1. Предмет и основные разделы кибернетики

Теория информации рассматривается как существенная часть кибернетики.

Кибернетика — это наука об общих законах получения, хранения, передачи и переработки информации. Ее основной предмет исследования — это так называемые кибернетические системы, рассматриваемые абстрактно, вне зависимости от их материальной природы. Примеры кибернетических систем: автоматические регуляторы в технике, ЭВМ, мозг человека или животных, биологическая популяция, социум. Часто кибернетику связывают с методами искусственного интеллекта, т.к. она разрабатывает общие принципы создания систем управления и систем для автоматизации умственного труда. Основными разделами (они фактически абсолютно самостоятельны и независимы) современной кибернетики считаются: теория информации, теория алгоритмов, теория автоматов, исследование операций, теория оптимального управления и теория распознавания образов.

Родоначальниками кибернетики (датой ее рождения считается 1948 год, год соответствующей публикации) считаются американские ученые Норберт Винер (Wiener, он — прежде всего) и Клод Шеннон (Shannon, он же основоположник теории информации).

Винер ввел основную категорию кибернетики — *управление,* показал существенные отличия этой категории от других, например, энергии, описал несколько задач, типичных для кибернетики, и привлек всеобщее внимание к особой роли вычислительных машин, считая их индикатором наступления новой НТР. Выделение категории управления позволило Винеру воспользоваться понятием информации, положив в основу кибернетики изучение законов передачи и преобразования информации.

Сущность принципа управления заключается в том, что движение и действие больших масс или передача и преобразование больших количеств энергии направляется и контролируется при помощи небольших количеств энергии, несущих информацию. Этот принцип управления лежит в основе организации и действия любых управляемых систем: автоматических устройств, живых организмов и т. п. Подобно тому, как введение понятия энергии позволило рассматривать все явления природы с единой точки зрения и отбросило целый ряд ложных теорий, так и введение понятия информации позволяет подойти с единой точки зрения к изучению самых различных процессов взаимодействия в природе.

В СССР значительный вклад в развитие кибернетики внесли академики Берг А. И. и Глушков В. М.

В нашей стране в 50-е годы кибернетика была объявлена лженаукой и была практически запрещена, что не мешало, однако, развиваться всем ее важным разделам (в том числе и теории информации) вне связи с обобщающим словом "кибернетика". Это было связано с тем, что сама по себе кибернетика представляет собой род философии, в кое-чем конфликтной с тогдашней официальной доктриной (марксистско-ленинской диалектикой).

Теория информации тесно связана с такими разделами математики как теория вероятностей и математическая статистика, а также прикладная алгебра, которые предоставляют для нее математический фундамент. С другой стороны теория информации исторически и практически представляет собой математический

фундамент теории связи. Часто теорию информации вообще рассматривают как одну из ветвей теории вероятностей или как часть теории связи. Таким образом, предмет "Теория информации" весьма узок, т. к. зажат между "чистой" математикой и прикладными (техническими) аспектами теории связи.

Теория информации представляет собой математическую теорию, посвященную измерению информации, ее потока, "размеров" канала связи и т.п., особенно применительно к радио, телеграфии, телевидению и к другим средствам связи. Первоначально теория была посвящена каналу связи, определяемому длиной волны и частотой, реализация которого была связана с колебаниями воздуха или электромагнитным излучением. Обычно соответствующий процесс был непрерывным, но мог быть и дискретным, когда информация кодировалась, а затем декодировалась. Кроме того, теория информации изучает методы построения кодов, обладающих полезными свойствами.

2. **Формальное представление знаний**

При формальном представлении знаний каждому описываемому объекту или понятию ставится в соответствие некоторый числовой код. Связи между кодируемыми сущностями также представляются кодами (адресами и указателями). Для такого перевода неформальных данных в формальный, цифровой вид должны использоваться специальные таблицы, сопоставляющие кодируемым сущностям их коды и называемые *таблицами кодировки.* Простейший пример такой таблицы — это ASCII (American Standard Code for Information Interchange), используемая повсеместно с вычислительной техникой. Она сопоставляет печатным и управляющим символам (управляющими являются, например, символы, отмечающие конец строки или страницы) числа от 0 до 127. Следующая программа на языке Паскаль выведет на экран все печатные символы этой таблицы и их коды:

```
var i: byte;
begin
  for i := 32 to 126 do
    write(i:6, chr(i):2);
  writeln
end.
```

На практике обычно используют не сам исходный ASCII, а так называемый расширенный ASCII (ASCII+), описывающий коды 256 символов (от 0 до 255). Первые 128 позиций расширенного ASCII совпадают со стандартом, а дополнительные 128 позиций определяются производителем оборудования или системного программного обеспечения. Кроме того, некоторым управляющим символам ASCII иногда назначают другое значение.

Хотя таблицы кодировки используются для формализации информации, сами они имеют неформальную природу, являясь мостом между реальными и формальными данными. Например, коду 65 в ASCII соответствует заглавная латинская буква A, но не конкретная, а любая. Этому коду будет соответствовать буква A, набранная жирным прямым шрифтом, и буква A, набранная нежирным с наклоном вправо на 9.5◦ шрифтом, и даже буква A готического шрифта. Задача сопоставления реальной букве ее кода в выбранной таблице кодировки очень сложна и частично решается программами распознания символов (например, Fine Reader).

3. **Виды информации**

Информация может быть двух видов: *дискретная (цифровая)* и *непрерывная (аналоговая)*. Дискретная информация характеризуется последовательными точными значениями некоторой величины, а непрерывная — непрерывным процессом изменения некоторой величины. Непрерывную информацию может, например, выдавать датчик атмосферного давления или датчик скорости автомашины. Дискретную информацию можно получить от любого цифрового индикатора: электронных часов, счетчика магнитофона и т.п.

Дискретная информация удобнее для обработки человеком, но непрерывная информация часто встречается в практической работе, поэтому необходимо уметь переводить непрерывную информацию в дискретную (дискретизация) и наоборот. Модем (это слово происходит от слов модуляция и демодуляция) представляет собой устройство для такого перевода: он переводит цифровые данные от компьютера в звук или электромагнитные колебания-копии звука и наоборот.

При переводе непрерывной информации в дискретную важна так называемая *частота дискретизации ν,* определяющая период *(T* = 1/ν) между измерениями значений непрерывной величины (см. рис. 1).

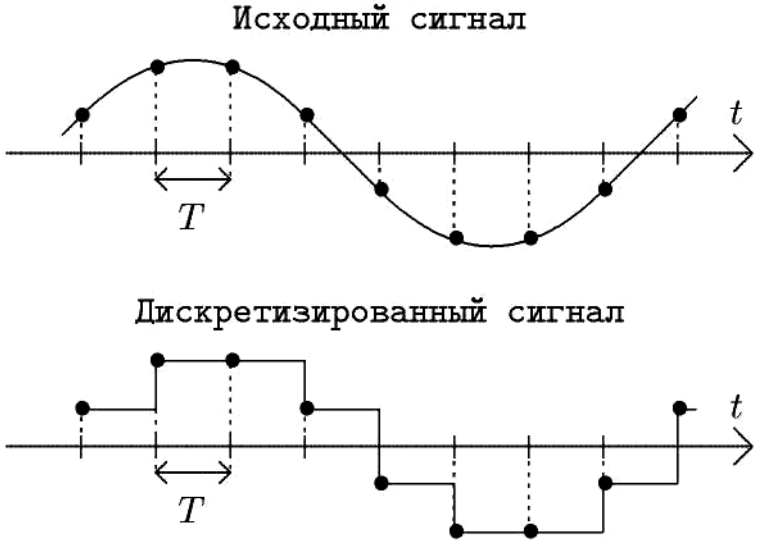


Рис. 1

Чем выше частота дискретизации, тем точнее происходит перевод непрерывной информации в дискретную. Но с ростом этой частоты растет и размер дискретных данных, получаемых при таком переводе, и, следовательно, сложность их обработки, передачи и хранения. Однако для повышения точности дискретизации необязательно безграничное увеличение ее частоты. Эту частоту разумно увеличивать только до предела, определяемого теоремой о выборках, называемой также теоремой Котельникова или законом Найквиста (Nyquist).

Любая непрерывная величина описывается множеством наложенных друг на друга волновых процессов, называемых гармониками, определяемых функциями вида $A\sin(\omega t + \varphi)$, где A — это амплитуда, ω — частота, t — время и φ — фаза.

Теорема о выборках утверждает, что для точной дискретизации ее частота должна быть не менее чем в два разы выше наибольшей частоты гармоники, входящей в дискретизируемую величину.

Примером использования этой теоремы являются лазерные компакт-диски, звуковая информация на которых хранится в цифровой форме. Чем выше будет

частота дискретизации, тем точнее будут воспроизводиться звуки и тем меньше их можно будет записать на один диск, но ухо обычного человека способно различать звуки с частотой до 20 КГц, поэтому точно записывать звуки с большей частотой бессмысленно. Согласно теореме о выборках частоту дискретизации нужно выбрать не меньшей 40 КГц (в промышленном стандарте на компакт-диске используется частота 44.1 КГц).

При преобразовании дискретной информации в непрерывную, определяющей является скорость этого преобразования: чем она выше, с тем более высокочастотными гармониками получится непрерывная величина. Но чем большие частоты встречаются в этой величине, тем сложнее с ней работать. Например, обычные телефонные линии предназначены для передачи звуков частотой до 3 КГц. Связь скорости передачи и наибольшей допустимой частоты подробнее будет рассмотрена далее.

Устройства для преобразования непрерывной информации в дискретную обобщающе называются АЦП (аналого-цифровой преобразователь) или ADC (Analog to Digital Convertor, A/D), а устройства для преобразования дискретной информации в аналоговую — ЦАП (цифро-аналоговый преобразователь) или DAC (Digital to Analog Convertor, D/A).

4. **Хранение, измерение, обработка и передача информации**

Для хранения информации используются специальные устройства памяти. Дискретную информацию хранить гораздо проще непрерывной, т. к. она описывается последовательностью чисел. Если представить каждое число в двоичной системе счисления, то дискретная информация предстанет в виде последовательностей нулей и единиц. Присутствие или отсутствие какого-либо признака в некотором устройстве может описывать некоторую цифру в какой-нибудь из этих последовательностей. Например, позиция на дискете описывает место цифры, а полярность намагниченности — ее значение. Для записи дискретной информации можно использовать ряд переключателей, перфокарты, перфоленты, различные виды магнитных и лазерных дисков, электронные триггеры и т. п. Одна позиция для двоичной цифры в описании дискретной информации называется *битом* (bit, binary digit). Бит служит для измерения информации. Информация размером в один бит содержится в ответе на вопрос, требующий ответа "да" или "нет". Непрерывную информацию тоже измеряют в битах.

Бит — это очень маленькая единица, поэтому часто используется величина в 8 раз большая — *байт (byte),* состоящая из двух 4-битных полубайт или тетрад. Байт обычно обозначают заглавной буквой В или Б. Как и для прочих стандартных единиц измерения для бита и байта существуют производные от них единицы, образуемые при помощи приставок кило (K), мега (M), гига (G или Г), тера (T), пета (P или П) и других. Но для битов и байтов они означают не степени 10, а степени двойки: кило — $2^{10} = 1024 \approx 10^3$, мега — $2^{20} \approx 10^6$, гига — $2^{30} \approx 10^9$, тера — $2^{40} \approx 10^{12}$, пета — $2^{50} \approx 10^{15}$. Например, 1 КВ = 8 Kbit = 1024 В = 8192 bit, 1 МБ = 1024 КБ = 1 048 576 Б = 8192 Кбит.

Для обработки информации используют вычислительные машины, которые бывают двух видов: ЦВМ (цифровая вычислительная машина) — для обработки дискретной информации, АВМ (аналоговая вычислительная машина) — для обработки непрерывной информации. ЦВМ — универсальны, на них можно решать

любые вычислительные задачи с любой точностью, но с ростом точности скорость их работы уменьшается. ЦВМ — это обычные компьютеры.

Каждая АВМ предназначена только для узкого класса задач, например, интегрирования или дифференцирования. Если на вход такой АВМ подать сигнал, описываемый функцией *f(t)*, то на ее выходе появится сигнал *F(t)* или *f'*(*t*). АВМ работают очень быстро, но их точность ограничена и не может быть увеличена без аппаратных переделок. Программа для АВМ — это электрическая схема из заданного набора электронных компонент, которую нужно физически собрать.

Бывают еще и гибридные вычислительные машины, сочетающие в себе элементы как ЦВМ, так и АВМ.

На рис. 2 изображена схема передачи информации.

Кодированием, например, является шифровка сообщения, декодированием — его дешифровка.

Процедуры кодирования и декодирования могут повторяться много раз. Ошибки при передаче информации происходят из-за шума в канале (атмосферные и технические помехи), а также при кодировании и декодировании. Теория информации изучает, в частности, способы минимизации количества таких ошибок.

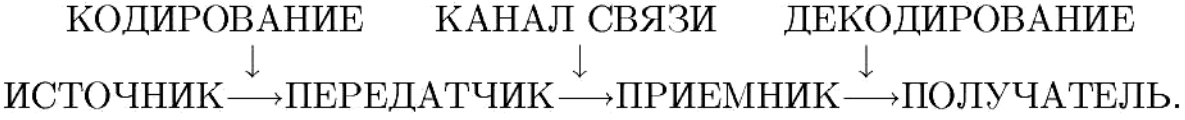


Рис. 2

Скорость передачи информации измеряется в количестве переданных за одну секунду бит или в бодах (baud): 1 бод = 1 бит/сек (bps). Производные единицы для бода такие же как и для бита и байта, например, 10 Kbaud = 10240 baud.

Информацию можно передавать *последовательно,* т. е. бит за битом, и *параллельно,* т.е. группами фиксированного количества бит. Параллельный способ быстрее, но он часто технически сложнее и дороже особенно при передаче данных на большие расстояния. Параллельный способ передачи используют, как правило, только на расстоянии не более 5 метров.

5. **Базовые понятия теории информации**

Информация — нематериальная сущность, при помощи которой с любой точностью можно описывать реальные (материальные), виртуальные (возможные) и понятийные сущности. Информация — противоположность неопределенности.

Канал связи — это среда передачи информации, которая характеризуется в первую очередь максимально возможной для нее скоростью передачи данных *(емкостью* канала связи).

Шум — это помехи в канале связи при передаче информации.

Кодирование — преобразование дискретной информации одним из следующих способов: шифрование, сжатие, защита от шума.

Общая схема передачи информации изображена на рис. 3.

Емкость канала связи без шума можно приблизительно вычислить, зная максимальную частоту волновых процессов, допустимую в этом канале. Можно считать, что скорость передачи данных может быть не меньше, чем эта частота. Например, при предельной частоте, равной 1000 Гц, можно обеспечить скорость передачи данных не меньше 1 Кбод.

Примеры каналов связи и связанных с ними предельных частот: телеграф — 140 Гц, телефон — до 3.1 КГц, короткие волны (10-100 м) — 3-30 МГц, УКВ (1-10 м) — 30-300 МГц, спутник (сантиметровые волны) — до 30 ГГц, оптический (инфракрасный диапазон) — 0.15— 400 ТГц, оптический (видимый свет) — 400-700 ТГц, оптический (ультрафиолетовый диапазон) — 0.7-1.75 ПГц.

Рис. 3

Типичные современные каналы: телеграфный и телефонный. Перспективные, внедряемые ныне: оптоволоконный (теработы) и цифровой телефонный (ISDN, Integrated Services Digital Networks) — 57-128 Кбод.

В реальных оптоволоконных системах скорость гораздо ниже теоретических пределов (редко превосходит 1-10 Гбод).

Наиболее широко пока используются телефонные линии связи. Здесь достигнута скорость более 50 Кбод!

6. **Способы измерения информации**

Понятие количества информации естественно возникает, например, в следующих типовых случаях:

1. Равенство вещественных переменных $a = b$, заключает в себе информацию о том, что a равно b. Про равенство $a^2 = b^2$ можно сказать, что оно несет меньшую информацию, чем первое, т.к. из первого следует второе, но не наоборот. Равенство $a^3 = b^3$ несет в себе информацию по объему такую же, как и первое;

2. Пусть происходят некоторые измерения с некоторой погрешностью. Тогда чем больше будет проведено измерений, тем больше информации об измеряемой сущности будет получено;

3. М.о. некоторой сл. в. содержит в себе информацию о самой сл. в. Для сл.в., распределенной по нормальному закону, с известной дисперсией знание м.о. дает полную информацию о сл.в.;

4. Рассмотрим схему передачи информации. Пусть передатчик описывается ел. в. X, тогда из-за помех в канале связи на приемник будет приходить сл.в. $Y = X + Z$, где Z — это сл.в., описывающая помехи. В этой схеме можно говорить о количестве информации, содержащейся в сл.в. Y, относительно X. Чем ниже уровень помех (дисперсия Z мала), тем больше информации можно получить из Y. При отсутствии помех Y содержит в себе всю информацию об X.

В 1865 г. немецкий физик Рудольф Клаузиус ввел в статистическую физику понятие энтропии или меры уравновешенности системы.

В 1921 г. основатель большей части математической статистики, англичанин Роналд Фишер впервые ввел термин "информация" в математику, но полученные им формулы носят очень специальный характер.

В 1948 г. Клод Шеннон в своих работах по теории связи выписывает формулы для вычисления количества информация и энтропии. Термин "энтропия" используется Шенноном по совету патриарха компьютерной эры фон Неймана, отметившего, что полученные Шенноном для теории связи формулы для ее расчета совпали с соответствующими формулами статистической физики, а также то, что "точно никто не знает" что же такое энтропия.

7. Вероятностный подход к измерению дискретной и непрерывной информации

В основе теории информации лежит предложенный Шенноном способ измерения количества информации, содержащейся в одной сл. в. относительно другой сл. в. Этот способ приводит к выражению количества информации числом.

Для д.с.в. X и Y, заданных законами распределения $P(X = X_i) = p_i$, $P(Y = Y_j) = q_j$ и совместным распределением $P(X = X_i, Y = Y_j) = p_{ij}$, *количество информации,* содержащейся в X относительно Y, равно

$$I(X,Y) = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i q_j}.$$

Для непрерывных сл. в. X и Y, заданных плотностями распределения вероятностей $p_X(t_1)$, $p_Y(t_2)$ и $P_{XY}(t_1,t_2)$, аналогичная формула имеет вид

$$I(X,Y) = \iint\limits_{\mathbb{R}^2} p_{XY}(t_1,t_2) \log_2 \frac{p_{XY}(t_1,t_2)}{p_X(t_1)p_Y(t_2)} dt_1 dt_2.$$

Очевидно, что

$$P(X = X_i, X = X_j) = \begin{cases} 0, & \text{при } i \neq j \\ P(X = X_i), & \text{при } i = j \end{cases}$$

и, следовательно,

$$I(X,X) = \sum_i p_i \log_2 \frac{p_i}{p_i p_i} = -\sum_i p_i \log_2 p_i.$$

Энтропия д. с. в. X в теории информации определяется формулой

$$H(X) = HX = I(X,X).$$

Свойства меры информации и энтропии:

1) $I(X,Y) \geqslant 0$, $I(X,Y) = 0 \quad \Leftrightarrow \quad X$ и Y независимы;
2) $I(X,Y) = I(Y,X)$;
3) $HX = 0 \quad \Leftrightarrow \quad X$ — константа;
4) $I(X,Y) = HX + HY - H(X,Y)$, где $H(X,Y) = -\sum_{i,j} p_{ij} \log_2 p_{ij}$;
5) $I(X,Y) \leqslant I(X,X)$. Если $I(X,Y) = I(X,X)$, то X — функция от Y.

1) Логарифмированием из очевидного для всех x неравенства $e^{x-1} \geq x$ (равенство устанавливается только при $x = 1$) получается неравенство $x - 1 \geq \ln x$ *или* $\frac{x-1}{\ln 2} \geq \log_2 x$

$$-I(X,Y) = \sum_{i,j} p_{ij} \log_2 \frac{p_i q_j}{p_{ij}} \leqslant \sum_{i,j} p_{ij} \frac{\frac{p_i q_j}{p_{ij}} - 1}{\ln 2} =$$

$$= \sum_{i,j} \frac{p_i q_j - p_{ij}}{\ln 2} = \frac{\sum_i p_i \sum_j q_j - \sum_{i,j} p_{ij}}{\ln 2} = \frac{1-1}{\ln 2} = 0,$$

т.е. $I(X, Y) = 0$ только при $p_{ij} = p_i q_j$ для всех i и j, т.е. при независимости X и Y. Если X и Y независимы, то $p_{ij} = p_i q_j$ и, следовательно, аргументы логарифмов равны 1 и, следовательно, сами логарифмы равны 0, что означает, что $I(X, Y) = 0$;

2) Следует из симметричности формул относительно аргументов;

3) Если $HX = 0$, то все члены суммы, определяющей HX, должны быть нули, что возможно только тогда и только тогда, когда X — константа;

4) Из четырех очевидных соотношений

$$\sum_j p_{ij} = p_i, \quad \sum_i p_{ij} = q_j,$$

$$HX = -\sum_i p_i \log_2 p_i = -\sum_{i,j} p_{ij} \log_2 p_i,$$

$$HY = -\sum_j q_j \log_2 q_j = -\sum_{i,j} p_{ij} \log_2 q_j$$

получается

$$HX + HY - H(X,Y) = \sum_{i,j} p_{ij} (\log_2 p_{ij} - \log_2 q_j - \log_2 p_i) = I(X,Y);$$

5) Нужно доказать

$$I(X,Y) = HX + HY - H(X,Y) \leqslant HX \text{ или } HY - H(X,Y) \leqslant 0$$

$$HY - H(X,Y) = -\sum_{i,j} p_{ij} \log_2 q_j + \sum_{i,j} p_{ij} \log_2 p_{ij} = \sum_{i,j} p_{ij} \log_2 (p_{ij}/q_j),$$

но $p_{ij} = P(X = X_i, Y = Y_j) \leqslant q_j = P(Y = Y_j)$,

а значит аргументы у всех логарифмов не больше 1 и, следовательно, значения логарифмов не больше 0, а это и значит, что вся сумма не больше 0.

Если $HX = I(X,X) = I(X,Y)$, то для каждого i p_{ij} равно либо q_j, либо 0. Но из

$p_{ij} = P(X = X_i, Y = Y_j) = P(X = X_i / Y = Y_j) P(Y = Y_j) \in \{q_j, 0\}$ следует $P(X = X_i / Y = Y_j) \in \{0,1\}$, что возможно только в случае, когда X — функция от Y.

При независимости сл. в. X и Y одна из них ничем не описывает другую, что и отражается в том, что для таких сл. в. $I(X, Y) = 0$.

Рассмотрим пример измерения количества информации при подбрасывании двух игральных костей.

Пусть заданы д. с. в. X_1, X_2 и Y. X_1 и X_2 — количества очков, выпавших соответственно на 1-й и 2-й игральной кости, а $Y = X_1 + X_2$. Найти $I(Y,X_1)$, $I(X_1,X_1)$, $I(Y, Y)$.

Законы распределения вероятностей для д. с. в. X_1 и X_2 совпадают, т.к. кости одинаковые и без изъянов.

X_1	1	2	3	4	5	6
p	$1/6$					

, т.е. при $j = 1...6$ $q_j = P(X_1 = j) = 1/6$.

Закон распределения вероятностей для д. с. в. *Y*,

$$P(Y = i) = P(X_1 + X_2 = i), \quad i = 2...12,$$

вследствие того, что X_1 X_2 — независимы и поэтому

$$P(X_1 = n, X_2 = m) = P(X_1 = n)P(X_2 = m),$$

будет

$$p_i = P(X_1 + X_2 = i) = \sum_{\substack{n+m=i \\ 1 \leqslant n,m \leqslant 6}} P(X_1 = n)P(X_2 = m) = \sum_{\substack{n+m=i \\ 1 \leqslant n,m \leqslant 6}} 1/36.$$

Таблицы, определяющие *Y:*

$X_2 \backslash X_1$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12,

$Y = X_1 + X_2$	2	3	4	5	6	7	8	9	10	11	12
p	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$,

т.е. при $i = 2...12$, $p_i = P(Y = i) = (6 - |7 - i|)/36$.

Закон совместного распределения вероятностей д. с. в. X_1 и *Y* будет

$$p_{ij} = P(Y = i, X_1 = j) = P(Y = i/X_1 = j)P(X_1 = j),$$

например,

$$P(Y = 2, X_1 = 1) = P(Y = 2/X_1 = 1)P(X_1 = 1) =$$

$$= P(X_2 = 1)P(X_1 = 1) = 1/36.$$

В общем случае получится

$$p_{ij} = P(Y = i, X_1 = j) = \begin{cases} 1/36, & \text{при } 1 \leqslant i - j \leqslant 6, \\ 0, & \text{иначе.} \end{cases}$$

$X_1 \backslash Y$	2	3	4	5	6	7	8	9	10	11	12
1	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0	0	0
2	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0	0
3	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0
4	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0
5	0	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0
6	0	0	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$

Тогда

$$I(Y,X_1)=\sum_{j=1}^{6}\sum_{1\leqslant i-j\leqslant 6} p_{ij}\log_2\frac{p_{ij}}{p_i q_j}=$$

$$=\frac{1}{36}\sum_{j=1}^{6}\sum_{1\leqslant i-j\leqslant 6}\log_2\frac{1}{6p_i}=$$

$$=\frac{1}{36}(\sum_{i=2}^{7}\log_2\frac{1}{6p_i}+\sum_{i=3}^{8}\log_2\frac{1}{6p_i}+\cdots+\sum_{i=6}^{11}\log_2\frac{1}{6p_i}+\sum_{i=7}^{12}\log_2\frac{1}{6p_i})=$$

$$=\frac{1}{36}((\log_2\frac{6}{1}+\log_2\frac{6}{2}+\cdots+\log_2\frac{6}{6})+\cdots+(\log_2\frac{6}{6}+\log_2\frac{6}{5}+\cdots+\log_2\frac{6}{1}))=$$

$$=\frac{1}{36}(\underline{2\log_2 6}+4\log_2 3+6\log_2 2+8\log_2\frac{3}{2}+10\log_2\frac{6}{5}+\underline{6\log_2 1})=$$

$$=(2+2\log_2 3+4\log_2 3+6+8\log_2 3-8+10\log_2 3+10-10\log_2 5)/36=$$

$$=(10+24\log_2 3-10\log_2 5)/36\approx 0.69 \text{ бит/символ.}$$

$I(X_1,X_1)=I(X_2,X_2)=-\sum_{j=1}^{6}q_j\log_2 q_j=\log_2 6=1+\log_2 3\approx 2.58$ бит/сим.

$I(Y,Y)=-\sum_{i=2}^{12}p_i\log_2 p_i=$

$$=\frac{1}{36}(2\log_2 36+4\log_2 18+6\log_2 12+8\log_2 9+10\log_2\frac{36}{5}+6\log_2 6)=$$

$$=(4+4\log_2 3+4+8\log_2 3+12+6\log_2 3+16\log_2 3+20+20\log_2 3-10\log_2 5+$$

$$+6+6\log_2 3)/36=(46+60\log_2 3-10\log_2 5)/36\approx 3.27 \text{ бит/сим.}$$

Здесь

$$0<I(Y,X_1)=I(Y,X_2)<I(X_1,X_1)=I(X_2,X_2)<I(Y,Y),$$

что соответствует свойствам информации.

Подчеркнутый член $\frac{1}{36}2\log_2 6=I(X_1,X_1)/18$ в расчете $I(X_1,Y)$ соответствует информации о двух случаях из 36, когда $Y = 2$ и $Y = 12$, которые однозначно определяют X_1. Шесть случаев, когда $Y = 7$, не несут никакой информации об X_1, что соответствует подчеркнутому члену $6\log_2 1=0$.

Расчеты можно проводить, используя 4-е свойство информации, через энтропию.

$H(Y,X_1)=-\sum_{i,j}p_{ij}\log_2 p_{ij}=\log_2 36=2(1+\log_2 3)=2HX_1\approx 5.17$ бит/сим.

$I(Y,X_1)=HX_1+HY-H(X_1,Y)=HY-HX_1\approx 3.27-2.58=0.69$ бит/сим.

Расчет количества информации с использованием 4-го свойства, а не определения, обычно требует меньше вычислений.

Рассмотрим более простой пример. Пусть д. с. в. *X* равна количеству очков, выпавших на игральной кости, а д. с. в. *Y* равна 0, если выпавшее количество очков нечетно, и 1, если выпавшее количество очков четно. Найти *I(X, Y)* и */(Y, Y).*

Составим законы распределения вероятностей д. с. в. *X* и *Y.*

X	1	2	3	4	5	6
p	1/6					

Y	0	1
p	1/2	

Таким образом, при $i = 1...6$ $p_i = P(X = i) = 1/6$ и, соответственно, при $j = 0...1$ $q_i = P(Y = j) = 1/2$.

Составим также закон совместного распределения вероятностей этих д. с. в.

X	1	3	5	2	4	6	1	3	5	2	4	6
Y	0	0	0	1	1	1	1	1	1	0	0	0
p	1/6						0					

$$\text{Таким образом, } p_{ij} = P(X = i, Y = j) = \begin{cases} 0, & \text{если } i + j \text{ — четно,} \\ 1/6, & \text{иначе.} \end{cases}$$

$$I(X,Y) = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i q_j} = 6\tfrac{1}{6} \log_2 2 = 1 \text{ бит/сим.}$$

$$I(Y,Y) = -\sum_{j=0}^{1} q_j \log_2 q_j = 2\tfrac{1}{2} \log_2 2 = 1 \text{ бит/сим.}$$

Точное количество выпавших очков дает точную информацию о четности, т.е. 1 бит. Из *I(X, Y) = I(Y,Y) = 1* бит/сим и 3-го свойства информации следует, что информация об *X* полностью определяет *Y*, но не наоборот, т.к. $I(X,Y) \neq I(X,X) = 1 + \log_2 3 \approx 2.58$ бит/сим. Действительно, *Y* функционально зависит от *X*, а *X* от *Y* функционально не зависит.

Расчеты через энтропию будут следующими

$$H(X,Y) = -\sum_{i,j} p_{ij} \log_2 p_{ij} = \log_2 6 = 1 + \log_2 3 = HX,$$
$$I(X,Y) = HX + HY - HX = HY = 1 \text{ бит/сим.}$$

8. Смысл энтропии Шеннона

Энтропия д. с. в. — это минимум среднего количества бит, которое нужно передавать по каналу связи о текущем значении данной д. с. в.

Рассмотрим пример (скачки). В заезде участвуют 4 лошади с равными шансами на победу, т.е. вероятность победы каждой лошади равна 1/4. Введем д. с. в. *X*, равную номеру победившей лошади. Здесь *HX = 2*. После каждого заезда по каналам связи достаточно будет передавать два бита информации о номере победившей лошади. Кодируем номер лошади следующим образом: 1—00, 2—01, 3—10, 4—11. Если ввести функцию *L(X),* которая возвращает длину сообщения, кодирующего заданное значение X, то м.о. *ML(X)* — это средняя длина сообщения, кодирующего *X.* Можно формально определить *L* через две функции *L(X) = len(code(X)),* где *code(X)* каждому значению *X* ставит в соответствие некоторый битовый код, причем, взаимно однозначно, а *len* возвращает длину в битах для любого конкретного кода. В этом примере *ML(X) = HX.*

Пусть теперь д. с. в. *X* имеет следующее распределение

$$P(X = 1) = \frac{3}{4},\ P(X = 2) = \frac{1}{8},\ P(X = 3) = P(X = 4) = \frac{1}{16},$$

т.е. лошадь с номером 1 — это фаворит. Тогда

$$HX = \frac{3}{4}\log_2\frac{4}{3} + \frac{1}{8}\log_2 8 + \frac{1}{8}\log_2 16 = \frac{19}{8} - \frac{3}{4}\log_2 3 \approx 1.186 \text{ бит/сим.}$$

Закодируем номера лошадей: 1—0, 2—10, 3—110, 4—111, — т.е. так, чтобы каждой код не был префиксом другого кода (подобное кодирование называют *префиксным)*. В среднем в 16 заездах 1-я лошадь должна победить в 12 из них, 2-я — в 2-х, 3-я — в 1-м и 4-я — в 1-м. Таким образом, средняя длина сообщения о победителе равна (1*12 + 2*2 + 3*1+3*1)/16 = 1.375 бит/сим или м.о. *L(X)*. Действительно, *L(X)* сейчас задается следующим распределением вероятностей

$$P(L(X) = 1) = 3/4,\ P(L(X) = 2) = 1/8,\ P(L(X) = 3) = 1/8.$$

Следовательно,

$$ML(X) = \frac{3}{4} + \frac{2}{8} + \frac{3}{8} = \frac{11}{8} = 1.375 \text{ бит/сим.}$$

Итак, *ML(X) > HX*.

Можно доказать, что более эффективного кодирования для двух рассмотренных случаев не существует.

То, что энтропия Шеннона соответствует интуитивному представлению о мере информации, может быть продемонстрировано в опыте по определению среднего времени психических реакций. Опыт заключается в том, что перед испытуемым человеком зажигается одна из *N* лампочек, которую он должен указать. Проводится большая серия испытаний, в которых каждая лампочка зажигается с определенной вероятностью p_i ($\sum_i^N p_i = 1$), где i — это номер лампочки. Оказывается, среднее время, необходимое для правильного ответа испытуемого, пропорционально величине энтропии — $\sum_{i=1}^N p_i \log_2 p_i$, а не числу лампочек *N,* как можно было бы подумать. В этом опыте предполагается, что чем больше информации будет получено человеком, тем дольше будет время ее обработки и, соответственно, реакции на нее.

9. Сжатие информации

Цель сжатия — уменьшение количества бит, необходимых для хранения или передачи заданной информации, что дает возможность передавать сообщения более быстро и хранить более экономно и оперативно (последнее означает, что операция извлечения данной информации с устройства ее хранения будет проходить быстрее, что возможно, если скорость распаковки данных выше скорости считывания данных с носителя информации). Сжатие позволяет, например, записать больше информации на дискету, "увеличить" размер жесткого диска, ускорить работу с модемом и т.д. При работе с компьютерами широко используются программы-архиваторы данных формата ZIP, GZ, ARJ и других. Методы сжатия информации были разработаны как математическая теория, которая долгое время (до первой половины 80-х годов), мало использовалась в компьютерах на практике.

Сжатие данных не может быть большим некоторого теоретического предела. Для формального определения этого предела рассматриваем любое информационное сообщение длины *n* как последовательность независимых, одинаково распределенных д. с. в. X_i или как выборки длины *n* значений одной д. с. в. *X*.

Доказано, что среднее количество бит, приходящихся на одно кодируемое значение д. с. в., не может быть меньшим, чем энтропия этой д.с.в., т.е. $ML(X) \geq HX$ для любой д.с.в. X и любого ее кода.

Кроме того, доказано утверждение о том, что существует такое кодирование (Шеннона-Фэно, Fano), что $HX \geq ML(X) - 1$.

Рассмотрим д.с.в. X_1 и X_2, независимые и одинаково распределенные. $HX_1 = HX_2$ и $I(X_1,X_2) = 0$, следовательно,

$$H(X_1, X_2) = HX_1 + HX_2 - I(X_1, X_2) = 2HX_1.$$

Вместо X_1 и X_2 можно говорить о двумерной д. с. в. $\vec{X} = (X_1, X_2)$.

Аналогичным образом для n-мерной д. с. в.

$$\vec{X} = (X_1, X_2, \ldots, X_n)$$

можно получить, что $H\vec{X} = nHX_1$.

Пусть $L_1(\vec{X}) = L(\vec{X})/n$, где $\vec{X} = (X_1, X_2, \ldots, X_n)$, т.е. $L_1(\vec{X})$ — это количество бит кода на единицу сообщения $\vec{X}$. Тогда $ML_1(\vec{X})$ — это среднее количество бит кода на единицу сообщения при передаче бесконечного множества сообщений $\vec{X}$. Из $ML(\vec{X}) - 1 \leqslant H\vec{X} \leqslant ML(\vec{X})$ для кода Шеннона-Фэно для $\vec{X}$ следует $ML_1(\vec{X}) - 1/n \leqslant HX_1 \leqslant ML_1(\vec{X})$ для этого же кода.

Таким образом, доказана *основная теорема о кодировании при отсутствии помех,* а именно то, что с ростом длины n сообщения, при кодировании методом Шеннона-Фэно всего сообщения целиком среднее количество бит на единицу сообщения будет сколь угодно мало отличаться от энтропии единицы сообщения. Подобное кодирование практически не реализуемо из-за того, что с ростом длины сообщения трудоемкость построения этого кода становится недопустимо большой. Кроме того, такое кодирование делает невозможным отправку сообщения по частям, что необходимо для непрерывных процессов передачи данных. Дополнительным недостатком этого способа кодирования является необходимость отправки или хранения собственно полученного кода вместе с его исходной длиной, что снижает эффект от сжатия. На практике для повышения степени сжатия используют *метод блокирования.*

По выбранному значению $\varepsilon > 0$ можно выбрать такое s, что если разбить все сообщение на блоки длиной s (всего будет n/s блоков), то кодированием Шеннона-Фэно таких блоков, рассматриваемых как единицы сообщения, можно сделать среднее количество бит на единицу сообщения большим энтропии менее, чем на ε. Действительно, пусть

$\vec{Y} = (\vec{Y_1}, \vec{Y_2}, \ldots, \vec{Y}_{n/s})$, $\vec{Y_1} = (X_1, X_2, \ldots, X_s)$, $\vec{Y_2} = (X_{s+1}, X_{s+2}, \ldots, X_{2s})$ и т.д., т.е. $\vec{Y_i} = (X_{s(i-1)+1}, X_{s(i-1)+2}, \ldots, X_{si})$. Тогда $H\vec{Y_1} = sHX_1$ и $sML_1(\vec{Y_1}) = ML(\vec{Y_1}) \leqslant H\vec{Y_1} + 1 = sHX_1 + 1$, следовательно,

$$ML_1(\vec{Y_1}) \leqslant HX_1 + 1/s,$$

т.е. достаточно брать $s = 1/\varepsilon$. Минимум s по заданному ε может быть гораздо меньшим $1/\varepsilon$.

Пример. Пусть д.с.в. X_1, X_2, ... X_n независимы, одинаково распределены и могут принимать только два значения $P(X_i = 0) = p = 3/4$ и $P(X_i = 1) = q = 1/4$ при i от 1 до n. Тогда

$$HX_i = \frac{3}{4}\log_2\frac{4}{3} + \frac{1}{4}\log_2 4 = 2 - \frac{3}{4}\log_2 3 \approx 0.811 \text{ бит/сим.}$$

Минимальное кодирование здесь — это коды 0 и 1 с длиной 1 бит каждый. При таком кодировании количество бит в среднем на единицу сообщения равно 1. Разобьем сообщение на блоки длины 2. Закон распределения вероятностей и кодирование для 2-мерной д. с. в. $\vec{X} = (X_1, X_2)$

$\vec{X}$	00	01	10	11
p	9/16	3/16	3/16	1/16
$code(\vec{X})$	0	10	110	111
$L(\vec{X})$	1	2	3	3.

Тогда при таком минимальном кодировании количество бит в среднем на единицу сообщения будет уже

$$ML_1(\vec{X}) = \big(1\frac{9}{16} + 2\frac{3}{16} + 3\frac{3}{16} + 3\frac{1}{16}\big)/2 = \frac{27}{32} = 0.84375,$$

т.е. меньше, чем для не блочного кодирования. Для блоков длины 3 количество бит в среднем на единицу сообщения можно сделать ≈ 0.823, для блоков длины 4 — ≈ 0.818 и т.д.

Все изложенное ранее подразумевало, что рассматриваемые д. с. в. кодируются только двумя значениями (обычно 0 и 1). Пусть д. с. в. кодируются ***m*** значениями. Тогда для д. с. в. $\vec{X}$ и любого ее кодирования верно, что $ML(\vec{X}) \geqslant H\vec{X}/\log_2 m$ и $ML_1(\vec{X}) \geqslant HX_1/\log_2 m$.

Кроме того, существует кодирование такое, что $ML(\vec{X}) - 1 \leqslant H\vec{X}/\log_2 m$ и $ML_1(\vec{X}) - 1/n \leqslant HX_1/\log_2 m$, где $n = \dim(\vec{X})$.

Формулы теоретических приделов уровня сжатия, рассмотренные ранее, задают предел для средней длины кода на единицу сообщений, передаваемых много раз, т.е. они ничего не говорят о нижней границе уровня сжатия, которая может достигаться на некоторых сообщениях и быть меньшей энтропии д.с.в., реализующей сообщение.

10. Простейшие алгоритмы сжатия информации

Метод Шеннона-Фэно состоит в следующем, значения д. с. в. располагают в порядке убывания их вероятностей, а затем последовательно делят на две части с приблизительно равными вероятностями, к коду первой части добавляют 0, а к коду второй — 1.

Для предшествующего примера получим

$\vec{X}$	p	$code(\vec{X})$
00	9/16	0
01	3/16	10
10	3/16	110
11	1/16	111,

$ML_1(\vec{X}) = 27/32 = 0.84375$ бит/сим.

Еще один пример. Код составляется после сортировки, т.е. после перестановки значений B и C.

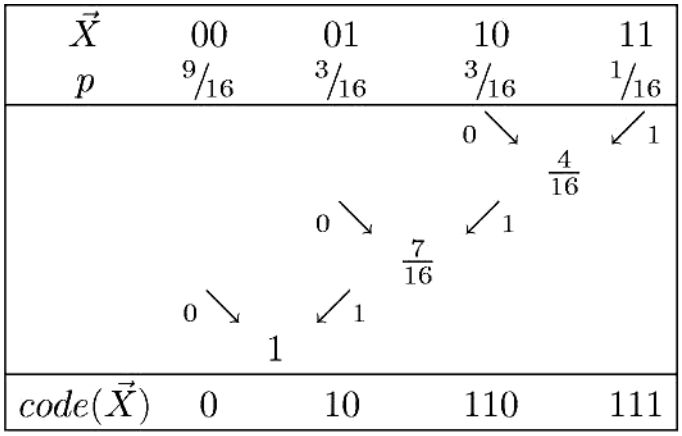
X	p	$code(X)$
A	0.4	0
B	0.2	11
C	0.4	10,

$ML(X) = ML_1(X) = 1.6$ бит/сим,
$HX = \log_2 5 - 0.8 \approx 1.523$ бит/сим.

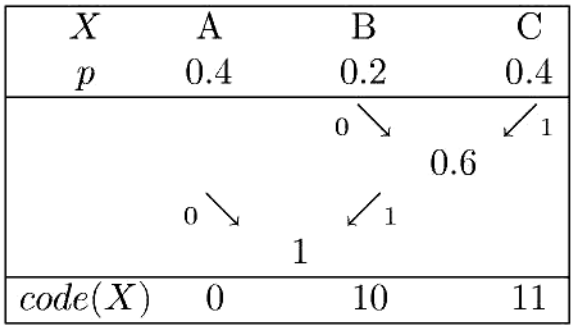
Метод Хаффмена (Huffman) разработан в 1952 г. Он более практичен и никогда по степени сжатия не уступает методу Шеннона-Фэно, более того, он сжимает максимально плотно. Код строится при помощи двоичного (бинарного) дерева. Вероятности значений д. с. в. приписываются его листьям; все дерево строится, опираясь на листья. Величина, приписанная к узлу дерева, называется весом узла. Два листа с наименьшими весами создают родительский узел с весом, равным сумме их весов; в дальнейшем этот узел учитывается наравне с оставшимися листьями, а образовавшие его узлы от такого рассмотрения устраняются. После постройки корня нужно приписать каждой из ветвей, исходящих из родительских узлов, значения 0 или 1. Код каждого значения д. с. в. — это число, получаемое при обходе ветвей от корня к листу, соответствующему данному значению.

Для методов Хаффмена и Шеннона-Фэно каждый раз вместе с собственно сообщением нужно передавать и таблицу кодов. Например, для случая из примера 2 нужно сообщить, что коду 10 соответствует символ C, коду 0 — A и т.д.

Построим коды Хаффмена для значений д. с. в. из двух предыдущих примеров.



$ML_1(\vec{X}) = ML(\vec{X})/2 = 27/32 = 0.84375$ бит/сим.



$ML_1(X) = ML(X) = 1.6$ бит/сим.

11. Арифметическое кодирование

Алгоритм кодирования Хаффмена, в лучшем случае, не может передавать на каждый символ сообщения менее одного бита информации. Предположим, известно, что в сообщении, состоящем из нулей и единиц, единицы встречаются в 10 раз чаще нулей. При кодировании методом Хаффмена и на 0 и на 1 придется тратить не менее одного бита. Но энтропия д.с.в., генерирующей такие сообщения ≈ 0.469 бит/сим. Метод Хаффмена дает для минимального среднего количества бит на один символ сообщения значение 1 бит. Хотелось бы иметь такую схему кодирования, которая позволяла бы кодировать некоторые символы менее чем одним битом. Одной из лучших среди таких схем является арифметическое кодирование, разработанное в 70-х годах XX века.

По исходному распределению вероятностей для выбранной для кодирования д. с. в. строится таблица, состоящая из пересекающихся только в граничных точках отрезков для каждого из значений этой д.с.в.; объединение этих отрезков должно образовывать отрезок [0,1], а их длины должны быть пропорциональны вероятностям соответствующих значений д. с. в. Алгоритм кодирования заключается в построении отрезка, однозначно определяющего данную последовательность значений д. с. в. Затем для построенного отрезка находится число, принадлежащее его внутренней части и равное целому числу, деленному на минимально возможную положительную целую степень двойки. Это число и будет кодом для рассматриваемой последовательности. Все возможные конкретные коды — это числа строго большие нуля и строго меньшие одного, поэтому можно отбрасывать лидирующий ноль и десятичную точку, но нужен еще один специальный код-маркер, сигнализирующий о конце сообщения. Отрезки строятся так. Если имеется отрезок для сообщения длины *n*—1, то для построения отрезка для сообщения длины *n,* разбиваем его на столько же частей, сколько значений имеет рассматриваемая д. с. в. Это разбиение делается совершенно также как и самое первое (с сохранением порядка). Затем выбирается из полученных отрезков тот, который соответствует заданной конкретной последовательности длины *n.*

Принципиальное отличие этого кодирования от рассмотренных ранее методов в его непрерывности, т. е. в ненужности блокирования. Эффективность арифметического кодирования растет с ростом длины сжимаемого сообщения (для кодирования Хаффмена или Шеннона-Фэно этого не происходит). Хотя арифметическое кодирование дает обычно лучшее сжатие, чем кодирование Хаффмена, оно пока используется на практике сравнительно редко, т.к. оно появилось гораздо позже и требует больших вычислительных ресурсов.

При сжатии заданных данных, например, из файла все рассмотренные методы требуют двух проходов. Первый для сбора частот символов, используемых как приближенные значения вероятностей символов, и второй для собственно сжатия.

Пример арифметического кодирования. Пусть д. с. в. X может принимать только два значения 0 и 1 с вероятностями 2/3 и 1/3 соответственно. Сопоставим значению 0 отрезок [0,2/3], а 1 — [2/3,1]. Тогда для д. с. в. $\vec{X}$,

$$\dim(\vec{X}) = 3,\ HX = H\vec{X}/3 = \log_2 3 - 2/3 \approx 0.9183 \text{ бит/сим.},$$

таблица построения кодов

Интервалы и коды			Вероятность	Код Хаффмена
		$111[\frac{26}{27}, 1] \ni \frac{31}{32} = 0.11111$	$^1/_{27}$	0000
	$11[\frac{8}{9}, 1]$	$110[\frac{8}{9}, \frac{26}{27}] \ni \frac{15}{16} = 0.1111$	$^2/_{27}$	0001
		$101[\frac{22}{27}, \frac{8}{9}] \ni \frac{7}{8} = 0.111$	$^2/_{27}$	010
$1[\frac{2}{3}, 1]$	$10[\frac{2}{3}, \frac{8}{9}]$	$100[\frac{2}{3}, \frac{22}{27}] \ni \frac{3}{4} = 0.11$	$^4/_{27}$	001
		$011[\frac{16}{27}, \frac{2}{3}] \ni \frac{5}{8} = 0.101$	$^2/_{27}$	011
	$01[\frac{4}{9}, \frac{2}{3}]$	$010[\frac{4}{9}, \frac{16}{27}] \ni \frac{1}{2} = 0.1$	$^4/_{27}$	100
		$001[\frac{8}{27}, \frac{4}{9}] \ni \frac{3}{8} = 0.011$	$^4/_{27}$	101
$0[0, \frac{2}{3}]$	$00[0, \frac{4}{9}]$	$000[0, \frac{8}{27}] \ni \frac{1}{4} = 0.01$	$^8/_{27}$	11.

$$ML_1(\vec{X}) = 65/81 \approx 0.8025 \text{ бит/сим. (арифметическое)},$$

$$ML_1(\vec{X}) = 76/81 \approx 0.9383 \text{ бит/сим. (блочный Хаффмена)},$$

$$ML_1(X) = ML(X) = 1 \text{ бит/сим. (Хаффмена)}.$$

Среднее количество бит на единицу сообщения для арифметического кодирования получилось меньше, чем энтропия. Это связано с тем, что в рассмотренной простейшей схеме кодирования, не описан код-маркер конца сообщения, введение которого неминуемо сделает это среднее количество бит большим энтропии.

Получение исходного сообщения из его арифметического кода происходит по следующему алгоритму.

Шаг 1. В таблице для кодирования значений д.с.в. определяется интервал, содержащий текущий код, — по этому интервалу однозначно определяется один символ исходного сообщения. Если этот символ — это маркер конца сообщения, то конец.

Шаг 2. Из текущего кода вычитается нижняя граница содержащего его интервала, полученная разность делится на длину этого же интервала. Полученное число считается новым текущим значением кода. Переход к шагу 1.

12. Адаптивные алгоритмы сжатия. Кодирование Хаффмена

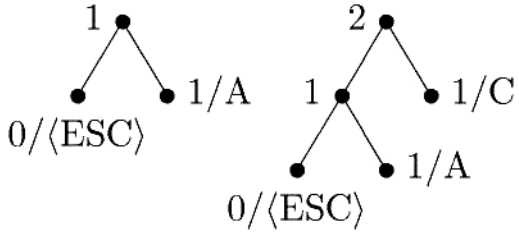
Является практичным, однопроходным, не требующим передачи таблицы кодов. Его суть в использовании *адаптивного алгоритма,* т.е. алгоритма, который при каждом сопоставлении символу кода, кроме того, изменяет внутренний ход вычислений так, что в следующий раз этому же символу может быть сопоставлен другой код, т.е. происходит адаптация алгоритма к поступающим для кодирования символам. При декодировании происходит аналогичный процесс.

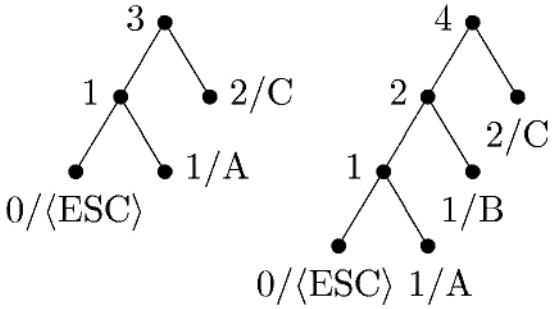
В начале работы алгоритма дерево кодирования содержит только один специальный символ, всегда имеющий частоту 0. Он необходим для занесения в дерево новых символов: после него код символа передается непосредственно. Обычно такой символ называют escape-символом ((ESC)). Расширенный ASCII кодируют каждый символ 8-битным числом, т.е. числом от 0 до 255. При построении дерева кодирования необходимо для возможности правильного декодирования как-то упорядочивать структуру дерева. Расположим листья дерева в

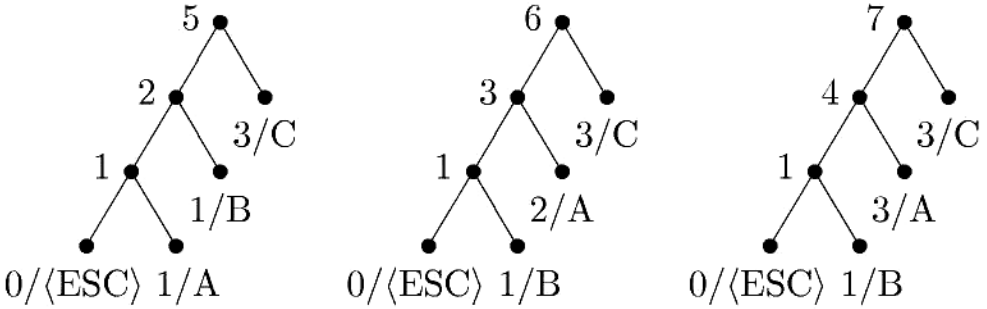
порядке возрастания частот и затем в порядке возрастания стандартных кодов символов. Узлы собираются слева направо без пропусков. Левые ветви помечаются 0, а правые — 1.

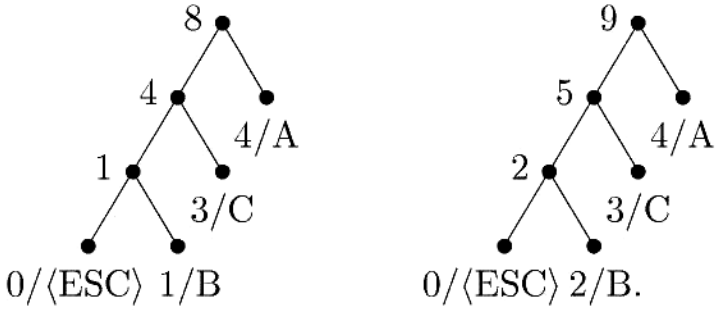
Рассмотрим процесс построения кодов по адаптивному алгоритму Хаффмена для сообщения ACCBCAAABC, которое соответствует выборке 10-и значений д. с. в. X из 2-го примера на построение неадаптивного кода Хаффмена:

входные данные	код	длина кода	№ дерева
A	'A'	8	1
C	0'C'	9	2
C	1	1	3
B	00'B'	10	4
C	1	1	5
A	001	3	6
A	01	2	7
A	01	2	8
B	001	3	9
C	01	2	









Здесь L_1(ACCBCAAABC) = 4.1 бит/сим. Если не использовать сжатия, то L_1(ACCBCAAABC) = 8 бит/сим. Для рассматриваемой д.с.в. ранее были получены значения $ML_1(X)$ = 1.6 бит/сим и $HX \approx 1.523$ бит/сим. Но с ростом длины сообщения среднее количество бит на символ сообщения при адаптивном алгоритме кодирования будет мало отличаться от значения, полученного при использовании неадаптивного метода Хаффмена или Шеннона-Фэно, т. к. алфавит символов ограничен и полный код каждого символа нужно передавать только один раз.

Теперь рассмотрим процесс декодирования сообщения 'A'0'C'100 'B'1001010100101. Здесь и далее символ в апостофах означает восемь бит, представляющих собой запись двоичного числа, номера символа, в таблице ASCII+.

В начале декодирования дерево Хаффмена содержит только escape-символ с частотой 0. С раскодированием каждого нового символа дерево заново перестраивается.

входные данные	символ	№ дерева
'A'	A	1
0'C'	C	2
1	C	3
00'B'	B	4
...	...	...

Выбранный способ адаптации алгоритма очень неэффективный, т.к. после обработки каждого символа нужно перестраивать все дерево кодирования. Существуют гораздо менее трудоемкие способы, при которых не нужно перестраивать все дерево, а нужно лишь незначительно изменять.

Бинарное дерево называется *упорядоченным,* если его узлы могут быть перечислены в порядке неубывания веса и в этом перечне узлы, имеющие общего родителя, должны находиться рядом, на одном ярусе. Причем перечисление должно идти по ярусам снизу-вверх и слева-направо в каждом ярусе.

На рис. 4 приведен пример упорядоченного дерева Хаффмена.

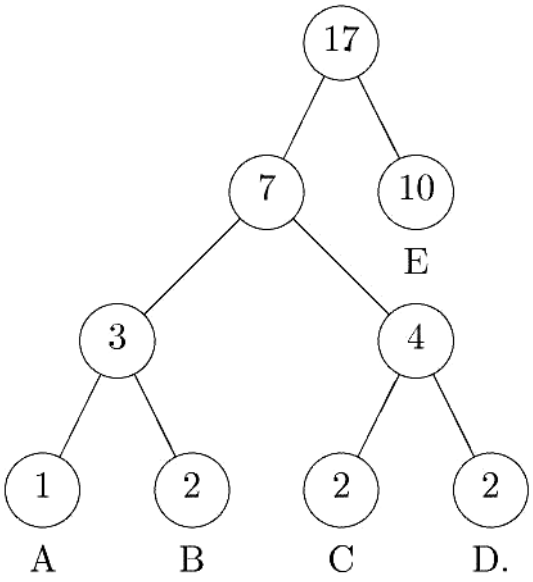


Рис. 4

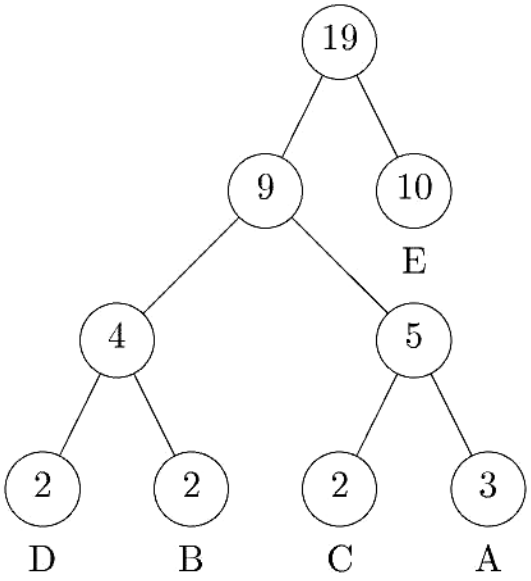


Рис. 5

Если дерево кодирования упорядоченно, то при изменении веса существующего узла дерево не нужно целиком перестраивать — в нем достаточно лишь поменять местами два узла: узел, вес которого нарушил упорядоченность, и последний из следующих за ним узлов меньшего веса. После перемены мест узлов, необходимо пересчитать веса всех их узлов-предков.

Например, если в дереве на рис. 4 добавить еще две буквы A, то узлы A и D должны поменяться местами (см. рис. 5).

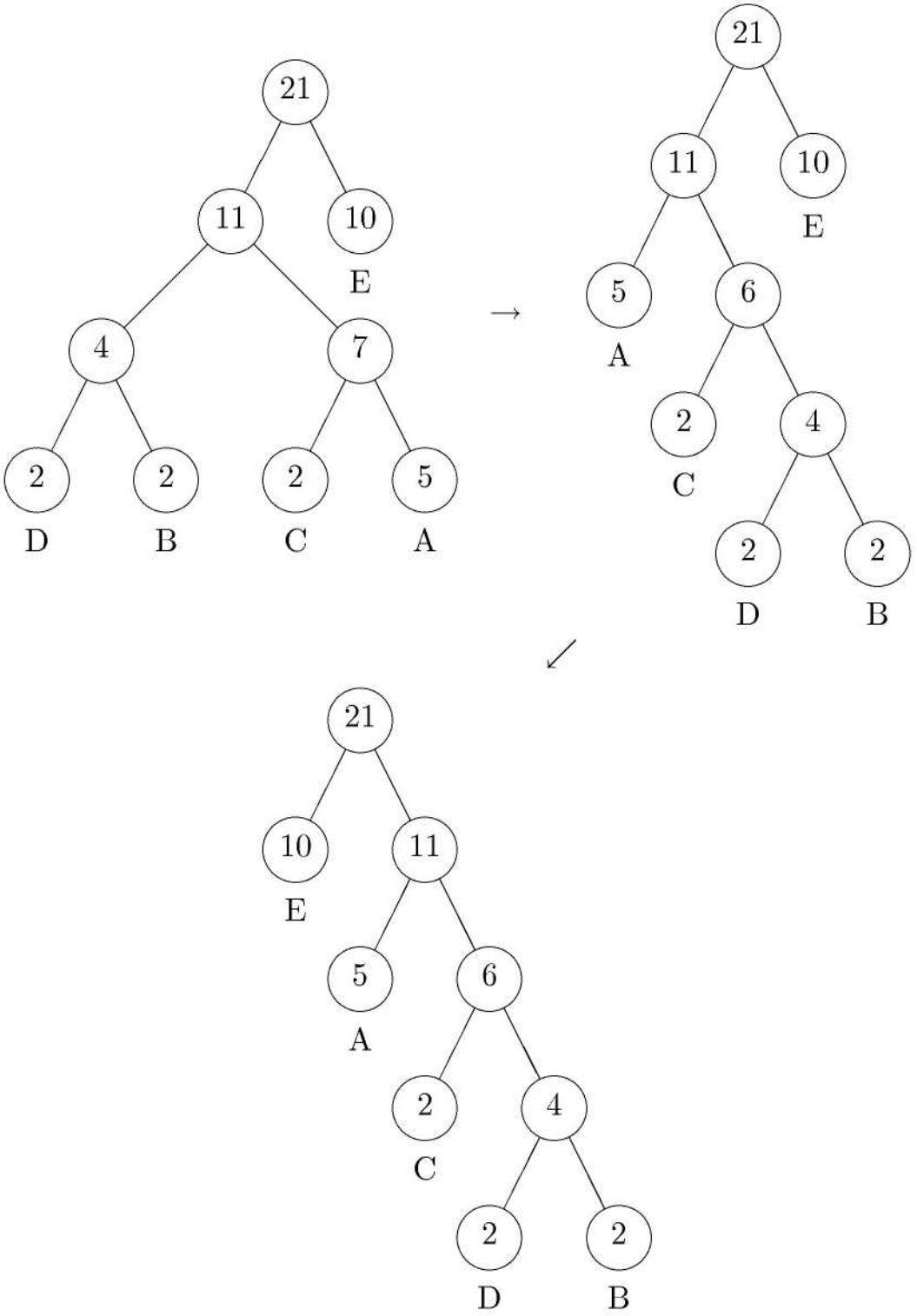


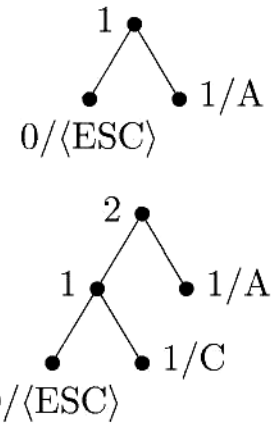
Рис. 6

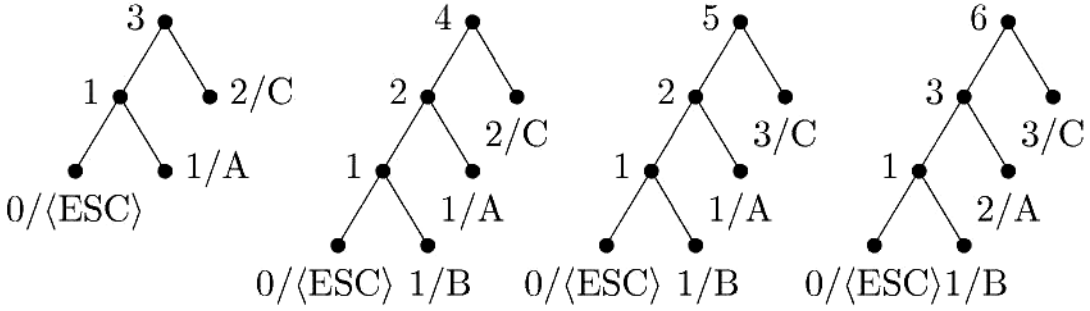
Если добавить еще две буквы A, то необходимо будет поменять местами сначала узел A и узел, родительский для узлов D и B, а затем узел E и узел-брат E (рис. 6).

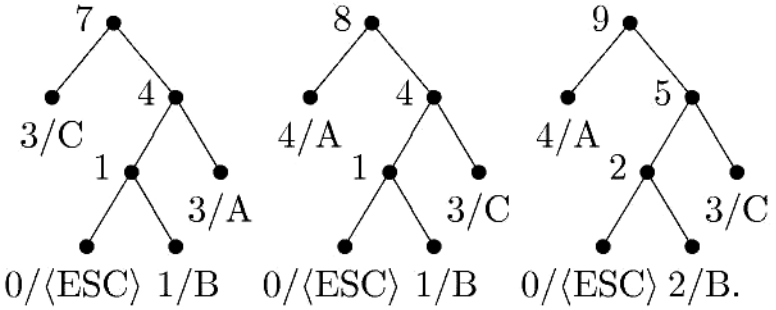
Дерево нужно перестраивать только при появлении в нем нового узла-листа. Вместо полной перестройки можно добавлять новый лист справа к листу (ESC) и упорядочивать, если необходимо, полученное таким образом дерево.

Процесс работы адаптивного алгоритма Хаффмена с упорядоченным деревом можно изобразить следующей схемой:

входные данные	код	длина кода	№ дерева	метод получения
A	'A'	8	1	добавление узла
C	0'C'	9	2	добавление узла
C	01	2	3	упорядочение
B	00'B'	10	4	добавление узла
C	1	1	5	не меняется
A	01	2	6	не меняется
A	01	2	7	упорядочение
A	11	2	8	упорядочение
B	101	3	9	не меняется
C	11	2		







Здесь получится $L_1(\text{ACCBCAAABC}) = 4.1$ бит/сим.

13. Адаптивное арифметическое кодирование

Для арифметического кодирования, как и для кодирования методом Хаффмена, существуют адаптивные алгоритмы. Реализация одного из них запатентована фирмой IBM.

Построение арифметического кода для последовательности символов из заданного множества можно реализовать следующим алгоритмом. Каждому символу сопоставляется его вес: вначале он для всех равен 1. Все символы располагаются в естественном порядке, например, по возрастанию. Вероятность каждого символа устанавливается равной его весу, деленному на суммарный вес всех символов. После получения очередного символа и постройки интервала для него, вес этого символа увеличивается на 1 (можно увеличивать вес любым регулярным способом).

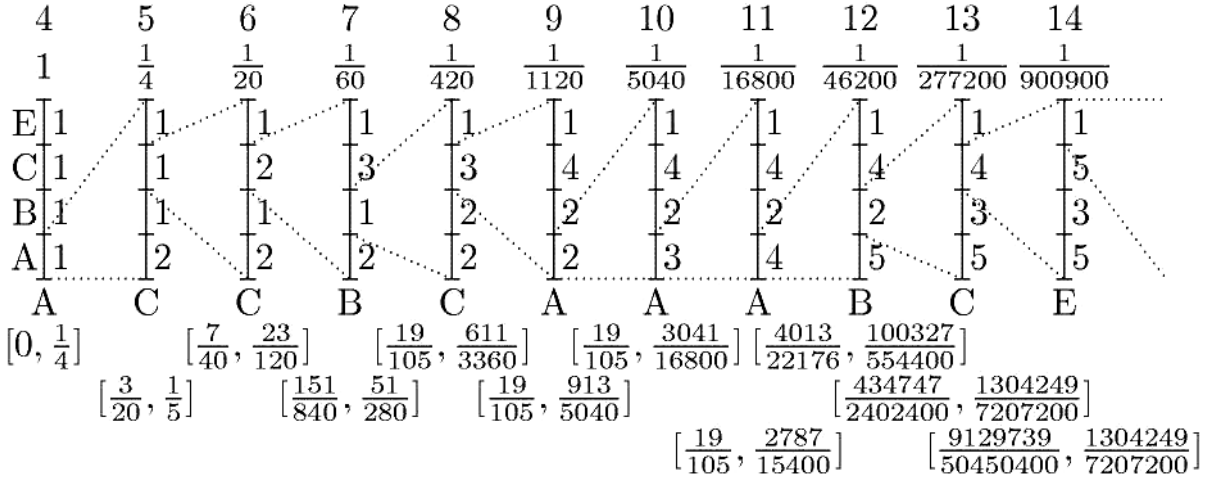


Рис. 7

Заданное множество символов — это, как правило, ASCII+. Для того, чтобы обеспечить остановку алгоритма распаковки вначале сжимаемого сообщения надо поставить его длину или ввести дополнительный символ-маркер конца сообщения. Если знать формат файла для сжатия, то вместо начального равномерного распределения весов можно выбрать распределение с учетом этих знаний. Например, в текстовом файле недопустимы ряд управляющих символов и их вес можно занулить.

Пример. Пусть заданное множество — это символы A, B, C. Сжимаемое сообщение — ACCBCAAABC. Введем маркер конца сообщения — E. Кодирование согласно приведенному алгоритму можно провести согласно схеме, приведенной на рис. 7.

Вследствие того, что

$$\frac{759021}{2^{22}=4194304} = 0.00101110010100111011 01_2 \in \left(\frac{9129739}{50450400}, \frac{1304249}{7207200}\right),$$

$$code(\text{ACCBCAAABC}) = 00101110010100111011 01 \text{ и}$$

$$L(\text{ACCBCAAABC}) = 22.$$

Поэтому L_1 (ACCBCAAABC) = 2.2 бит/сим. Результат, полученный адаптивным алгоритмом Хаффмена — 4.1 бит/сим, но если кодировать буквы не 8 битами, а 2, то результат будет 2.3 бит/сим. В первой строчке схемы выписаны суммарные веса символов, а во второй — длины текущих отрезков.

Способ распаковки адаптивного арифметического кода почти аналогичен приведенному для неадаптивного. Отличие только в том, что на втором шаге после получения нового кода нужно перестроить разбиение единичного отрезка согласно новому распределению весов символов. Получение маркера конца или заданного началом сообщения числа символов означает окончание работы.

Пример. Распакуем код 001011100101001110110 1, зная, что множество символов сообщения состоит из A, B, C и E, причем последний — это маркер конца сообщения.

$$0.00101110010100111011012 = \frac{759021}{4194304}.$$

Веса				Число-код и его интервал	Символ	Длина интервала
A	B	C	E			
1	1	1	1	$\frac{759021}{4194304} \in (0, \frac{1}{4})$	A	$\frac{1}{4}$
2	1	1	1	$\frac{759021}{1048576} \in (\frac{3}{5}, \frac{4}{5})$	C	$\frac{1}{5}$
2	1	2	1	$\frac{649377}{1048576} \in (\frac{1}{2}, \frac{5}{6})$	C	$\frac{1}{3}$
2	1	3	1	$\frac{375267}{1048576} \in (\frac{2}{7}, \frac{3}{7})$	B	$\frac{1}{7}$
2	2	3	1	$\frac{529717}{1048576} \in (\frac{1}{2}, \frac{7}{8})$	C	$\frac{3}{8}$
2	2	4	1	$\frac{5429}{393216} \in (0, \frac{2}{9})$	A	$\frac{2}{9}$
3	2	4	1	$\frac{16287}{262144} \in (0, 0.3)$	A	0.3
4	2	4	1	$\frac{27145}{131072} \in (0, \frac{4}{11})$	A	$\frac{4}{11}$
5	2	4	1	$\frac{298595}{524288} \in (\frac{5}{12}, \frac{7}{12})$	B	$\frac{1}{6}$
5	3	4	1	$\frac{240425}{262144} \in (\frac{8}{13}, \frac{12}{13})$	C	$\frac{4}{13}$
5	3	5	1	$\frac{1028373}{1048576} \in (\frac{13}{14}, 1)$	E	

14. Подстановочные или словарно-ориентированные алгоритмы сжатия информации. Методы Лемпела-Зива

Методы Шеннона-Фэно, Хаффмена и арифметическое кодирование обобщающе называются *статистическими* методами. Словарные алгоритмы носят менее математически обоснованный, но более практичный характер.

Алгоритм LZ77 был опубликован в 1977 г. Разработан израильскими математиками Якобом Зивом (Ziv) и Авраамом Лемпелом (Lempel). Многие программы сжатия информации используют ту или иную модификацию LZ77. Одной из причин популярности алгоритмов LZ является их исключительная простота при высокой эффективности сжатия.

Основная идея LZ77 состоит в том, что второе и последующие вхождения некоторой строки символов в сообщении заменяются ссылками на ее первое вхождение.

LZ77 использует уже просмотренную часть сообщения как словарь. Чтобы добиться сжатия, он пытается заменить очередной фрагмент сообщения на указатель в содержимое словаря.

LZ77 использует "скользящее" по сообщению окно, разделенное на две неравные части. Первая, большая по размеру, включает уже просмотренную часть сообщения. Вторая, намного меньшая, является буфером, содержащим еще незакодированные символы входного потока.

Обычно размер окна составляет несколько килобайт, а размер буфера — не более ста байт. Алгоритм пытается найти в словаре (большей части окна) фрагмент, совпадающий с содержимым буфера.

Алгоритм LZ77 выдает коды, состоящие из трех элементов:

- смещение в словаре относительно его начала подстроки, совпадающей с началом содержимого буфера;
- длина этой подстроки;
- первый символ буфера, следующий за подстрокой.

Пример. Размер окна — 20 символ, словаря — 12 символов, а буфера — 8. Кодируется сообщение "ПРОГРАММНЫЕ ПРОДУКТЫ ФИРМЫ MICROSOFT". Пусть словарь уже заполнен. Тогда он содержит строку "ПРОГРАММНЫЕ ", а буфер строку "ПРОДУКТЫ". Просматривая словарь, алгоритм обнаружит, что совпадающей подстрокой будет "ПРО", в словаре она расположена со смещением 0 и имеет длину 3 символа, а следующим символом в буфере является "Д". Таким образом, выходным кодом будет тройка (0,3,'Д'). После этого алгоритм сдвигает влево все содержимое окна на длину совпадающей подстроки +1 и одновременно считывает столько же символов из входного потока в буфер. Получаем в словаре строку "РАММНЫЕ ПРОД", в буфере — "УКТЫ ФИР". В данной ситуации совпадающей подстроки обнаружить не удаться и алгоритм выдаст код (0,0,'У'), после чего сдвинет окно на один символ. Затем словарь будет содержать "АММНЫЕ ПРО-ДУ", а буфер — "КТЫ ФИРМ". И т.д.

Декодирование кодов LZ77 проще их получения, т.к. не нужно осуществлять поиск в словаре.

Недостатки LZ77:

1) с ростом размеров словаря скорость работы алгоритма-кодера пропорционально замедляется;
2) кодирование одиночных символов очень неэффективно.

Пример. Закодировать по алгоритму LZ77 строку "КРАСНАЯ КРАСКА".

СЛОВАРЬ(8)	БУФЕР(5)	КОД
"........"	"КРАСН"	<0,0,'К'>
".......К"	"РАСНА"	<0,0,'Р'>
"......КР"	"АСНАЯ"	<0,0,'А'>
".....КРА"	"СНАЯ "	<0,0,'С'>
"....КРАС"	"НАЯ К"	<0,0,'Н'>
"...КРАСН"	"АЯ КР"	<5,1,'Я'>
".КРАСНАЯ"	" КРАС"	<0,0,' '>
"КРАСНАЯ "	"КРАСК"	<0,4,'К'>
"АЯ КРАСК"	"А...."	<0,0,'А'>

В последней строчке, буква "А" берется не из словаря, т. к. она последняя.

Длина кода вычисляется следующим образом: длина подстроки не может быть больше размера буфера, а смещение не может быть больше размера словаря − 1. Следовательно, длина двоичного кода смещения будет округленным в большую сторону $\log_2$(размер словаря), а длина двоичного кода для длины подстроки будет округленным в большую сторону $\log_2$ (размер буфера+1). А символ кодируется 8 битами (например, ASCII+).

В последнем примере длина полученного кода равна 9 * (3 + 3 + 8) = 126 бит, против 14 * 8 = 112 бит исходной длины строки.

В 1982 г. Сторером (Storer) и Шиманским (Szimanski) на базе LZ77 был разработан алгоритм LZSS, который отличается от LZ77 производимыми кодами.

Код, выдаваемый LZSS, начинается с однобитного префикса, различающего собственно код от незакодированного символа. Код состоит из пары: смещение и длина, такими же как и для LZ77. В LZSS окно сдвигается ровно на длину найденной подстроки или на 1, если не найдено вхождение подстроки из буфера в словарь. Длина подстроки в LZSS всегда больше нуля, поэтому длина двоичного кода для длины подстроки — это округленный до большего целого двоичный логарифм от длины буфера.

Пример. Закодировать по алгоритму LZSS строку "КРАСНАЯ КРАСКА".

СЛОВАРЬ(8)	БУФЕР(5)	КОД	ДЛИНА КОДА
"........"	"КРАСН"	0'К'	9
".......К"	"РАСНА"	0'Р'	9
"......КР"	"АСНАЯ"	0'А'	9
".....КРА"	"СНАЯ "	0'С'	9
"....КРАС"	"НАЯ К"	0'Н'	9
"...КРАСН"	"АЯ КР"	1<5,1>	7
"..КРАСНА"	"Я КРА"	0'Я'	9
".КРАСНАЯ"	" КРАС"	0' '	9
"КРАСНАЯ "	"КРАСК"	1<0,4>	7
"НАЯ КРАС"	"КА..."	1<4,1>	7
"АЯ КРАСК"	"А...."	1<0,1>	7

Здесь длина полученного кода равна 7*9 + 4*7 = 91 бит.

LZ77 и LZSS обладают следующими очевидными недостатками:

1) невозможность кодирования подстрок, отстоящих друг от друга на расстоянии, большем длины словаря;

2) длина подстроки, которую можно закодировать, ограничена размером буфера.

Если механически чрезмерно увеличивать размеры словаря и буфера, то это приведет к снижению эффективности кодирования, т.к. с ростом этих величин будут расти и длины кодов для смещения и длины, что сделает коды для коротких подстрок недопустимо большими. Кроме того, резко увеличится время работы алгоритма-кодера.

В 1978 г. авторами LZ77 был разработан алгоритм LZ78, лишенный названных недостатков.

LZ78 не использует "скользящее" окно, он хранит словарь из уже просмотренных фраз. При старте алгоритма этот словарь содержит только одну пустую строку (строку длины нуль). Алгоритм считывает символы сообщения до тех пор, пока накапливаемая подстрока входит целиком в одну из фраз словаря. Как только эта строка перестанет соответствовать хотя бы одной фразе словаря, алгоритм генерирует код, состоящий из индекса строки в словаре, которая до последнего введенного символа содержала входную строку, и символа,

нарушившего совпадение. Затем в словарь добавляется введенная подстрока. Если словарь уже заполнен, то из него предварительно удаляют менее всех используемую в сравнениях фразу.

Ключевым для размера получаемых кодов является размер словаря во фразах, потому что каждый код при кодировании по методу LZ78 содержит номер фразы в словаре. Из последнего следует, что эти коды имеют постоянную длину, равную округленному в большую сторону двоичному логарифму размера словаря +8 (это количество бит в байт-коде расширенного ASCII).

Пример. Закодировать по алгоритму LZ78 строку "КРАСНАЯ КРАСКА", используя словарь длиной 16 фраз.

ВХОДНАЯ ФРАЗА (В СЛОВАРЬ)	КОД	ПОЗИЦИЯ СЛОВАРЯ
""		0
"К"	<0,'К'>	1
"Р"	<0,'Р'>	2
"А"	<0,'А'>	3
"С"	<0,'С'>	4
"Н"	<0,'Н'>	5
"АЯ"	<3,'Я'>	6
" "	<0,' '>	7
"КР"	<1,'Р'>	8
"АС"	<3,'С'>	9
"КА"	<1,'А'>	10

Указатель на любую фразу такого словаря — это число от 0 до 15, для его кодирования достаточно четырех бит.

В последнем примере длина полученного кода равна 10 * (4 + 8) = 120 битам.

В 1984 г. Уэлчем (Welch) был путем модификации LZ78 создан алгоритм LZW.

Пошаговое описание алгоритма-кодера.

Шаг 1. Инициализация словаря всеми возможными односимвольными фразами (обычно 256 символами расширенного ASCII). Инициализация входной фразы w первым символом сообщения.

Шаг 2. Считать очередной символ K из кодируемого сообщения.

Шаг 3. Если КОНЕЦ_СООБЩЕНИЯ

Выдать код для w

Конец

Если фраза wK уже есть в словаре

Присвоить входной фразе значение Wk

Перейти к Шагу 2

Иначе

Выдать код w

Добавить wK в словарь

Присвоить входной фразе значение К

Перейти к Шагу 2.

Как и в случае с LZ78 для LZW ключевым для размера получаемых кодов является размер словаря во фразах: LZW-коды имеют постоянную длину, равную округленному в большую сторону двоичному логарифму размера словаря.

Пример. Закодировать по алгоритму LZW строку "КРАСНАЯ КРАСКА". Размер словаря — 500 фраз.

ВХОДНАЯ ФРАЗА, wK (В СЛОВАРЬ)	КОД для w	ПОЗИЦИЯ СЛОВАРЯ
ASCII+		0-255
"КР"	0'К'	256
"РА"	0'Р'	257
"АС"	0'А'	258
"СН"	0'С'	259
"НА"	0'Н'	260
"АЯ"	0'А'	261
"Я "	0'Я'	262
" К"	0' '	263
"КРА"	<256>	264
"АСК"	<258>	265
"КА"	0'К'	266
"А"	0'А'	

В этом примере длина полученного кода равна 12 * 9 = 108 битам.

При переполнении словаря, т. е. когда необходимо внести новую фразу в полностью заполненный словарь, из него удаляют либо наиболее редко используемую фразу, либо все фразы, отличающиеся от одиночного символа.

Алгоритм LZW является запатентованным и, таким образом, представляет собой интеллектуальную собственность. Его безлицензионное использование особенно на аппаратном уровне может повлечь за собой неприятности.

Любопытна история патентования LZW. Заявку на LZW подали почти одновременно две фирмы — сначала IBM и затем Unisys, но первой была рассмотрена заявка Unisys, которая и получила патент. Однако, еще до патентования LZW был использован в широко известной в мире Unix программе сжатия данных compress.

15. LZ-алгоритмы распаковки данных. Примеры

1. LZ77, длина словаря — 8 байт (символов). Коды сжатого сообщения — (0,0,'К') (0,0,'Р') (0,0,'А') (0,0,'С) (0,0,'Н') (5,1,'Я') (0,0,' ') (0,4,'К') (0,0,'А').

ВХОДНОЙ КОД	ПЕЧАТЬ	СЛОВАРЬ
<0,0,’К’>	"К"	".......К"
<0,0,’Р’>	"Р"	"......КР"
<0,0,’А’>	"А"	".....КРА"
<0,0,’С’>	"С"	"....КРАС"
<0,0,’Н’>	"Н"	"...КРАСН"
<5,1,’Я’>	"АЯ"	".КРАСНАЯ"
<0,0,’ ’>	" "	"КРАСНАЯ "
<0,4,’К’>	"КРАСК"	"АЯ КРАСК"
<0,0,’А’>	"А"	"Я КРАСКА"

2. LZSS, длина словаря — 8 байт (символов). Коды сжатого сообщения — 0'К' ОТ' О'А' 0'С ’0'Н' 1(5,1) 0'Я' 0' ' 1(0,4) 1(4,1) 1(0,1).

ВХОДНОЙ КОД	ПЕЧАТЬ	СЛОВАРЬ
0 ’К’	"К"	".......К"
0 ’Р’	"Р"	"......КР"
0 ’А’	"А"	".....КРА"
0 ’С’	"С"	"....КРАС"
0 ’Н’	"Н"	"...КРАСН"
1 <5,1>	"А"	"..КРАСНА"
0 ’Я’	"Я"	".КРАСНАЯ"
0 ’ ’	" "	"КРАСНАЯ "
1 <0,4>	"КРАС"	"НАЯ КРАС"
1 <4,1>	"К"	"АЯ КРАСК"
1 <0,1>	"А"	"Я КРАСКА"

3. LZ78, длина словаря — 16 фраз. Коды сжатого сообщения —
(0,'К') (0,'Р') (0,'А') (0,'С) (0,'Н') (3,'Я') (0,' ') (1,'Р') (3,'С) (1,'А').

ВХОДНОЙ КОД	ПЕЧАТЬ (СЛОВАРЬ)	ПОЗИЦИЯ СЛОВАРЯ
	""	0
<0,’К’>	"К"	1
<0,’Р’>	"Р"	2
<0,’А’>	"А"	3
<0,’С’>	"С"	4
<0,’Н’>	"Н"	5
<3,’Я’>	"АЯ"	6
<0,’ ’>	" "	7
<1,’Р’>	"КР"	8
<3,’С’>	"АС"	9
<1,’А’>	"КА"	10

4. LZW, длина словаря — 500 фраз. Коды сжатого сообщения —
0'К' 0'Р' 0'А' 0'С 0'Н' 0'А' 0'Я' 0' ' (256) (258) О'К' О'А'.

При распаковке нужно придерживаться следующего правила. Словарь пополняется после считывания первого символа идущего за текущим кода, т.е. из фразы, соответствующей следующему после раскодированного коду, берется первый символ. Это правило позволяет избежать бесконечного цикла при раскодировании сообщений вида wKwK, где w — фраза, а К — символ. Конкретным примером такого сообщения является любая последовательность трех одинаковых символов, пары которых ранее не встречались.

ВХОДНОЙ КОД	ПЕЧАТЬ	СЛОВАРЬ	ПОЗИЦИЯ СЛОВАРЯ
		ASCII+	0-255
0'К'	"К"	"КР"	256
0'Р'	"Р"	"РА"	257
0'А'	"А"	"АС"	258
0'С'	"С"	"СН"	259
0'Н'	"Н"	"НА"	260
0'А'	"А"	"АЯ"	261
0'Я'	"Я"	"Я "	262
0' '	" "	" К"	263
<256>	"КР"	"КРА"	264
<258>	"АС"	"АСК"	265
0'К'	"К"	"КА"	266
0'А'	"А"		

16. Особенности программ-архиваторов

Если коды алгоритмов типа LZ передать для кодирования (адаптивному) алгоритму Хаффмена или арифметическому, то полученный двухшаговый (конвейерный, а не двухпроходный) алгоритм даст результаты сжатия подобные широко известным программам: GZIP, ARJ, PKZIP, ...

Наибольшую степень сжатия дают двухпроходные алгоритмы, которые исходные данные последовательно сжимают два раза, но они работают до двух раз медленнее однопроходных при незначительном увеличении степени сжатия.

Большинство программ-архиваторов сжимает каждый файл по отдельности, но некоторые сжимают файлы в общем потоке, что дает увеличение степени сжатия, но одновременно усложняет способы работы с полученным архивом, например, замена в таком архиве файла на его более новую версию может потребовать перекодирования всего архива. Примером программы, имеющей возможность сжимать файлы в общем потоке, является RAR. Архиваторы ОС Unix (gzip, bzip2, ...*)* сжимают файлы в общем потоке практически всегда.

В 1992 году фирма WEB Technologies объявила о выходе новой программы сжатия DataFiles/16, которая якобы может при неоднократном использовании сжать любое количество данных до 1024 байт. Информация об этом прошла из солидного издания, журнала Byte.

Конечно же никакой алгоритм сжатия не может уплотнить произвольные данные. Для доказательства этого проделаем следующий мысленный эксперимент. Предположим, что на жестком диске компьютера хранятся все возможные разные файлы длиной ровно 100 байт (таких файлов будет всего 2^{800}). И пусть существует

идеальная программа сжатия данных, которая сожмет каждый из них хотя бы на один байт. Но тогда, так как всего разных файлов длиной меньшей 100 байт существует не более чем $1 + 2^8 + 2^{16} + \cdots + 2^{792} = (2^{800} - 1)/255 < 2^{800}$, то неизбежно получится, что два разных файла упакуются в идентичные файлы. Следовательно, не может существовать программы сжатия данных, которая может сжать любые исходные данные.

Формат файла, содержащего данные, которые перед использованием требуется распаковать соответствующей программой-архиватором, как правило, может быть идентифицирован расширением имени файла.

В следующей таблице приводятся некоторые типичные расширения, соответствующие им программы-архиваторы и методы сжатия данных.

Расширения	Программы	Тип кодирования
Z	compress	LZW
arc	arc, pkarc	LZW, Хаффмена
zip	zip, unzip, pkzip, pkunzip	LZW, LZ77, Хаффмена, Шеннона-Фэно
gz	gzip	LZ77, Хаффмена
bz2	bzip2	Берроуза-Уиллера, Хаффмена
arj	arj	LZ77, Хаффмена
ice, lzh	lha, lharc	LZSS, Хаффмена
pak	pak	LZW

Практически все форматы файлов для хранения графической информации используют сжатие данных. Формат графического файла также, как правила, идентифицируется расширением имени файла.

В следующей таблице приводятся некоторые типичные расширения графических файлов и соответствующие им методы сжатия данных.

Расширения	Тип кодирования
gif	LZW
jpeg, jpg	сжатие с потерями, Хаффмена или арифметическое
bmp, pcx	RLE
tiff, tif	CCITT/3 для факсов, LZW или другие

Сжатие RLE (Run Length Encoding — кодирование переменной длины) — это простейший метод сжатия, в общем случае очень неэффективный, но дающий неплохие результаты на типичной графической информации. Оно основано в основном на выделении специального кода-маркера, указывающего сколько раз повторить следующий байт.

Сжатие и распаковка в реальном времени используется в программах-драйверах для "уплотнения" носителей информации, позволяющих увеличить емкость носителя приблизительно в 2 раза. Наиболее известной программой такого рода является DriverSpace для MS-DOS и Microsoft Windows.

17. **Сжатие информации с потерями**

Все ранее рассмотренные алгоритмы сжатия информации обеспечивали возможность полного восстановления исходных данных. Но иногда для повышения степени сжатия можно отбрасывать часть исходной информации, т. е. производить

сжатие с потерями. Естественно, что такое сжатие нельзя проводить, например, на финансовой базе данных банка. Но в тех случаях, когда сжимается информация, используемая лишь для качественной оценки (это, как правило, аналоговая информация), сжатие с потерями является очень подходящим.

Сжатие с потерями используется в основном для трех видов данных: полноцветная графика ($2^{24} \approx 16$ млн. цветов), звук и видеоинформация.

Сжатие с потерями обычно проходит в два этапа. На первом из них исходная информация приводится (с потерями) к виду, в котором ее можно эффективно сжимать алгоритмами 2-го этапа сжатия без потерь.

Основная идея сжатия графической информации с потерями заключается в следующем. Каждая точка в картинке характеризуется тремя равноважными атрибутами: яркостью, цветом и насыщенностью. Но глаз человека воспринимает эти атрибуты не как равные. Глаз воспринимает полностью только информацию о яркости и в гораздо меньшей степени о цвете и насыщенности, что позволяет отбрасывать часть информации о двух последних атрибутах без потери качества изображения. Это свойство зрения используется, в частности, в цветном телевизоре, в котором на базовое черно-белое изображение наносят цветовую раскраску.

Для сжатия графической информации с потерями в конце 1980-х установлен один стандарт — формат JPEG (Joint Photographic Experts Group — название объединения его разработчиков). В этом формате можно регулировать степень сжатия, задавая степень потери качества.

Сжатие видеоинформации основано на том, что при переходе от одного кадра фильма к другому на экране обычно почти ничего не меняется. Таким образом, сжатая видеоинформация представляет собой запись некоторых базовых кадров и последовательности изменений в них. При этом часть информации может отбрасываться. Сжатую подобным образом информацию можно далее сжимать и другими методами. Хотя существует не один стандарт для сжатия видеоданных, наиболее распространенными являются стандарты MPEG (Motion Picture Experts Group), первый из которых был опубликован в 1988 году. MPEG — практически единственный стандарт для записи видео и звуковой информации на CD-ROM, DVD-ROM и в цифровом спутниковом телевидении. Видеоинформацию можно сжать необыкновенно плотно, до 100 и более раз, что позволяет, например, на одну видеокассету, записать более ста различных художественных фильмов. Но из-за очень сложных проблем, связанных с правами на интеллектуальную собственность, реально возможности сжатия информации таким образом используются сравнительно редко.

Для сжатии звуковой информации с потерями существует несколько стандартов. Наиболее широко используемый из них — это MPEG без видеоданных. Стандарт LPC (Linear Predictive Coding) используется для сжатия речи. Алгоритм LPC пытается промоделировать речевой тракт человека и выдает на выходе буквально текущее состояние участвующих в формировании звуков органов.

18. **Информационный канал**

Канал информационный — это совокупность устройств, объединенных линиями связи, предназначенных для передачи информации от источника информации *(начального устройства канала)* до ее приемника *(конечного устройства канала)*.

Линии связи обеспечивают прохождение информационных сигналов между устройствами канала. Информация обычно передается при помощи электрического тока (по проводам), света (по оптоволокну), электромагнитных волн радиодиапазона (в пространстве) и, редко, звука (в плотной среде: атмосфере, воде и т.п.) и прочих.

Устройства канала — это, как правило, *репитеры,* просто передающие усиленным принятый сигнал (пример, радиорелейные линии). К устройствам канала иногда относят и кодеры/декодеры, но в только тех случаях, когда кодирование/декодирование происходит с высокой скоростью, не требующей ее специального учета, как замедляющего фактора; обычно же кодеры/декодеры относят к источникам или приемникам информации.

Технические характеристики канала определяются принципом действия входящих в него устройств, видом сигнала, свойствами и составом физической среды, в которой распространяются сигналы, свойствами применяемого кода.

Эффективность канала характеризуется скоростью и достоверностью передачи информации, надежностью работы устройств и задержкой сигнала во времени.

Задержка сигнала во времени — это интервал времени от отправки сигнала передатчиком до его приема приемником.

Математически канал задается множеством допустимых сообщений на входе, множеством допустимых сообщений на выходе и набором условных вероятностей *P(y/x)* получения сигнала *y* на выходе при входном сигнале *x.* Условные вероятности описывают статистические свойства "шумов" (или помех), искажающих сигнал в процессе передачи. В случае, когда *P(y/x)* = 1 при $y = x$ и *P(y/x)* = 0 при $y \neq x$, канал называется *каналом без "шумов"*. В соответствии со структурой входных и выходных сигналов выделяют дискретные и непрерывные каналы. В дискретных каналах сигналы на входе и выходе представляют собой последовательность символов одного или двух (по одному для входа и выхода) алфавитов. В непрерывных каналах входной и выходной сигналы представляют собой функции от непрерывного параметра-времени. Бывают также смешанные или гибридные каналы, но тогда обычно рассматривают их дискретные и непрерывные компоненты раздельно. Далее рассматриваются только дискретные каналы.

Способность канала передавать информацию характеризуется числом — *пропускной способностью* или *емкостью канала* (обозначение -C).

Для случая канала без шума формула расчета емкости канала имеет вид $C = \lim\limits_{T\to\infty} \frac{\log_2 N(T)}{T}$, где *N(T)* — число всех возможных сигналов за время *T*.

Пример. Пусть алфавит канала без "шумов" состоит из двух символов — 0 и 1, длительность τ секунд каждый. За время *T* успеет пройти $n = T/\tau$ сигналов, всего возможны 2^n различных сообщений длиной *n*.

В этом случае $C = \lim\limits_{T\to\infty} \frac{\log_2 2^{T/\tau}}{T} = 1/\tau$ бод.

На рис. 8 приведена схема, на которой изображен процесс прохождения информации по каналу с описанными в примере характеристиками.

Здесь для кодирования используется уровень сигнала: низкий для 0 и высокий для 1. Недостатки этого способа проявляются в случаях, когда нужно передавать много сплошных нулей или единиц. Малейшее рассогласование синхронизации

между приемником и передатчиком приводит тогда к неисправимым ошибкам. Кроме того, многие носители информации, в частности, магнитные, не могут поддерживать длительный постоянный уровень сигнала.

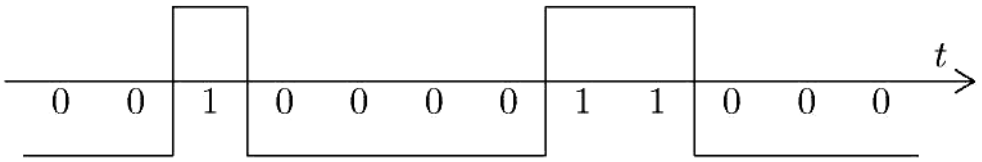


Рис. 8

Для передачи информации используется обычно другой способ, когда для представления 0 и 1 используются две разные частоты, отличающиеся друг от друга ровно в два раза (см. рис. 9) — это так называемая *частотная модуляция* (ЧМ или FM).

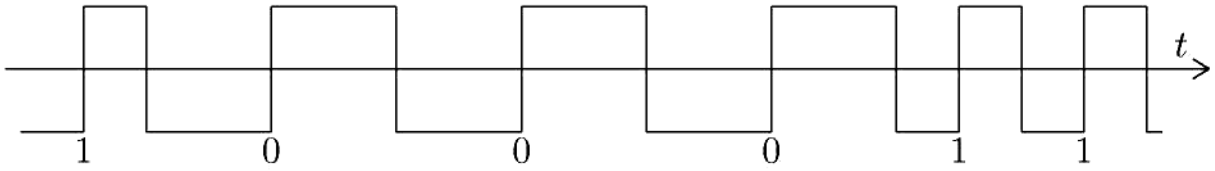


Рис. 9

Таким образом, при таком кодировании, если сигнал 1 имеет длительность τ, то 0 — 2τ.

Рассчитаем емкость этого канала. Нужно рассчитать *N(T).* Пусть $n = T/\tau$, тогда получается, что нужно рассчитать сколькими способами можно разбить отрезок длины n отрезками длины 2 и 1. Получаем, что $N(T) = S_n = C_n^n + C_{n-1}^{n-2} + C_{n-2}^{n-4} + \cdots$, где первое слагаемое — это количество способов, которыми можно разбить отрезок длины n n отрезками длины 1, второе слагаемое — это количество способов, которыми можно разбить отрезок длины n $(n$ — 2) отрезками длины 1 и одним отрезком длины 2, третье слагаемое — это количество способов, которыми можно разбить отрезок длины n $(n$ — 4) отрезками длины 1 и двумя отрезками длины 2 и т.д. Таким образом, $S_1 = 1$. Вследствие того, что $C_m^k + C_m^{k+1} = C_{m+1}^{k+1}$ для любых $k < m$, получается, что

$$\begin{array}{rcccccccc} S_{n-1} & = & & C_{n-1}^{n-1} & + & C_{n-2}^{n-3} & + & C_{n-3}^{n-5} & + \cdots \\ S_n & = & C_n^n & + & C_{n-1}^{n-2} & + & C_{n-2}^{n-4} & + & C_{n-3}^{n-6} & + \cdots, \\ S_{n+1} & = & C_{n+1}^{n+1} & + & C_n^{n-1} & + & C_{n-1}^{n-3} & + & C_{n-2}^{n-5} & + \cdots \end{array}$$

т.е. $S_{n+1} = S_n + S_{n-1}$ при $n > 1$. Если положить, что $S_0 = 1$, то S_0, S_1,... — это последовательность 1, 1, 2, 3, 5, 8, 13, 21, 34,..., т.е. числа Фибоначчи. С XIX века для вычисления n-го члена последовательности Фибоначчи известна формула

$$S_n = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right).$$

Таким образом,

$$C = \lim_{T\to\infty} \frac{\log_2 N(T)}{T} = \lim_{n\to\infty} \frac{\log_2 S_n}{n\tau} = \frac{\log_2 \frac{1+\sqrt{5}}{2}}{\tau} \approx 0.69/\tau \text{ бод.}$$

При использовании частотной модуляции на практике нули, как правило, кодируются в два раза плотнее. Это достигается тем, что учитываются не уровни сигнала, а смена уровня (полярности). Если частота ν соответствует 1, то с частотой 2ν производится проверка уровня сигнала. Если он меняется, то это сигнал 1, если

нет, то — 0. На практике частота v — это частота синхронизации, т. е. частота импульса, который независимо от данных меняет полярность сигнала. 0 не генерирует импульса смены полярности, а 1 генерирует (см. рис. 10).

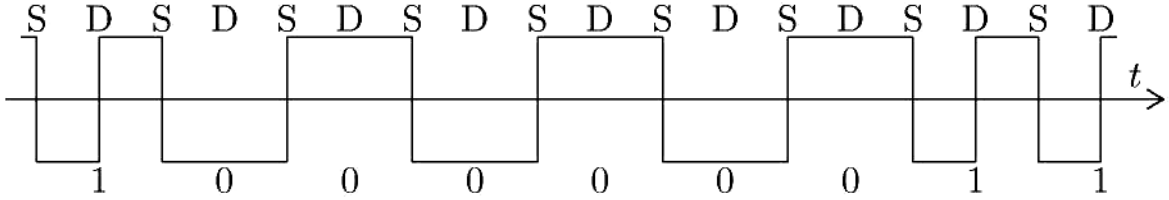


Рис. 10

Для записи информации на первые магнитные диски и ленты использовался метод FM. На гибкие диски 5.25" и 3.5" информация записывается методом MFM (Modified FM) — модификацией метода FM, позволяющей в 2 раза повысить плотность записи. Это достигается тем, что частота синхронизации увеличивается вдвое. MFM можно использовать с теми же физическими каналами, что и FM, потому что импульсы синхронизации не передаются перед 1 и первым 0 в серии нулей (см. рис. 11).

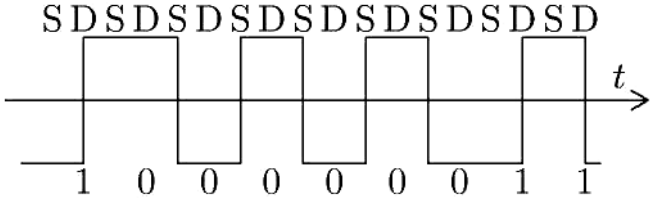


Рис. 11

Метод записи с групповым кодированием, RLL — Run Limited Length, не использует импульсы синхронизации, применяется, в частности, в жестких дисках "винчестер" и существует в нескольких разновидностях. Одна из них основана на замене тетрад байта на 5-битные группы. Эти группы подбираются таким образом, чтобы при передаче данных нули не встречались подряд более двух раз, что делает код самосинхронизирующимся. Например, тетрада 0000 заменяется группой бит 11001, тетрада 1000 — 11010, тетрада 0001 — 11011, тетрада 1111 — 01111 (см. рис. 12). Существуют разновидности RLL, в которых заменяются последовательности бит различной длины. Кодирование MFM или FM можно представить как частный случай RLL.

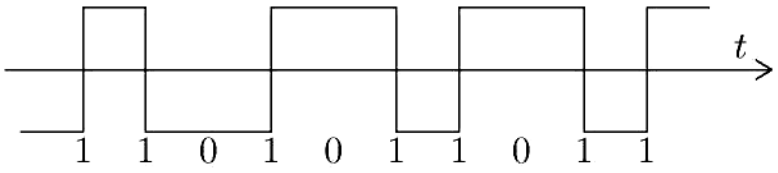


Рис. 12

При необходимости передачи записанных с помощью некоторого кода сообщений по данному каналу приходиться преобразовывать эти сообщения в допустимые сигналы канала, т. е. производить надлежащее кодирование, а при приеме данных — декодирование. Кодирование целесообразно производить так, чтобы среднее время, затрачиваемое на передачу, было как можно меньше. Получается, что исходному входному алфавиту нужно однозначно сопоставить новый алфавит, обеспечивающий большую скорость передачи.

Следующий, *основной факт теории передачи информации* или *основная теорема о кодировании при наличии помех* позволяет при знании емкости канала и

энтропии передатчика вычислить максимальную скорость передачи данных в канале.

Теорема Шеннона. Пусть источник характеризуется д.с.в. *X.* Рассматривается канал с шумом, т.е. для каждого передаваемого сообщения задана вероятность ε его искажения в процессе передачи (вероятность ошибки). Тогда существует такая скорость передачи *u,* зависящая только от *X*, что $\forall \varepsilon > 0 \ \exists\, u' < u$ сколь угодно близкая к *u* такая, что существует способ передавать значения *X* со скоростью *u′* и с вероятностью ошибки меньшей ε, причем *u = C/HX.* Упомянутый способ образует помехоустойчивый код.

Кроме того, Фэно доказана следующая *обратная теорема о кодировании при наличии помех.* Для *u′ > u* можно найти такое положительное число ε, что в случае передачи информации по линии связи со скоростью *u′* вероятность ошибки ε передачи каждого символа сообщения при любом методе кодирования и декодирования будет не меньше ε *(*ε очевидно растет вслед за ростом *u′).*

19. **Помехозащитное кодирование**

Простейший код для борьбы с шумом — это контроль четности, он, в частности, широко используется в модемах. Кодирование заключается в добавлении к каждому байту девятого бита таким образом, чтобы дополнить количество единиц в байте до заранее выбранного для кода четного (even) или нечетного (odd) значения. Используя этот код, можно лишь обнаруживать большинство ошибок.

Простейший код, исправляющий ошибки, — это тройное повторение каждого бита. Если с ошибкой произойдет передача одного бита из трех, то ошибка будет исправлена, но если случится двойная или тройная ошибка, то будут получены неправильные данные. Часто коды для исправления ошибок используют совместно с кодами для обнаружения ошибок. При тройном повторении для повышения надежности три бита располагают не подряд, а на фиксированном расстоянии друг от друга. Использование тройного повторения значительно снижает скорость передачи данных.

Двоичный симметричный канал изображен на рис. 13, где *p* — это вероятность безошибочной передачи бита, а *q* — вероятность передачи бита с ошибкой. Предполагается, что в таком канале ошибки происходят независимо. Далее рассматриваются только такие каналы.

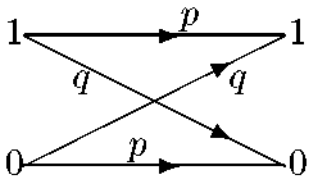


Рис. 13

Двоичный симметричный канал реализует схему Бернулли, поэтому вероятность передачи *n* бит по двоичному симметричному каналу с *k* ошибками равна $P_n(k) = C_n^k p^{n-k} q^k$.

Пример. Вероятность передачи одного бита информации с ошибкой равна *q* = 0.01 и нас интересует вероятность безошибочной передачи 1000 бит (125 байт). Искомую вероятность можно подсчитать по формуле $P_{1000}(0) = C_{1000}^0 p^{1000} q^0 = 0.99^{1000} \approx 4.32 * 10^{-5}$, т.е. она ничтожно мала.

Добиться минимальности вероятности ошибки при передаче данных можно используя специальные коды. Обычно используют *систематические* помехозащитные коды. Идея систематических кодов состоит в добавлении к символам исходных кодов, предназначенных для передачи в канале, нескольких контрольных символов по определенной схеме кодирования. Принятая такая

удлиненная последовательность кодов декодируется по схеме декодирования в первоначально переданную. Приемник способен распознавать и / или исправлять ошибки, вызванные шумом, анализируя дополнительную информацию, содержащуюся в удлиненных кодах.

Двоичным (m,n)-кодом называется пара, состоящая из схемы кодирования *E:* $\mathbb{Z}_2^m \to \mathbb{Z}_2^n$ и схемы декодирования *D:* $\mathbb{Z}_2^n \to \mathbb{Z}_2^m$, где $\mathbb{Z}_2^n$ — множество всех двоичных последовательностей длины *n*, *m* < *n* (случай *m* = *n* рассматривается в криптографии).

Функции *D* и *E* выбираются так, чтобы функция $H = D \circ T \circ E$, где *T* — *функция ошибок,* с вероятностью, близкой к единице, была тождественной. Функции *D* и *E* считаются безошибочными, т. е. функция *D* ∘ E — тождественная (см. рис. 14).

	E		T		D	
Исходное сообщение	⟶	Кодированное сообщение	⟶	Двоичный симметричный канал	⟶	Принятое сообщение	⟶	Декодированное сообщение

Рис. 14

20. Математическая модель системы связи

Коды делятся на два больших класса. Коды с исправлением ошибок имеют целью восстановить с вероятностью, близкой к единице, посланное сообщение. Коды с обнаружением ошибок имеют целью выявить с вероятностью, близкой к единице, наличие ошибок.

Простой код с обнаружением ошибок основан на схеме проверки четности, применимой к сообщениям $a_1 \ldots a_m$ любой фиксированной длины *m*. Схема кодирования определяется следующими формулами:

$$E(a_1 \ldots a_m) = a_1 \ldots a_m a_{m+1},$$

$$a_{m+1} = \begin{cases} 0, & \text{если } \sum_{i=1}^{m} a_i \text{ — четная;} \\ 1, & \text{если } \sum_{i=1}^{m} a_i \text{ — нечетная.} \end{cases}$$

Таким образом, $\sum_{i=1}^{m+1} a_i$ должна быть четной.

Соответствующая схема декодирования тривиальна:

$$D(a_1 \ldots a_m a_{m+1}) = \begin{cases} a_1 \ldots a_m, & \text{если } \sum_{i=1}^{m+1} a_i \text{ — четна;} \\ \langle \textit{ошибка} \rangle, & \text{если } \sum_{i=1}^{m+1} a_i \text{ — нечетна.} \end{cases}$$

Разумеется, что четность $\sum_{i=1}^{m+1} a_i$ не гарантирует безошибочной передачи.

Пример. Проверка четности при *m* = 2 реализуется следующим кодом (функцией *E):* 00 → 000, 01 → 011, 10 → 101, 11 → 110. В двоичном симметричном канале доля неверно принятых сообщений для этого кода (хотя бы с одной ошибкой) равна $q^3 + 3pq^2 + 3p^2q$ (три, две или одна ошибка соответственно). Из них незамеченными окажутся только ошибки точно в двух битах, не изменяющие четности. Вероятность таких ошибок $3pq^2$. Вероятность ошибочной передачи сообщения из двух бит равна $2pq + q^2$. При малых *q* верно, что $3pq^2 \ll 2pq + q^2$.

Рассмотрим (*m, 3m*)-код с *тройным повторением.* Коды с повторениями очень неэффективны, но полезны в качестве теоретического примера кодов, исправляющих ошибки. Любое сообщение разбивается на блоки длиной *m* каждое и каждый блок передается трижды — это определяет функцию *E.* Функция *D* определяется следующим образом. Принятая строка разбивается на блоки длиной *3m.* Бит с номером *i* (1 << *i* << *m)* в декодированном блоке получается из анализа битов с номерами *i, i+m, i+2m* в полученном блоке: берется тот бит из трех, который встречается не менее двух раз. Вероятность того, что бит в данной позиции будет принят трижды правильно равна p^3. Вероятность одной ошибки в тройке равна $3p^2q$. Поэтому вероятность правильного приема одного бита равна $p^3 + 3p^2q$. Аналогичным образом получается, что вероятность приема ошибочного бита равна $q^3 + 3pq^2$.

Пример. Предположим $q = 0.1$. Тогда вероятность ошибки при передачи одного бита — 0.028, т.е. этот код снижает вероятность ошибки с 10% до 2.8%. Подобным образом организованная передача с пятикратным повторением даст вероятность ошибки на бит

$$q^5 + 5pq^4 + 10p^2q^3 = 0.00856 = 0.856\%,$$

т.е. менее 1%. В результате вероятность правильной передачи строки длиной 10 возрастет с $0.9^{10} \approx 35\%$ до $0.972^{10} \approx 75\%$ при тройных повторениях и до $0.99144^{10} \approx 92\%$ при пятикратных повторениях.

Тройное повторение обеспечивает исправление одной ошибки в каждой позиции за счет трехкратного увеличения времени передачи.

Рассмотрим (2048, 2313)-код, используемый при записи данных на магнитофонную ленту компьютерами Apple II. К каждому байту исходных данных прибавляется бит четности и, кроме того, после каждых таких расширенных битом четности 256 байт добавляется специальный байт, также расширенный битом четности. Этот специальный байт, который называют *контрольной суммой* (check sum), есть результат применения поразрядной логической операции "исключающее ИЛИ" (XOR) к 256 предшествующим расширенным байтам. Этот код способен как обнаруживать ошибки нечетной кратности в каждом из отдельных байт, так и исправлять до 8 ошибок в блоке длиной 256 байт. Исправление ошибок основано на том, что если в одном из бит одного из байт 256 байтового блока произойдет сбой, обнаруживаемый проверкой четности, то этот же сбой проявится и в том, что результат операции "исключающее ИЛИ" над всеми соответствующими битами блока не будет соответствовать соответствующему биту контрольной суммы. Сбойный бит однозначно определяется пересечением сбойных колонки байта и строки бита контрольной суммы. На рис. 15 изображена схема участка ленты, содержащего ровно 9 ошибок в позициях, обозначенных $p_1, p_2, \ldots, p_9$. Расширенный байт контрольной суммы обозначен CS, а бит паритета (в данном случае четности) — PB (parity bit). Ошибка в позиции p_1 может быть исправлена. Ошибки в позициях p_4, p_5, p_6, p_7 можно обнаружить, но не исправить. Ошибки в позициях p_2, p_3, p_8, p_9 невозможно даже обнаружить.

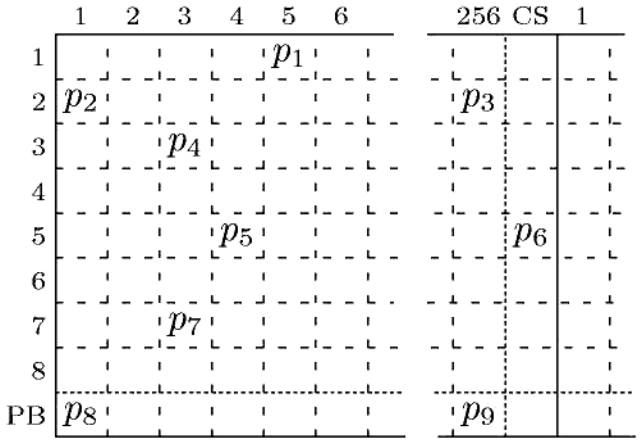


Рис. 15

Приведенные ранее примеры простейших кодов принадлежат к классу *блочных.* По определению, блочный код заменяет каждый блок из *m* символов более длинным блоком из *n* символов. Следовательно, (*m, n*)-коды являются блочными. Существуют также *древовидные* или *последовательные коды,* в которых значение очередного контрольного символа зависит от всего предшествующего фрагмента сообщения. Работа с древовидным шумозащитным кодом имеет сходство с работой с арифметическим кодом для сжатия информации.

Расстоянием (Хэмминга) между двоичными словами длины *n* называется количество позиций, в которых эти слова различаются. Это одно из ключевых понятий теории кодирования. Если обозначить двоичные слова как $a = a_1 \ldots a_n$ и $b = b_1 \ldots b_n$ то расстояние между ними обозначается *d(a, b).*

Весом двоичного слова $a = a_1 \ldots a_n$ называется количество единиц в нем. Обозначение *w(a).* Можно сказать, что $w(a) = \sum_{i=1}^{n} a_i$.

Пример. Пусть $a = 1001$ и $b = 0011$, тогда $w(a) = w(b) = 2$, $d(a, b) = 2$.

Далее операция + при применении к двоичным словам будет означать поразрядное сложение без переноса, т. е. сложение по модулю 2 или "исключающее ИЛИ" (XOR).

Расстояние между двоичными словами *a* и *b* равно весу их поразрядной суммы, т.е. $d(a, b) = w(a + b)$.

Если два слова различаются в каком-либо разряде, то это добавит единицу к весу их поразрядной суммы.

Следовательно, если *a* и *b* — слова длины *n*, то вероятность того, что слово *a* будет принято как *b*, равна $p^{n-d(a,b)}q^{d(a,b)}$.

Наример, вероятность того, что слово 1011 будет принято как 0011, равна p^3q.

Для возможности обнаружения ошибки в одной позиции минимальное расстояние между словами кода должно быть большим 1.

Иначе ошибка в одной позиции сможет превратить одно кодовое слово в другое, что не даст ее обнаружить.

Для того, чтобы код давал возможность обнаруживать все ошибки кратности, не большей *k,* необходимо и достаточно, чтобы наименьшее расстояние между его словами было $k + 1$.

Достаточность доказывается конструктивно: если условие утверждения выполнено для *E,* то в качестве декодирующей функции *D* следует взять функцию, сообщающую об ошибке, если декодируемое слово отличается от любого из слов из образа *E.*

Необходимость доказывается от противного: если минимальное расстояние $k' < k + 1$, то ошибка в k' позициях сможет превратить одно кодовое слово в другое.

Для такого кода вероятность того, что ошибки в сообщении останутся необнаруженными, равна

$$\sum_{i=k+1}^{n} C_n^i p^{n-i} q^i = C_n^{k+1} p^{n-k-1} q^{k+1} + \cdots + C_n^{n-1} p q^{n-1} + q^n \approx$$

$$\approx [\text{при малых } q \text{ и не слишком маленьких } k] \approx C_n^{k+1} p^{n-k-1} q^{k+1}.$$

Для того, чтобы код давал возможность исправлять все ошибки кратности, не большей k, необходимо и достаточно, чтобы наименьшее расстояние между его словами было $2k + 1$.

Достаточность доказывается конструктивно: если условие утверждения выполнено для E, то в качестве декодирующей функции D следует взять функцию, возвращающую ближайшее к декодируемому слово из образа E. Необходимость доказывается от противного. Пусть расстояние между выбранными словами в коде равно $2k$. Тогда если при передаче каждого из этих слов случится k ошибок, которые изменят биты, в которых различаются эти слова, то приемник получит два идентичных сообщения, что свидетельствует о том, что в данной ситуации исправление k ошибок невозможно. Следовательно, минимальное расстояние между словами кода должно быть большим $2k$.

Пример. Рассмотрим $(1,3)$-код, состоящий из E, задающей отображение $0 \to 000$ и $1 \to 111$, и D, задающей отображение $000 \to 0$, $001 \to 0$, $010 \to 0$, $011 \to 1$, $100 \to 0$, $101 \to 1$, $110 \to 1$, $111 \to 1$. Этот код (с тройным повторением) исправляет ошибки в одной позиции, т.к. минимальное расстояние между словами кода равно 3.

Если код исправляет все ошибки кратности k и меньшей, то вероятность ошибочного приема слова длины n очевидно не превосходит $\sum_{i=k+1}^{n} C_n^i p^{n-i} q^i$. Вероятность правильного приема в этом случае не меньше, чем

$$\sum_{i=0}^{k} C_n^i p^{n-i} q^i = p^n + C_n^1 p^{n-1} q + \cdots + C_n^k p^{n-k} q^k.$$

Передачу данных часто удобно рассматривать следующим образом. Исходное сообщение $a = a_1 \ldots a_m$ кодируется функцией E в кодовое слово $b = b_1 \ldots b_n$. Канал связи при передаче добавляет к нему функцией T *строку ошибок* $e = e_1 \ldots e_n$ так, что приемник получает сообщение $r = r_1 \ldots r_n$, где $r_i = b_i + e_i$ $(1 \le i \le n)$. Система, исправляющая ошибки, переводит r в некоторое (обычно ближайшее) кодовое слово. Система, только обнаруживающая ошибки, лишь проверяет, является ли принятое слово кодовым, и сигнализирует о наличии ошибки, если это не так.

Пример. Пусть передаваемое слово $a = 01$ кодируется словом $b = 0110$, а строка ошибок — $e = 0010$. Тогда будет принято слово $r = 0100$. Система, исправляющая ошибки, переведет его в 0110 и затем восстановит переданное слово 01.

Если система только обнаруживает ошибки и расстояние между любыми кодовыми словами $k \ge 2$, то любая строка ошибок e с единственной единицей приведет к слову $r = b + e$, которое не является кодовым.

Пример. Рассмотрим $(2, 3)$-код с проверкой четности. Множество кодовых слов — {000,011,101, 110}. Ни одна из строк ошибок 001, 010, 100, 111 не переводит

одно кодовое слово в другое. Поэтому однократная и тройная ошибки могут быть обнаружены.

Пример. Следующий (*2, 5*)-код обнаруживает две ошибки:

$$a_1 = 00 \to 00000 = b_1, \qquad a_2 = 01 \to 01011 = b_2,$$

$$a_3 = 10 \to 10101 = b_3, \qquad a_4 = 11 \to 11110 = b_4.$$

Этот же код способен исправлять однократную ошибку, потому что любые два кодовых слова отличаются по меньшей мере в трех позициях. Из того, что $d(b_i,b_j) \geq 3$ при $i \neq j$, следует, что однократная ошибка приведет к приему слова, которое находится на расстоянии 1 от кодового слова, которое было передано. Поэтому схема декодирования, состоящая в том, что принятое слово переводится в ближайшее к нему кодовое, будет исправлять однократную ошибку. В двоичном симметричном канале вероятность правильной передачи одного блока будет не меньше чем $p^5 + 5p^4q$.

Установлено, что в ($n - r$, n)-коде, минимальное расстояние между кодовыми словами которого $2k + 1$, числа n, r (число дополнительных разрядов в кодовых словах) и k должны соответствовать неравенству

$$r \geqslant \log_2(C_n^k + C_n^{k-1} + \cdots + C_n^1 + 1),$$

называемому *неравенством* или *нижней границей Хэмминга.* Кроме того, если числа n, r и k соответствуют неравенству

$$r > \log_2(C_{n-1}^{2k-1} + C_{n-1}^{2k-2} + \cdots + C_{n-1}^1 + 1),$$

называемому *неравенством* или *верхней границей Варшамова- Гильберта,* то существует ($n - r$, n)-код, исправляющий все ошибки веса k и менее.

Нижняя граница задает необходимое условие для помехозащитного кода с заданными характеристиками, т.е. любой такой код должен ему соответствовать, но не всегда можно построить код по подобранным, удовлетворяющим условию характеристикам. Верхняя граница задает достаточное условие для существования помехозащитного кода с заданными характеристиками, т. е. по любым подобранным, удовлетворяющим условию характеристикам можно построить им соответствующий код.

21. Матричное кодирование

Ранее каждая схема кодирования описывалась таблицами, задающими кодовое слово длины n для каждого исходного слова длины m. Для блоков большой длины этот способ требует большого объема памяти и поэтому непрактичен. Например, для (*16, 33*)-кода потребуется $33 * 2^{16} = 2\,162\,688$ бит.

Гораздо меньшего объема памяти требует *матричное кодирование.* Пусть E матрица размерности $m \times n$, состоящая из элементов e_{ij}, где i — это номер строки, а j — номер столбца. Каждый из элементов матрицы e_{ij} может быть либо 0, либо 1. Кодирование реализуется операцией $b = aE$ или $b_j = a_1e_{1j} + a_2e_{2j} + \cdots + a_me_{mj}$, где кодовые слова рассматриваются как векторы, т.е как матрицы-строки размера $1 \times n$.

Пример. Рассмотрим следующую 3×6-матрицу:

$$E = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Тогда кодирование задается такими отображениями: 000 → 000000, 001 → 001111, 010 → 010011, 011 → 011100, 100 → 100110, 101 → 101001, 110 → 110101, 111 → 111010.

Рассмотренный пример показывает преимущества матричного кодирования: достаточно запомнить m кодовых слов вместо 2^m слов. Это общий факт.

Кодирование не должно приписывать одно и то же кодовое слово разным исходным сообщениям. Простой способ добиться этого состоит в том, чтобы m столбцов (в предыдущем примере — первых) матрицы E образовывали единичную матрицу. При умножении любого вектора на единичную матрицу получается этот же самый вектор, следовательно, разным векторам-сообщениям будут соответствовать разные вектора систематического кода.

Матричные коды называют также *линейными кодами.* Для линейных $(n - r, n)$-кодов с минимальным расстоянием Хэмминга d существует *нижняя граница Плоткина* (Plotkin) для минимального количества контрольных разрядов r при $n \geq 2d - 1$,

$$r \geqslant 2d - 2 - \log_2 d.$$

22. Групповые коды

Множество всех двоичных слов $a = a_1 \ldots a_m$ длины m образует абелеву (коммутативную) группу относительно поразрядного сложения.

Пусть E — кодирующая $m \times n$-матрица, у которой есть $m \times m$-подматрица с отличным от нуля определителем, например, единичная. Тогда отображение $a \to a\,E$ переводит группу всех двоичных слов длины m в группу кодовых слов длины n.

Предположим, что $a = a_1 \ldots a_m = a' + a''$. Тогда для $b = b_1 \cdots b_n = aE$, $b' = a'E$, $b'' = a''E$, получаем

$$b_j = a_1 e_{1j} + a_2 e_{2j} + \cdots + a_m e_{mj} =$$

$$= (a'_1 + a'_2)e_{1j} + (a'_2 + a''_2)e_{2j} + \cdots + (a'_m + a''_m)e_{mj} = b'_j + b''_j,$$

т. е. $b = b' + b''$. Следовательно, взаимно-однозначное отображение группы двоичных слов длины m при помощи заданной матрицы E сохраняет свойства групповой операции, что означает, что кодовые слова образуют группу.

Блочный код называется *групповым,* если его кодовые слова образуют группу.

Если код является групповым, то наименьшее расстояние между двумя кодовыми словами равно наименьшему весу ненулевого слова.

Это следует из соотношения $d(b_i, b_j) = w(b_i + b_j)$.

В предыдущем примере наименьший вес ненулевого слова равен 3. Следовательно, этот код способен исправлять однократную ошибку или обнаруживать однократную и двойную.

При использовании группового кода незамеченными остаются те и только те ошибки, которые отвечают строкам ошибок, в точности равным кодовым словам.

Такие строки ошибок переводят одно кодовое слово в другое.

Следовательно, вероятность того, что ошибка останется необнаруженной, равна сумме вероятностей всех строк ошибок, равных кодовым словам.

В рассмотренном примере вероятность ошибки равна $4p^3q^3 + 3p^2q^4$.

Рассмотрим задачу оптимизации декодирования группового кода с двоичной матрицей кодирования *E*. Требуется минимизировать вероятность того, что $D(T(aE)) \neq a$.

Схема декодирования состоит из группы *G* всех слов, которые могут быть приняты $(\#G = 2^n)$. Так как кодовые слова *B* образуют нормальную (нормальность следует из коммутативности *G)* подгруппу G, то множеству *G* можно придать структуру таблицы: будем записывать в одну строку те элементы *G*, которые являются членами одного смежного класса *G* по *B*. Первая строка, соответствующая нулевому слову из *G*, будет тогда всеми кодовыми словами из *B,* т.е. $b_0, b_1, \ldots, b_{2^m-1}$. В общем случае, если $g_i \in G$, то строка, содержащая g_i (смежный класс g_i *B)* имеет вид $b_0 + g_i, b_1 + g_i, \ldots, b_{2^m-1} + g_i$.

Лидером каждого из таких построенных смежных классов называется слово минимального веса.

Каждый элемент g из *G* однозначно представляется в виде суммы $g_i + b_j$, где $g_i \in G$ — лидер соответствующего смежного класса и $b_j \in B$.

Множество классов смежности группы образуют фактор-группу, которая есть фактор-множество множества *G* по отношению эквивалентности-принадлежности к одному смежному классу, а это означает, что множества, составляющие это фактор-множество, образуют разбиение *G*. Отсюда следует, что строки построенной таблицы попарно либо не пересекаются, либо совпадают.

Если в рассматриваемой таблице в первом столбце записать лидеры, то полученная таблица называется *таблицей декодирования.* Она имеет вид:

$$\begin{array}{lllll}
b_0 & b_1 & b_2 & \cdots & b_{2^m-1} \\
g_1 & g_1 + b_1 & g_1 + b_2 & \cdots & g_1 + b_{2^m-1} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
g_{2^{n-m}-1} & g_{2^{n-m}-1} + b_1 & g_{2^{n-m}-1} + b_2 & \cdots & g_{2^{n-m}-1} + b_{2^m-1}.
\end{array}$$

То, что строк будет 2^{n-m} следует из теоремы Лагранжа, т.к. 2^{n-m} — это порядок фактор-группы *G/B, #(G/B)* = #(G)/#(B), $\#B = 2^m$.

Декодирование слова $g = b_j + g_i$ состоит в выборе кодового слова b_j в качестве переданного и последующем применении операции, обратной умножению на *E*. Такая схема декодирования сможет исправлять ошибки.

Для (*3, 6*)-кода из рассматриваемого примера таблица декодирования будет следующей:

000000	100110	010011	110101	001111	101001	011100	111010
100000	000110	110011	010101	101111	001001	111100	011010
010000	110110	000011	100101	011111	011001	001100	101010
001000	101110	011011	111101	000111	100001	010100	110010
000100	100010	010111	110001	001011	101101	011000	111110
000010	100100	010001	110111	001101	101011	011110	111000
000001	100111	010010	110100	001110	101000	011101	111011
000101	100011	010110	110000	001010	101100	011001	111111.

Первая строка в ней — это строка кодовых слов, а первый столбец — это лидеры.

Чтобы декодировать слово $b_i + e$, следует отыскать его в таблице и выбрать в качестве переданного слово в том же столбце и в первой строке.

Например, если принято слово 110011 (2-я строка, 3-й столбец таблицы), то считается, что было передано слово 010011; аналогично, если принято слово 100101 (3-я строка, 4-й столбец таблицы), переданным считается слово 110101, и т.д.

Групповое кодирование со схемой декодирования посредством лидеров исправляет все ошибки, строки которых совпадают с лидерами. Следовательно, вероятность правильного декодирования переданного по двоичному симметричному каналу кода равна сумме вероятностей всех лидеров, включая нулевой.

В рассмотренной схеме вероятность правильной передачи слова будет $p^6 + 6p^5q + p^4q^2$.

Кодовое слово любого столбца таблицы декодирования является ближайшим кодовым словом ко всем прочим словам данного столбца.

Пусть переданное слово b_i принято как $b_i + e$, $d(b_i, b_i + e) = w(e)$, т.е. это расстояние равно весу соответствующего лидера. Расстояние от $b_i + e$ до любого другого кодового слова b_j равно весу их поразрядной суммы, т.е.

$$d(b_j, b_i + e) = w(b_j + b_i + e) = d(b_j + b_i, e) = d(b_k, e) = w(b_k + e) \geqslant w(e),$$

т. к. e — лидер смежного класса, к которому принадлежат как $b_k + e$, так и $b_i + e$.

Доказано, при схеме декодирования лидерами по полученному слову берется ближайшее к нему кодовое.

23. Совершенные и квазисовершенные коды

Групповой (m, n)-код, исправляющий все ошибки веса, не большего k, и никаких других, называется *совершенным.* Свойства совершенного кода:

1. Для совершенного (m, n)-кода, исправляющего все ошибки веса, не большего k, выполняется соотношение $\sum_{i=0}^{k} C_n^i = 2^{n-m}$. Верно и обратное утверждение;
2. Совершенный код, исправляющий все ошибки веса, не большего k, в столбцах таблицы декодирования содержит все слова, отстоящие от кодовых на расстоянии, не большем k. Верно и обратное утверждение;
3. Таблица декодирования совершенного кода, исправляющего все ошибки в не более чем k позициях, имеет в качестве лидеров все строки, содержащие не более k единиц. Верно и обратное утверждение.

Совершенный код — это лучший код, обеспечивающий максимум минимального расстояния между кодовыми словами при минимуме длины кодовых слов. Совершенный код легко декодировать: каждому полученному слову однозначно ставится в соответствие ближайшее кодовое. Чисел m, n и k $(1 < k < \frac{n-1}{2})$, удовлетворяющих условию совершенности кода очень мало. Но и при подобранных m, n и k совершенный код можно построить только в исключительных случаях.

Если m, n и k не удовлетворяют условию совершенности, то лучший групповой код, который им соответствует называется *квазисовершенным,* если он исправляет все ошибки кратности, не большей k, и некоторые ошибки кратности $k + 1$. Квазисовершенных кодов также очень мало.

Двоичный блочный (*m, n*)-код называется *оптимальным,* если он минимизирует вероятность ошибочного декодирования. Совершенный или квазисовершенный код — оптимален. Общий способ построения оптимальных кодов пока неизвестен.

Для любого целого положительного числа *r* существует совершенный (*m, n*)-код, исправляющий одну ошибку, называемый *кодом Хэмминга* (Hamming), в котором $m = 2^r - r - 1$ и $n = 2^r - 1$.

Действительно, $\sum_{i=0}^{1} C_n^i = 1 + 2^r - 1 = 2^r = 2^{n-m}$.

Порядок построения кода Хэмминга следующий:

1. Выбираем целое положительное число *r*. Сообщения будут словами длины $m = 2^r - r - 1$, а кодовые слова — длины $n = 2^r - 1$;

2. В каждом кодовом слове $b = b_1 b_2 ... b_n$ *r* бит с индексами-степенями двойки (2^0, 2^1,..., 2^{r-1}) — являются контрольными, остальные — в естественном порядке — битами сообщения. Например, если *r* = 4, то биты b_1, b_2, b_4, b_8 — контрольные, а— $b_3, b_5, b_6, b_7, b_9, b_{10}, b_{11}, b_{12}, b_{13}, b_{14}, b_{15}$ из исходного сообщения;

3. Строится матрица *M* из $2^r - 1$ строк и *r* столбцов. В *i*-ой строке стоят цифры двоичного представления числа *i.* Матрицы для *r* = 2, 3 и 4 таковы:

$$M_{3\times 2} = \begin{bmatrix} 01 \\ 10 \\ 11 \end{bmatrix} \qquad M_{7\times 3} = \begin{bmatrix} 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{bmatrix} \qquad M_{15\times 4} = \begin{bmatrix} 0001 \\ 0010 \\ 0011 \\ 0100 \\ 0101 \\ 0110 \\ 0111 \\ 1000 \\ 1001 \\ 1010 \\ 1011 \\ 1100 \\ 1101 \\ 1110 \\ 1111 \end{bmatrix};$$

4. Записывается система уравнений $bM = 0$, где *M* — матрица из предыдущего пункта. Система состоит из *r* уравнений. Например, для *r* = 3:

$$\begin{cases} b_4 + b_5 + b_6 + b_7 = 0 \\ b_2 + b_3 + b_6 + b_7 = 0 \\ b_1 + b_3 + b_5 + b_7 = 0 \end{cases};$$

5. Чтобы закодировать сообщение *a*, берутся в качестве b_j, *j* не равно степени двойки, соответствующие биты сообщения и отыскиваются, используя полученную систему уравнений, те b_j, для которых *j* — степень двойки. В каждое уравнение входит только одно b_j, $j = 2^i$ В выписанной системе b_4 входит в 1-е уравнение, b_2 — во второе и b_1 — в третье. В рассмотренном примере сообщение *a* = 0111 будет закодировано кодовым словом *b* = 0001111.

Декодирование кода Хэмминга проходит по следующей схеме. Пусть принято слово $b + e$, где b — переданное кодовое слово, а e — строка ошибок. Так как $bM = 0$, то $(b + e)M = bM + eM = eM$. Если результат нулевой, как происходит при правильной передаче, считается, что ошибок не было. Если строка ошибок имеет единицу в *i-й* позиции, то результатом произведения eM будет *i-я* строка матрицы M или двоичное представление числа i. В этом случае следует изменить символ в *i-й* позиции слова $b + e$, считая позиции слева, с единицы.

Пример. (*4, 7*)-код Хэмминга имеет в качестве одного из кодовых слов $b = 0001111$. Матрица $M_{7\times3}$ приведена на шаге 3 хода построения кода Хэмминга. Ясно, что $bM = 0$. Добавим к b строку ошибок $e = 0010000$. Тогда $b + e = 0011111$ и $(b + e)M = 011 = 3_{10}$, т.е. ошибка находится в третьей позиции. Если $e = 0000001$, то $b + e = 0001110$ и позиция ошибки — $(b+e)M = 111 = 7_{10}$ и т.п. Если ошибка допущена в более чем в одной позиции, то декодирование даст неверный результат.

Код Хэмминга — это групповой код.

Это следует из того, что (*m, n*)-код Хэмминга можно получить матричным кодированием, при помощи $m \times n$-матрицы, в которой столбцы с номерами не степенями 2 образуют единичную подматрицу. Остальные столбцы соответствуют уравнениям шага 4 построения кода Хэмминга, т. е. 1-му столбцу соответствует уравнение для вычисления 1-го контрольного разряда, 2-му — для 2-го, 4-му — для 4-го и т.д. Такая матрица будет при кодировании копировать биты сообщения в позиции не степени 2 кода и заполнять другие позиции кода согласно схеме кодирования Хэмминга.

Пример. Кодирующая матрица для (*4, 7*)-кода Хэмминга —

$$E = \begin{bmatrix} 1110000 \\ 1001100 \\ 0101010 \\ 1101001 \end{bmatrix}.$$

Ее столбцы с номерами 3, 5, 6 и 7 образуют единичную подматрицу. Столбцы с номерами 1, 2 и 4 соответствуют уравнениям для вычисления контрольных бит, например, уравнению $b_1 = b_3 + b_5 + b_7$ соответствует столбец 1101, т.е. для вычисления первого контрольного разряда берутся 1, 2 и 4 биты исходного сообщения или биты 3, 5 и 7 кода.

К (*m, n*)-коду Хэмминга можно добавить проверку четности. Получится (*m, n + 1*)-код с наименьшим весом ненулевого кодового слова 4, способный исправлять одну и обнаруживать две ошибки.

Коды Хэмминга накладывают ограничения на длину слов сообщения: эта длина может быть только числами вида $2^r - r - 1$: 1, 4, 11, 26, 57, ... Но в реальных системах информация передается байтам или машинными словами, т.е. порциями по 8, 16, 32 или 64 бита, что делает использование совершенных кодов не всегда подходящим. Поэтому в таких случаях часто используются квазисовершенные коды.

Квазисовершенные (*m, n*)-коды, исправляющие одну ошибку, строятся следующим образом. Выбирается минимальное n так, чтобы

$$\frac{2^n}{n+1} \geqslant 2^m.$$

Каждое кодовое слово такого кода будет содержать $k = n - m$ контрольных разрядов. Из предыдущих соотношений следует, что

$$2^k = 2^{n-m} \geqslant n+1 = C_n^1 + C_n^0 = m+k+1.$$

Каждому из *n* разрядов присваивается слева-направо номер от 1 до *n*. Для заданного слова сообщения составляются *k* контрольных сумм S_1,..., S_k по модулю 2 значений специально выбранных разрядов кодового слова, которые помещаются в позиции-степени 2 в нем: для S_i $(1 \leqslant i \leqslant k)$ выбираются разряды, содержащие биты исходного сообщения, двоичные числа-номера которых имеют в *i*-м разряде единицу. Для суммы S_1 это будут, например, разряды 3, 5, 7 и т.д., для суммы S_2 — 3, 6, 7 и т.д. Таким образом, для слова сообщения $a = a_1 \,..\, .a_m$ будет построено кодовое слово $b = S_1 S_2 a_1 S_3 a_2 a_3 a_4 S_4 a_5 \ldots a_m$. Обозначим S_i^* сумму по модулю 2 разрядов полученного слова, соответствующих контрольной сумме S_i и самой этой контрольной суммы. Если $S_k^* \ldots S_1^* = 0,$ то считается, что передача прошла без ошибок. В случае одинарной ошибки $S_k^* \ldots S_1^*$ будет равно двоичному числу-номеру сбойного бита. В случае ошибки, кратности большей 1, когда $S_k^* \ldots S_1^* > n,$ ее можно обнаружить. Подобная схема декодирования не позволяет исправлять некоторые двойные ошибки, чего можно было бы достичь, используя схему декодирования с лидерами, но последняя значительно сложнее в реализации и дает незначительное улучшение качества кода.

Пример построения кодового слова квазисовершенного (*9, n*)-кода, исправляющего все однократные ошибки, для сообщения 100011010.

$$\frac{2^{12}}{13} = \frac{4096}{13} < 2^9 = 512 \quad \text{и} \quad \frac{2^{13}}{14} = \frac{4096}{7} > 512, \text{ т.е. } n = 13.$$

Искомое кодовое слово имеет вид $\overset{1}{S_1}\overset{2}{S_2}\,1\,\overset{4}{S_3}\,0\,0\,0\,\overset{8}{S_4}\,1\,1\,0\,1\,0$. Далее нужно вычислить контрольные суммы.

$$\begin{aligned}
1_{10} &= 0001_2 \\
2_{10} &= 0010_2 \\
3_{10} &= 0011_2 \\
4_{10} &= 0100_2 \\
5_{10} &= 0101_2 \\
6_{10} &= 0110_2 \\
7_{10} &= 0111_2 \\
8_{10} &= 1000_2 \\
9_{10} &= 1001_2 \\
10_{10} &= 1010_2 \\
11_{10} &= 1011_2 \\
12_{10} &= 1100_2 \\
13_{10} &= 1101_2
\end{aligned}$$

$$\begin{aligned}
S_1 &= b_3 + b_5 + b_7 + b_9 + b_{11} + b_{13} = 0 \\
S_2 &= b_3 + b_6 + b_7 + b_{10} + b_{11} = 0 \\
S_3 &= b_5 + b_6 + b_7 + b_{12} + b_{13} = 1 \\
S_4 &= b_9 + b_{10} + b_{11} + b_{12} + b_{13} = 1
\end{aligned}$$

Таким образом, искомый код — 0011000111010. Если в процессе передачи этого кода будет испорчен его пятый бит, то приемник получит код 0011100111010. Для его декодирования опять вычисляются контрольные суммы:

$$S_1^* = b_1 + b_3 + b_5 + b_7 + b_9 + b_{11} + b_{13} = 1$$
$$S_2^* = b_2 + b_3 + b_6 + b_7 + b_{10} + b_{11} = 0$$
$$S_3^* = b_4 + b_5 + b_6 + b_7 + b_{12} + b_{13} = 1$$
$$S_4^* = b_8 + b_9 + b_{10} + b_{11} + b_{12} + b_{13} = 0$$

$$S_4^* S_3^* S_2^* S_1^* = 0101_2 = 5_{10}.$$

Приемник преобразует изменением пятого бита полученное сообщение в отправленное передатчиком, из которого затем отбрасыванием контрольных разрядов восстанавливает исходное сообщение.

Совершенный код Хэмминга также можно строить по рассмотренной схеме, т.к. для него $2^n/(n + 1) = 2^m$.

Для исправление одинарной ошибки к 8-разрядному коду достаточно приписать 4 разряда ($2^{12}/13 > 2^8$), к 16-разрядному — 5, к 32-разрядному — 6, к 64-разрядному — 7.

24. **Полиномиальные коды**

При *полиномиальном кодировании* каждое сообщение отождествляется с многочленом, а само кодирование состоит в умножении на фиксированный многочлен. Полиномиальные коды — блочные и отличаются от рассмотренных ранее только алгоритмами кодирования и декодирования.

Пусть $a = a_0 \ldots a_{m-1}$ — двоичное сообщение. Тогда сопоставим ему многочлен $a(x) = a_0 + a_1 x + \ldots + a_{m-1} x^{m-1}$. Все вычисления происходят в поле классов вычетов по модулю 2, т. е. от результата любой арифметической операции берется остаток от его деления на 2.

Например, последовательности 10011 при $m = 5$ соответствует многочлен $1 + x^3 + x^4$.

Зафиксируем некоторый многочлен степени k,

$$g(x) = g_0 + g_1 x + \cdots + g_k x^k, \quad g_0 \neq 0, \quad g_k \neq 0.$$

Полиномиальный код с кодирующим многочленом $g(x)$ кодирует слово сообщения $a(x)$ многочленом $b(x) = a(x)g(x) = b_0 + b_1 x + \cdots + b_{n-1} x^{n-1}$ или кодовым словом из коэффициентов этого многочлена $b = b_0 \ldots b_{n-1}$. Условия $g_0 \neq 0$ и $g_k \neq 0$ необходимы, потому что в противном случае b_0 и b_{n-1} не будут нести никакой информации, т.к. они всегда будут нулями.

Пример. Рассмотрим кодирующий многочлен $g(x) = 1 + x^2 + x^3$.

Сообщение 01011, отвечающее многочлену $a(x) = x + x^3 + x^4$, будет закодировано коэффициентами многочлена $b(x) = g(x)a(x) = x + x^5 + x^7$, т.е. b = 01000101.

Полиномиальный код с кодирующим многочленом $g(x)$ степени k является матричным кодом с кодирующей матрицей G размерности $m \times (m + k)$:

$$G = \begin{bmatrix} g_0 & g_1 & g_2 & \cdots & g_k & 0 & 0 & \cdots & 0 \\ 0 & g_0 & g_1 & \cdots & g_{k-1} & g_k & 0 & \cdots & 0 \\ 0 & 0 & g_0 & \cdots & g_{k-2} & g_{k-1} & g_k & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & g_k \end{bmatrix}.$$

Т.е. ненулевые элементы в *j-й* строке — это последовательность коэффициентов кодирующего многочлена, расположенных с *j-го* по *(j + k)-й* столбцах.

Например, (*3, 6*)-код с кодирующим многочленом $1+x+x^3$ отвечает матрице

$$G = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

или отображению: 000 → 000000; 001 → 001101; 010 → 011010; 011 → 010111; 100 → 110100; 101 → 111001; 110 → 101110; 111 → 100011.

Полиномиальные коды являются групповыми.

Это следует из того, что коды, получаемые матричным кодированием, — групповые.

Рассмотрим *(m,n)-код* с кодирующим многочленом *g (x).* Строка ошибок $e = e_0 \ldots e_{n-1}$ останется необнаруженной в том и только в том случае, если соответствующий ей многочлен $e(x) = e_0 + e_1x + \cdots + e_{n-1}x^{n-1}$ делится на *g (x).*

Действительно, *a(x)g(x) + e(x)* делится на *g(x)* тогда и только тогда, когда *e(x)* делится на *g(x).* Поэтому любая ошибка, многочлен которой не делится на *g(x),* будет обнаружена и, соответственно, любая ошибка, многочлен которой делится на *g(x),* не может быть обнаружена.

Таким образом, обнаружение ошибки при использовании полиномиального кода с кодирующим многочленом *g(x)* может быть реализовано при помощи алгоритма деления многочленов с остатком: если остаток ненулевой, то при передаче произошло искажение данных.

Коды Хэмминга можно строить как полиномиальные, например, кодирующий многочлен $x^3 + x^2 + 1$ определяет совершенный (*4, 7*)-код, отличный от рассмотренного ранее.

Вообще же, если кодирующий многочлен *g(x),* порождающий соответствующий (*m, n*)-код, не является делителем ни одного из многочленов вида $x^j + 1$ при $j < n,$ то минимальное расстояние между кодовыми словами порожденного им кода не меньше 3.

Пусть *d* — минимальное расстояние между кодовыми словами, оно равно минимуму среди весов ненулевых кодовых слов. Предположим $d = 2$. Тогда существует *a(x)* такой, что $a(x)g(x) = b(x)$ и степень *b(x)* не больше *n.* Вес *b* равен 2, поэтому $b(x) = x^m + x^l$ и $l < m < n.$ Следовательно, $b(x)=x^l(x^{m-l} + 1),$ что означает, что $x^{m-l} + 1$ должен делиться на *g(x),* а это невозможно по условию. Если предположить, что d = 1, то это приведет к утверждению о том, что x^m должен делиться на *g(x),* что тоже противоречит условию. Итак, $d \geq 3$.

Кодирующий многочлен $x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$ определяет совершенный (*12, 23*)-код Голея (Golay) с минимальным расстоянием между кодовыми словами 7.

В 1971 году финскими и советскими математиками было доказано, что кроме кодов Хэмминга и Голея других совершенных кодов нет.

Наиболее интересными среди полиномиальных кодов являются *циклические коды,* в которых вместе с любым кодовым словом вида $b_0 \dots b_{n-2} b_{n-1}$ есть кодовое слово $b_{n-1} b_0 \dots b_{n-2}$.

25. **Понятие о кодах Боуза-Чоудхури-Хоккенгема**

Остался открытым вопрос о методике построения кодов, минимальное расстояние между кодовыми словами которых равно заданному числу. В 1960 году независимо Боуз (Bose), Чоудхури (Chaudhuri) и Хоккенгем (Hocquengem) открыли способ построения полиномиальных кодов, удовлетворяющих таким требованиям. Эти коды получили названия кодов Боуза-Чоудхури-Хоккенгема или БЧХ-кодов (BCH codes).

БЧХ-коды могут быть не только двоичными, например, на практике достаточно широко используются недвоичные коды Рида-Соломона (Reed, Solomon), но далее будут рассматриваться только двоичные.

Многочлен $g(x)$ степени k называется *примитивным,* если $x^j + 1$ делится на $g(x)$ без остатка для $j = 2^k - 1$ и не делится ни для какого меньшего значения j.

Например, многочлен $1 + x^2 + x^3$ примитивен: он делит $x^7 + 1$, но не делит $x^j + 1$ при $j < 7$. Примитивен также многочлен $1 + x^3 + x^4$ — он делит $x^{15} + 1$, но не делит $x^j + 1$ при $j < 15$.

Кодирующий многочлен $g(x)$ для БЧХ-кода, длина кодовых слов которого n, строится так. Находится примитивный многочлен минимальной степени q такой, что $n \leqslant 2^q - 1$ или $q \geqslant \log_2(n+1)$. Пусть a — корень этого многочлена, тогда рассмотрим кодирующий многочлен $g(x) = \text{НОК}(m_1(x), \dots, m_{d-1}(x))$, где $m_1(x), \dots, m_{d-1}(x)$ — многочлены минимальной степени, имеющие корнями соответственно $\alpha, \alpha^2, \dots, \alpha^{d-1}$.

Построенный кодирующий многочлен производит код с минимальным расстоянием между кодовыми словами, не меньшим d, и длиной кодовых слов n.

Пример. Нужно построить БЧХ-код с длиной кодовых слов $n = 15$ и минимальным расстоянием между кодовыми словами $d = 5$. Степень примитивного многочлена равна $q = \log_2(n+1) = 4$ и сам он равен $x^4 + x^3 + 1$. Пусть α — его корень, тогда α^2 и α^4 — также его корни. Минимальным многочленом для α^3 будет $x^4 + x^3 + x^2 + x + 1$. Следовательно,

$$g(x) = \text{НОК}(x^4 + x^3 + 1, x^4 + x^3 + x^2 + x + 1) =$$

$$= (x^4 + x^3 + 1)(x^4 + x^3 + x^2 + x + 1) = x^8 + x^4 + x^2 + x + 1.$$

Степень полученного многочлена равна 8, следовательно, построенный БЧХ-код будет (*7,15*)-кодом. Слово 1000100 или $a(x) = x^4 + 1$ будет закодировано кодовым словом $a(x)g(x) = x^{12} + x^6 + x^5 + x^2 + x + 1$ или 111001100000100.

Можно построить двоичный БЧХ-код с кодовыми словами длины $n = 2^q - 1$ и нечетным минимальным расстоянием d, у которого число контрольных символов не больше $\frac{q(d-1)}{2}$. Первый БЧХ-код, примененный напрактике, был (*92,127*)-кодом, исправляющим ошибки кратности до 5, но наиболее широкое распространение получил (*231, 255*)-код, обнаруживающий ошибки кратности до 6.

БЧХ-коды умеренной длины не слишком далеки от совершенных или квазисовершенных кодов. Коды Хэмминга, например, являются БЧХ-кодами, а БЧХ-коды с минимальным весом кодового слова 5 — квазисовершенны. Но с ростом длины кодовых слов качество БЧХ-кодов падает. Код Голея, например, — это не код БЧХ.

26. **Циклические избыточные коды**

Циклический избыточный код (Cyclical Redundancy Check — CRC) имеет фиксированную длину и используется для обнаружения ошибок. Наибольшее распространения получили коды CRC-16 и CRC-32, имеющие длину 16 и 32 бита соответственно. Код CRC строится по исходному сообщению произвольной длины, т.е. этот код не является блочным в строгом смысле этого слова. Но при каждом конкретном применении этот код — блочный, *(m, m* + 16)-код для CRC-16 или *(m, m* + 32)-код для CRC-32.

Вычисление значения кода CRC происходит посредством деления многочлена, соответствующего исходному сообщению (полином-сообщение), на фиксированный многочлен (полином-генератор). Остаток от такого деления и есть код CRC, соответствующий исходному сообщению. Для кода CRC-16 полином-генератор имеет степень 16, а для CRC-32 — 32. Полиномы-генераторы подбираются специальным образом и для кодов CRC-16/32 стандартизированы Международным консультативным комитетом по телеграфной и телефонной связи (CCITT). Для CRC-16, например, стандартным является полином-генератор $x^{16} + x^{12} + x^5 + 1$.

Пример построения CRC-4 кода для сообщения 11010111, используя полином-генератор $x^4+x^3+x^2+1$. Исходному сообщению соответствует полином $x^7 + x^6 + x^4 + x^2 + x + 1$, т. е. нумерация битов здесь начинается справа.

$$\begin{array}{l|l} x^7+x^6+x^4+x^2+x+1 & x^4+x^3+x^2+1 \\ \underline{x^7+x^6+x^5+x^3} & \hline x^3+x \\ x^5+x^4+x^3+x^2+x+1 & \\ \underline{x^5+x^4+x^3+x} & \\ x^2+1 & \end{array}$$

Полиному $x^2 + 1$ соответствуют биты 0101 — это и есть CRC-4 код.

Существуют быстрые алгоритмы для расчета CRC-кодов, использующие специальные таблицы, а не деление многочленов с остатком.

CRC-коды способны обнаруживать одиночную ошибку в любой позиции и, кроме того, многочисленные комбинации кратных ошибок, расположенных близко друг от друга. При реальной передаче или хранении информации ошибки обычно группируются на некотором участке, а не распределяются равномерно по всей длине данных. Таким образом, хотя для идеального случая двоичного симметричного канала CRC-коды не имеют никаких теоретических преимуществ по сравнению, например, с простыми контрольными суммами, для реальных систем эти коды являются очень полезными.

Коды CRC используются очень широко: модемами, телекоммуникационными программами, программами архивации и проверки целостности данных и многими другими программными и аппаратными компонентами вычислительных систем.