Избыточность сообщений

Чем больше энтропия, тем большее количество информации содержит в среднем каждый элемент сообщения

Пусть энтропии двух источников сообщений H1<H2, а количество информации, получаемое от них одинаковое, т.е.

I = n1H1 = n2H2,

где n1 и n2 - длина сообщения от первого и второго источников

Чем больше энтропия, тем большее количество информации содержит в среднем каждый элемент сообщения

Обозначим $$\mu = \frac{n_2}{n_1} = \frac{H_1}{H_2}$$

При передаче одинакового количества информации **сообщение тем длиннее, чем меньше его энтропия**.

Величина **μ**, называемая *коэффициентом сжатия*, характеризует степень укорочения сообщения при переходе к кодированию состояний элементов, характеризующихся большей энтропией.

При этом доля излишних элементов оценивается коэффициентом избыточности:

$$r = \frac{H_2 - H_1}{H_2} = 1 - \frac{H_1}{H_2} = 1 - \mu$$

- Русский алфавит, включая пропуски между словами, содержит 32 элемента, следовательно, при одинаковых вероятностях появления всех 32 элементов алфавита, неопределенность , приходящаяся на один элемент, составляет H0 = log 32 = 5 бит

- Анализ показывает, что с учетом неравномерного появления различных букв алфавита H = 4,42 бит,
- а с учетом зависимости двухбуквенных сочетаний H’ = 3,52 бит,
- т.е. **H’< H < H0**

Обычно применяют три коэффициента избыточности:

1) *частная избыточность*, обусловленная взаимосвязью **r’ = 1 - H’/H**;
2) частная избыточность, зависящая от распределения r’’ = 1 - H/ H0;
3) полная избыточность r0 = 1 - H’/ H0

Эти три величины связаны зависимостью

r0 = r’ + r’’ - r’r’’

Вследствие зависимости между сочетаниями, содержащими две и больше букв, а также смысловой зависимости между словами, избыточность русского языка (как и других европейских языков) превышает 50%

(r0 =1 - H’/ H0 = 1 - 3,52/5 = 0,30).

Избыточность играет положительную роль, т.к. благодаря ней сообщения защищены от помех.

Это используют при помехоустойчивом кодировании

Алгоритмы обратимого сжатия делятся на две группы:

- 1) Алгоритмы частотного анализа - подсчет частоты различных символов в данных и преобразование кодов символов с соответствии с их частотой.

- 2) Алгоритмы корреляционного анализа - поиск корреляций (в простейшем случае точных повторов) между различными участками данных и замена коррелирующих данных на код(ы), позволяющая восстановить данные на основе предшествующих данных.

Можно отметить следующие алгоритмы обратимого сжатия данных из первой группы:

- 1) Метод Хаффмана - замена кода равной длины для символов на коды неравной длины в соответствии с частотой появления символов в данных, коды минимальной длины присваиваются наиболее часто встречающимся символам.

- 2) Метод Шеннона-Фано - сходен с методом Хаффмана, но использует другой алгоритм генерации кодов и не всегда дает оптимальные коды (оптимальный код - код дающий наибольшее сжатие данных из всех возможных для данного типа преобразования).

- 3) Арифметическое кодирование - усовершенствование метода Хаффмана, позволяющее получать оптимальные коды для данных, где частоты появления символов не являются степенью двойки (2n). Этот метод достигает теоретической энтропийной границы эффективности сжатия этого типа для любого источника.

Для второй группы можно назвать следующие алгоритмы:

- 1) Метод Running (или RLE) - заменяет цепочки повторяющихся символов на код символа и число повторов. Это пример элементарного и очень понятного метода сжатия, но, к сожалению, он не обладает достаточной эффективностью.

- 2) Методы Лемпеля-Зива - это группа алгоритмов сжатия объединенная общей идеей: поиск повторов фрагментов текста в данных и замена повторов ссылкой (кодом) на первое (или предыдущее) вхождение этого фрагмента в данные. Отличаются друг от друга методом поиска фрагментов и методом генерации ссылок (кодов).

Пример. CD-диск (Тема доклада!!)

Нормальный на вид лазерный диск может содержать дефекты.

И даже при наличии разрушений поверхности диска

корректирующие коды C1, C2, Q - и P - уровней восстанавливают все известные приводы, и их корректирующая способность может достигать двух ошибок на каждый из уровней C 1 и C 2 и до 86 и 52 ошибок на уровни Q и P соответственно.

По мере разрастания дефектов, корректирующей способности кодов Рида—Соломона неожиданно перестает хватать, и диск без всяких видимых причин отказывает читаться.

Избыточность устраняют построением оптимальных кодов

которые укорачивают сообщения по сравнению с равномерными кодами. Это используют при архивации данных. Действие средств архивации основано на использовании алгоритмов сжатия, имеющих достаточно длинную историю развития, начавшуюся задолго до появления первого компьютера

- —/еще в 40-х гг. XX века.

Алгоритмы сжатия

Из разработок того времени практическое применение нашли алгоритмы сжатия Хаффмана и Шеннона-Фано.

А в 1977 г. математики Якоб Зив и Абрахам Лемпел придумали новый алгоритм сжатия (Зива-Лемпела), который позже доработал Терри Велч.

Основой для архивации послужили алгоритмы сжатия Я. Зива и А. Лемпела.

Суть работы архиваторов

- они находят в файлах избыточную информацию (повторяющиеся участки и пробелы),
- кодируют их,
- а при распаковке восстанавливают исходные файлы по особым отметкам.

Архиваторы

- Первым широкое признание получил архиватор Zip.

Со временем завоевали популярность и другие программы:
RAR, ARJ, ACE, TAR, LHA и т. д.

В операционной системе Windows обозначились два лидера:
WinZip (http://www.winzip.com)
и WinRAR, созданный российским программистом Евгением Рошалем
(http://www.rarlab.com).

Архивация

При архивации надо иметь в виду, что качество сжатия файлов сильно зависит от степени избыточности хранящихся в них данных, которая определяется их типом.

К примеру, степень избыточности у видеоданных обычно в несколько раз больше, чем у графических,

а степень избыточности графических данных в несколько раз больше, чем текстовых.

Программы-архиваторы можно разделить на три категории

1. Программы, используемые для сжатия исполняемых файлов.
2. Программы, используемые для сжатия мультимедийных файлов
3. Программы, используемые для сжатия любых видов файлов и каталогов.

Хотя имеются программы, которые " видят" некоторые типы архивов как самые обычные каталоги, но они имеют ряд неприятных нюансов, например, сильно нагружают центральный процессор, что

исключает их использование на "слабых машинах".

Принцип работы архиваторов

основан на

- поиске в файле "избыточной" информации и
- последующем ее кодировании с целью получения минимального объема.

Самым известным методом архивации файлов является сжатие последовательностей одинаковых символов.

Рассмотрим пример

Внутри вашего файла находятся последовательности байтов, которые часто повторяются.

Вместо того, чтобы хранить каждый байт, фиксируется количество повторяемых символов и их позиция.

Например, архивируемый файл занимает 15 байт и состоит из следующих символов:

B B B B B L L L L L A A A A A

В шестнадцатеричной системе

42 42 42 42 42 4C 4C 4C 4C 4C 41 41 41 41 41

Архиватор может представить этот файл

в следующем виде (шестнадцатеричном):

01 05 42 06 05 4C 0A 05 41

Это значит: с первой позиции пять раз повторяется символ "B", с позиции 6 пять раз повторяется символ "L" и с позиции 11 пять раз повторяется символ "A".

Для хранения файла в такой форме потребуется всего **9 байт**, что на 6 байт меньше исходного.

Этот подход не работает если обрабатываемый текст содержит небольшое количество последовательностей повторяющихся символов.

Более изощренный метод сжатия данных

используемый в том или ином виде практически любым архиватором, — это так называемый **оптимальный префиксный код** и,

в частности,

- **кодирование символами переменной длины (алгоритм Хаффмана).**

Код переменной длины

позволяет записывать наиболее часто встречающиеся символы и группы символов всего лишь несколькими битами, в то время как редкие символы и фразы будут записаны более длинными битовыми строками.

Например

В любом английском тексте буква E встречается чаще, чем Z, а X и Q относятся к наименее встречающимся.

Таким образом, используя специальную таблицу соответствия, можно закодировать каждую букву E меньшим числом битов и использовать более длинный код для более редких букв.

Архиваторы ARJ, PAK, PKZIP работают на основе алгоритма Лемпела-Зива

Это адаптивные словарные кодировщики.

Важнейшей отличительной чертой этого алгоритма является использование грамматического разбора предшествующего текста с расположением его на фразы, которые записываются в словарь.

Указатели позволяют сделать ссылки на любую фразу в окне установленного размера, предшествующего текущей фразе. Если соответствие найдено, текущая фраза заменяется указателем на своего предыдущего двойника.

КОДИРОВАНИЕ ИЗОБРАЖЕНИЯ, ЗВУКА И ВИДИО

ГРАФИЧЕСКИЕ ФОРМАТЫ ДЛЯ СОХРАНЕНИЯ ИЗОБРАЖЕНИЯ

Все графические форматы делятся на две большие группы:

растровые и векторные.

Файлы растровых форматов

содержат описание каждой точки изображения.

Они представляют собой прямоугольную матрицу (bitmap), состоящую из пикселей. (GIF и JPEG, BMP (стандартный формат Windows) и TIFF, применяющийся в полиграфии.)

В растровом формате хранятся фотографии, рисунки и обои Рабочего стола.

Видео также является последовательностью растровых изображений.

Файлы векторных форматов

Содержат математические формулы, описывающие координаты кривых.

Например, прямая линия представлена координатами двух точек, а окружность — координатами центра и радиуса, поэтому достигается очень небольшой размер файла.

.

В векторных форматах сохраняются логотипы, схематические рисунки, текст, предназначенный для вывода на печать, и пр.

Формат BMP

BMP (Windows Device Independent Bitmap) — это один из старейших форматов, к тому же
являющийся «родным» форматом Windows

Записанный в нем файл представляет собой массив данных, содержащий информацию о цвете каждого пиксела

Основной недостаток формата, ограничивающий его применение, — большой размер BMP-файлов.

Конечно, можно попробовать сохранить
изображение в формате BMP со сжатием, однако это часто вызывает проблемы при работе с некоторыми графическими пакетами.

GIF (CompuServe Graphics Interchange Format)

- В 1989 г. CompuServe выпустила усовершенствованную версию формата, названную GIF89a на основе алгоритма LWZ.

В нее были добавлены две функции:

- альфа-канал, где может храниться маска прозрачности,
- GIF стал анимированным, т. е. в один файл можно поместить несколько изображений

- чересстрочная развертка. Во время загрузки изображения создается эффект постепенного проявления картинки на экране.

Форматы на основе LZW не справлялись с эффективной обработкой фотографий, и

- потому появилась идея сжатия с потерей качества.

Суть его заключается в том, что

- часть малозаметных для глаза деталей изображения опускается,
- а восстановленный после сжатия
- цифровой массив не полностью соответствует оригиналу.

Таким образом, можно добиться
довольно большой степени сжатия данных —
в 10—20 раз вместо двукратного, производимого
без потерь.

JPEG

- В 1991 г. группа Joint Photographic Experts Group, опирающаяся на более чем полувековой опыт исследований в области человеческого зрения, представила первую спецификацию формата JPEG.

- Через три года она была признана индустриальным стандартом кодирования неподвижных изображений, зарегистрированным как ISO/IEC 10918-1.

- Впоследствии JPEG лег в основу стандарта сжатия видео **MPEG**.

JPEG (Joint Photographic Experts Group) (тема доклада!!!)

- признания он достиг благодаря одноименному алгоритму сжатия, который показал отличные результаты в соотношении размер/качество.

JPEG сжимает файлы весьма оригинальным образом:

- он ищет *не одинаковые пиксели, а вычисляет разницу между соседними квадратами размером 9x9 пикселей.*

- Информация, не заслуживающая внимания, отбрасывается, а ряд полученных значений усредняется.
- В результате получают файл в десятки или сотни раз меньшего размера, чем BMP.

- Естественно, чем выше уровень компрессии, тем ниже качество.

Формат TIFF (Tagged Image File Format)

- используется в издательствах, поскольку при сохранении изображения в этом формате не происходит потерь качества.
- Формат поддерживают практически все программы как PC, так и Macintosh.

- TIFF — это наиболее универсальный формат, обеспечивающий среднюю степень сжатия файла.

- Храня работы в TIFF, можно быть уверенным, что при просмотре файла вы увидите именно то изображение, которое сохраняли.
- За качество придется расплачиваться достаточно большим объемом файлов.

ХАРАКТЕРИСТИКИ ВИДЕОФАЙЛОВ И ВИДЕОПЛЕЕРОВ

Стандарт MPEG-4 (тема доклада!!!)

Файлы с расширением AVI, которые используют этот кодек, не имеют постоянных параметров, так как существуют в двух ипостасях:

- без компрессии и
- файлы, сжатые, например, кодеком DivX, которые также имеют расширение AVI.

Типы воспроизводимых файлов

Тема доклада!!!!

Обзор популярны форматов файлов и видеоплееров, которые могут их воспроизвести

КОДИРОВАНИЕ ЗВУКА

Кодеки (Тема доклада!!)

Существует два основных стандарта:
MP3 и WMA.

Если стандарт WMA разрабатывается исключительно фирмой Microsoft,
то кодек для сжатия цифрового звука в стандарте MP3 может создать любой программист.

В результате это привело к появлению большого количества алгоритмов кодирования.

В условиях конкурентной борьбы за качество MP3-файлов на первое место вышел проект нескольких программистов — Lame.

ФОРМАТ PDF

Portable Document Format или просто PDF, был создан специально для ликвидации проблем с отображением информации в файлах.

- PDF использует качественные алгоритмы сжатия (LWZ): если объем файла Word, содержащего пару картинок, вряд ли получится меньше мегабайта, то точно такой же PDF вполне уместится в 300-400 Кбайт.
- PDF умеет встраивать в себя все используемые в документе шрифты.
-
- в формат PDF можно преобразовать любой электронный документ.

Тема доклада!!

- «Обзор алгоритмов сжатия в популярных форматах»

Все это плавный переход

- К теме

Эффективное кодирование