) и строчные, от «**a**» до «**z**» (с кодами от 97 до 122);

• знаки препинания: **. , : ; ! ?**

• скобки: **[] {} ()**

• математические символы: **+ - * / = < >**

• некоторые другие знаки: **" ' # $ % & ^ | @ \ _ ~**

Минимальная единица памяти, имеющая собственный адрес – это 8-битный байт. Поэтому для хранения кодов ASCII в памяти можно добавить к ним еще один (старший) нулевой бит, таким образом, получая 8-битную кодировку (**1 байт на символ**). Кроме того, дополнительный бит можно использовать: он дает возможность добавить в таблицу еще 128 символов с кодами от 128 до 255. Такое расширение ASCII часто называют кодовой страницей.

Первую половину кодовой страницы (коды от 0 до 127) занимает стандартная таблица ASCII, а вторую – символы национальных алфавитов (например, кириллица).

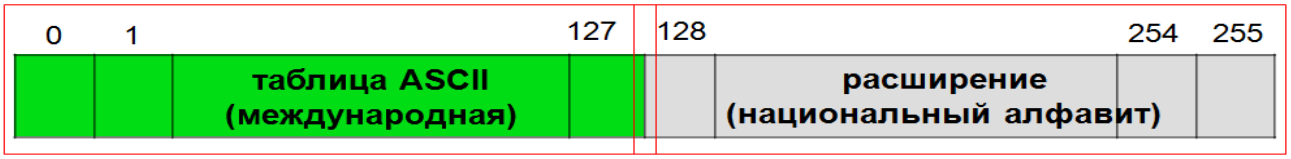


Рисунок 1 – Кодовая страница ASCII

Для русского языка существуют несколько кодовых страниц, которые были разработаны для разных операционных систем. Наиболее известны:

• кодовая страница Windows-1251 (CP-1251, см. рисунок 2) – в ОС Windows;

• кодовая страница KOI8-R – в ОС Unix;

• альтернативная кодировка (CP-1251) – в системе MS DOS;

• MacCyrillic – на компьютерах фирмы Apple.

128 Ъ	144 ђ	160	176 °	192 А	208 Р	224 а	240 р
129 Ѓ	145 ‘	161 Ў	177 ±	193 Б	209 С	225 б	241 с
130 ‚	146 ’	162 ў	178 І	194 В	210 Т	226 в	242 т
131 ѓ	147 “	163 Ј	179 і	195 Г	211 У	227 г	243 у
132 „	148 ”	164 ¤	180 ґ	196 Д	212 Ф	228 д	244 ф
133 …	149 •	165 Ґ	181 µ	197 Е	213 Х	229 е	245 х
134 †	150 –	166 ¦	182 ¶	198 Ж	214 Ц	230 ж	246 ц
135 ‡	151 —	167 §	183 ·	199 З	215 Ч	231 з	247 ч
136	152	168 Ё	184 ё	200 И	216 Ш	232 и	248 ш
137 ‰	153 ™	169 ©	185 №	201 Й	217 Щ	233 й	249 щ
138 Љ	154 љ	170 Є	186 є	202 К	218 Ъ	234 к	250 ъ
139 ‹	155 ›	171 «	187 »	203 Л	219 Ы	235 л	251 ы
140 Њ	156 њ	172 ¬	188 ј	204 М	220 Ь	236 м	252 ь
141 Ќ	157 ќ	173 -	189 S	205 Н	221 Э	237 н	253 э
142 Ћ	158 ћ	174 ®	190 s	206 О	222 Ю	238 о	254 ю
143 Џ	159 џ	175 Ї	191 ї	207 П	223 Я	239 п	255 я

Рисунок 2 – Кодовая страница Windows-1251

Стандарт UNICODE

Любая 8-битная кодовая страница имеет ограничение – она может включать только 256 символов. Поэтому не получится набрать в одном документе часть текста на русском языке, а часть – на испанском. Кроме того, существует проблема чтения документов, набранных с использованием другой кодовой страницы. Все это привело к принятию в 1991 году нового стандарта кодирования символов UNICODE, который позволяет одновременно записывать знаки любых существующих языков, математические и музыкальные символы и др.

Если расширить число используемых знаков, то необходимо увеличивать место, которое отводится под каждый символ. Компьютер работает сразу с одним или несколькими байтами, прочитанными из памяти. Поэтому место, отводимое на каждый символ, расширили сразу с одного байта до двух. Это позволило закодировать 2^{16} = 65 536 символов в одном наборе. В современной версии UNICODE можно кодировать до 2^{31} = 2 147 483 648 различных знаков, однако реально используются немногим более 100 000 символов.

В ОС Windows используется кодировка UNICODE, называемая UTF-16 (от англ. UNICODE Transformation Format – формат преобразования UNICODE). В ней на каждый символ отводится 16 бит (**2 байта**). В Unix-подобных системах, например, в Linux, чаще применяют кодировку UTF-8. В ней все символы, входящие в таблицу ASCII, кодируются в виде 1 байта, а другие символы могут занимать от 2 до 4 байт. Если значительную часть текста составляют латинские буквы и цифры, такой подход позволяет значительно уменьшить объем файла в сравнении с UTF-16. Текст, состоящий только из символов таблицы ASCII, кодируется точно так же, как и в кодировке ASCII.

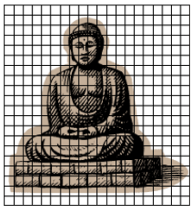
Достоинства кодировки UNICODE состоят в том, что она позволяет использовать символы разных языков в одном документе и решает проблему правильного отображения текста, вызванную использованием разных кодовых страниц. Но при этом увеличивается объем файлов.

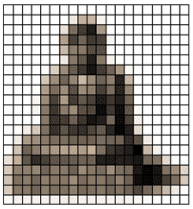
3 Кодирование графических данных

Графическая информация может быть представлена в аналоговой или дискретной форме. Примером аналогового представления графической информации является, например, фотография, а примером дискретного представления – изображение на экране монитора, состоящее из отдельных точек – *пикселей* (pixel – PIcture ELement) разного цвета.

Получение цифрового представления изображения основано на выполнении пространственной дискретизации аналогового изображения (осуществлении аналого-цифрового преобразования). Данный процесс заключается в разбиении непрерывного (аналогового) изображения на отдельные мелкие фрагменты, после чего цвет каждого фрагмента (а точнее – код цвета, например, в цветовой системе *RGB*) записывается в ячейку таблицы с координатами, соответствующими координатам фрагмента исходного изображения.

Растровое кодирование





Шаг 1. Дискретизация: разбивка на *пиксели*.

Пиксель – это наименьший элемент рисунка, для которого можно независимо установить цвет.

Шаг 2. Для каждого пикселя определяется **единый цвет**.

Есть потеря информации!
- **почему?**
- **как ее уменьшить?**

Разрешение: число пикселей на дюйм, *pixels per inch (ppi)*
экран 96 ppi, **печать 300-600** ppi, **типография 1200** ppi

Одним из устройств, которое выполняет дискретизацию изображения, является сканер. Сканер – это устройство для ввода в ЭВМ графической информации. Сканер осуществляет аналого-цифровое преобразование. К основным параметрам, определяющим результат работы сканера, относятся:

1) оптическое разрешение измеряется в точках на дюйм (dots per inch – dpi). Обычно указывается два значения, например, 600x1200 dpi, где горизонтальное разрешение (первое число) определяется CCD-матрицей[1] сканера, а вертикальное (второе число) определяется количеством шагов двигателя на дюйм;

2) глубина цвета определяется качеством CCD-матрицы и разрядностью АЦП. Измеряется количеством оттенков, которые устройство способно распознать (например, 24 бита соответствуют 16 777 216 оттенкам). В настоящее время сканеры выпускают с глубиной цвета 24, 30 и 36 бит.

Цифровое изображение обычно описывается следующими параметрами:

1) **глубина цвета** – количество битов, используемых для представления цвета при кодировании одного пикселя изображения:

$$\mathbf{I = \log_2 N},$$

где N – количество цветов в изображении, I – глубина цвета;

2) **цветовой диапазон** – максимальное количество цветов в изображении:

$$\mathbf{N = 2^I};$$

3) **размер изображения** – количество пикселей по вертикали (w) и по горизонтали (h);

4) **объем памяти**, занимаемой изображением:

$$\mathbf{I_I = I \cdot h \cdot w},$$

где I_I – объем памяти, занимаемый изображением, I – глубина цвета.

Описание цветов в ЭВМ основано на использовании цветовых моделей и соответствующих им способов кодирования цвета.

---

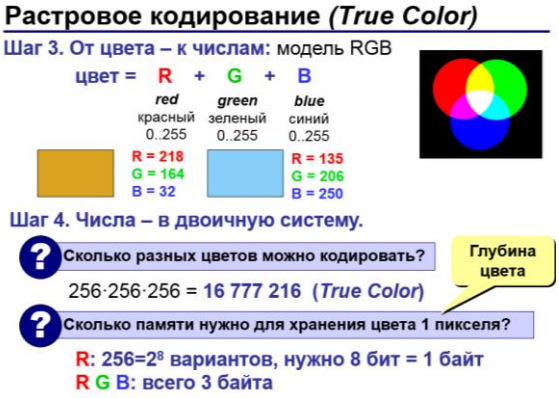
[1] CCD-матрица (Charge-Coupled Device) – специализированная аналоговая интегральная микросхема, состоящая из светочувствительных фотодиодов, выполненная на основе кремния и использующая технологию ПЗС (прибор с зарядовой связью).

Модель *RGB*

Согласно современному представлению о цветном зрении глаз человека содержит чувствительные элементы трех типов. Каждый из них воспринимает весь поток света, но первые наиболее чувствительны в области красного цвета, вторые – области зеленого, а третьи – в области синего цвета. Цвет – это результат возбуждения всех трех типов рецепторов. Поэтому считается, что любой цвет (т.е. ощущения человека, воспринимающего волны определенной длины) можно имитировать, используя только три световых луча (красный, зеленый, синий) разной яркости. Следовательно, любой цвет (в том числе и «белый») приближенно раскладывается на три составляющих – красную, зеленую и синюю. Меняя силу этих составляющих, можно составить любые цвета. Эта модель получила название *RGB* – Red (красный), Green (зеленый) и Blue (синий). Данная модель является аддитивной, т. е. требуемый произвольный цвет получается при сложении трех базовых цветов. Яркость каждого базового цвета может при этом принимать значения от 0 до 255 (256 значений); таким образом, данная модель позволяет кодировать $256 \cdot 256 \cdot 256 = 2^8 \cdot 2^8 \cdot 2^8 = 2^{24}$ цветов. Если значения яркостей всех базовых цветов равны, то образуемый цвет представляет собой один из оттенков серого.

Кроме цветовой модели RGB также используются CMYK, HSB, Lab и другие.

Размер файла не зависит от сложности изображения, а определяется только разрешением и глубиной цвета.



4 Кодирование аудиоданных

Звук – волна с непрерывно меняющейся амплитудой и частотой. Частоту звука измеряют в герцах – количество колебаний с секунду. Человек способен воспринимать звук от 16 Гц до 20 кГц.

Число ***T*** называется **интервалом дискретизации**, а обратная ему величина ***f*** – **частотой дискретизации** (один Гц – один отсчет в секунду, 1 кГц – 1000 отсчетов в секунду).

Чем больше частота дискретизации, тем точнее записан сигнал, тем меньше информации теряется. Но возрастает количество отсчетов, т.е. информационный объем кодированного звука.

В памяти есть только значения, снятые с интервалом *T*, остальная информация «теряется» при кодировании. В простейшем случае по ним можно восстановить ступенчатый сигнал. В современных звуковых картах для повышения качества этот ступенчатый сигнал сглаживается с помощью специальных фильтров.

Глубина кодирования – количество бит, которые выделяются на один отсчет или на кодирование различных уровней звука (количество уровней $N=2^i$).

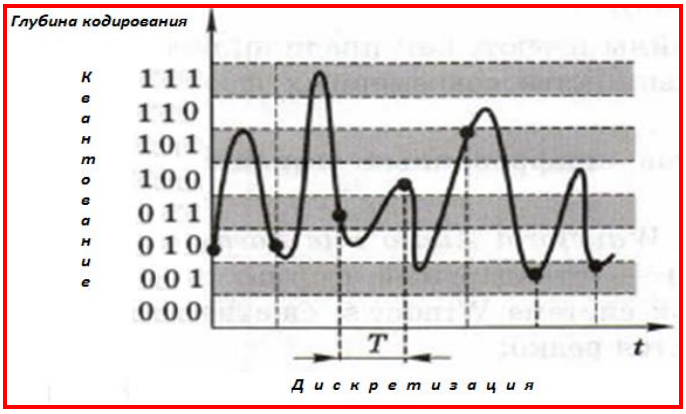


Рисунок – Дискретизация аналогово сигнала

Для хранения информации о звуке длительностью *t* секунд, закодированном с частотой дискретизации *f Гц* и глубиной кодирования *B* бит требуется следующее количество информации:

$$I = f \cdot t \cdot B$$

При двухканальной записи (*стерео*) объем памяти, необходимый для хранения данных одного канала, умножается на 2. *Квадро* – это 4 канала, поэтому результат умножается на 4.

5 Кодирование видеоданных

Для того чтобы сохранить видео в памяти компьютера, нужно закодировать звук и изменяющееся изображение, причем требуется обеспечить их синхронность. Для кодирования звука чаще всего используют оцифровку с частотой 48 кГц и глубиной кодирования 16 бит. Изображение состоит из отдельных растровых рисунков, которые меняются с ча-

стотой не менее 25 кадров в секунду, так что глаз человека воспринимает смену кадров как непрерывное движение. Это значит, что для каждой секунды видео нужно хранить в памяти 25 изображений.

Если используется размер 768 на 576 точек (стандарты PAL/SECAM) и глубина цвета 24 бита на пиксель, то закодированная 1 секунда видео будет занимать примерно 32 Мбайта, а 1 минута – около 1,85 Гбайт. Это недопустимо много, поэтому в большинстве форматов видеоизображений используется сжатие с потерями.

На практике используются различные алгоритмы сжатия для уменьшения скорости и объема потока видеоинформации – **кодеки –** метод сжатия аудио и видео и обратного восстановления данных. Это значит, что некоторые незначительные детали теряются, но «обычный» человек (непрофессионал) не почувствует существенного ухудшения качества. Например, один из алгоритмов сжатия заключается в том, что за короткое время изображение изменяется очень мало, поэтому можно запомнить «исходный» кадр, а затем сохранять только изменения. Через 10-15 секунд изображение изменяется настолько, что необходим новый исходный кадр.