Кафедра КиИИ

ЛЕКЦИЯ

Тема: Теория сжатия информации и оптимальные методы кодирования

Дисциплина: Основы кодирования

ОП Системы информационной безопасности

Авторы: Ассоц. проф. КиИИ Даненова Г.Т.
Ассоц. проф. КиИИ Коккоз М.М.
Ассоц. проф. КиИИ Кан О.А.
Ассоц. проф. СМиТ Ахметжанов Т.Б.

План лекции

1. Сжатие информации.
2. Основы оптимального кодирования.
3. Задача статистического кодирования.
4. Алгоритм RLE.
5. Кодирование методом LZW.

Сжатие информации

Одним из методов криптографического преобразования, наряду с шифрованием, кодированием и стеганографией, является сжатие данных.

Сжатие данных (архивирование) — это алгоритмическое преобразование данных, в результате которого уменьшается ее избыточность и, соответственно, требуется меньший объем памяти для хранения.

Сжатие без потерь может применяться для кодирования любой информации, поскольку обеспечивает абсолютно точное восстановление данных после декодирования.
Сжатие без потерь основано на простом принципе преобразования данных из одной группы символов в другую, более компактную.
Наиболее известны два алгоритма сжатия без потерь: это кодирование методом Хаффмана и LZW - кодирование (по начальным буквам имен создателей Lempel , Ziv , Welch).

Кодирование методом Хаффмана является простым алгоритмом для построения кодов переменной длины. Этот популярный алгоритм служит основой многих компьютерных программ сжатия текстовой и графической информации.
Принцип кодирования Хаффмана заключается в уменьшении количества бит, используемых для кодирования часто встречающихся символов и, соответственно, в увеличении количества битов, используемых для редко встречающихся символов.

Метод LZW анализирует входные строки символов начального алфавита для построения расширенного алфавита и дальнейшего кодирования текста.
Обычно в качестве начального алфавита используется всем известная стандартная таблица символов ASCII.
Алгоритм основан на идее расширения алфавита, что позволяет использовать дополнительные символы для представления группы исходных символов.

Таким образом, сжатие с потерями применяется в основном для графики, звука и видео, т.е. там, где в силу огромных размеров файлов степень сжатия очень важна, и можно пожертвовать деталями, несущественными для восприятия этой информации человеком.
Например, алгоритмы сжатия графических изображений с потерями, обеспечивают очень высокие степени компрессии. При этом изменения в изображениях практически незаметны для человека.

В настоящее время существует много различных архиваторов. Они имеют разную распространенность и эффективность. Разработано большое количество разнообразных методов, их модификаций и подвидов для сжатия данных.
Самым популярным архиватором является WinZip. Объясняется это тем, что формат ZIP считается мировым стандартом архивирования и имеет самую длительную историю развития.
За ним следуют многими любимый формат WinRAR и набирающий обороты WinAce.

Например, программа WinZip 9.0. работает в двух режимах: классическом и режиме мастера, рассчитанном на новичков. Она ориентирована преимущественно на ZIP-архивы, но при этом поддерживает и другие популярные архивные форматы.
В числе возможностей WinZip — поддержка технологии перетаскивания (drag & drop), создание самораспаковывающихся файлов, поддержка антивирусных программ, отправка архива по электронной почте и др.

Алгоритм шифрования AES

Алгоритм шифрования AES (Advanced Encryption Standard) — симметричный алгоритм блочного шифрования (размер блока 128 бит, длина ключа 128/192/256 бит), принят в качестве стандарта шифрования правительством США в 2000-ом году.
Этот алгоритм хорошо проанализирован и сейчас широко используется. По состоянию на 2009 год AES является одним из самых распространённых алгоритмов симметричного шифрования.

Основы оптимального кодирования

В последние годы были разработаны различные методы повышения эффективности передачи информации. Эти методы получили название методов оптимального кодирования, которые можно разделить на:

- адаптивные методы;
- разностные методы;
- статистические методы.

Адаптивное кодирование

Системы связи, в которых при изменении состояния канала меняется и метод кодирования, называются системами с адаптивным кодированием.

При хорошем состоянии канала используется простой код с малой избыточностью, с ухудшением состояния канала переходят к коду с большей избыточностью, замедляя скорость передачи информации, но повышая надежность и помехозащищенность.

Разностное кодирование

Разностный метод кодирования заключается в следующем:

- кодируются разности между предыдущими и текущими значениями передаваемых сообщений;
- прием «разностных» кодов осуществляется последовательным суммированием и декодированием результата суммирования.

Рассмотрим пример.
Пусть имеется случайная величина $X(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8)$, имеющая восемь состояний с распределением вероятностей P(1/4, 1/4, 1/8, 1/8, 1/16, 1/16, 1/16, 1/16).

Для кодирования алфавита из восьми букв равномерным двоичным кодом без учета вероятностей нам понадобятся три бита:

$$K = \log_2 8 = 3$$

Это 000, 001, 010, 011, 100, 101, 110, 111

Чтобы ответить, оптимальный этот код или нет, необходимо определить энтропию

$$H = -\sum_{i=1}^{8} p_i \log p_i = -2 \cdot \frac{1}{4}\log\frac{1}{4} - 2 \cdot \frac{1}{8}\log\frac{1}{8} - 4 \cdot \frac{1}{16}\log\frac{1}{16} = 2.75$$

Определив избыточность **L** по формуле

$$\mathbf{L = 1 - H/H_0 = 1 - 2,75/3 = 0,084}$$

видим, что возможно сокращение длины кода на 8,4%.

Принципы построения оптимальных кодов

1. Каждый элементарный символ должен переносить максимальное количество информации, для этого необходимо, чтобы элементарные символы (0 и 1) в закодированном тексте встречались в среднем одинаково часто. Энтропия в этом случае будет максимальной.
2. Символам первичного алфавита, имеющим большую вероятность появления в сообщении, присваиваются более короткие кодовые комбинации.

При таком кодировании избыточность кода, которая вызвана различными вероятностями символов алфавита, сводится к минимуму. Оптимальные коды являются неравномерными блочными кодами, поэтому при их построении необходимо обеспечить однозначность декодирования.

Префиксным - называется код, в котором ни одна кодовая комбинация не является началом другой комбинации.

Пример.

01 11 0010 10101 110

Алгоритм RLE

Одним из простейших методов сжатия изображений является алгоритм RLE (Run Length Encoding – кодирование с переменной длиной строки). Основной идеей этого метода является поиск одинаковых пикселей в одной строке. Найденные цепочки одинаковых элементов заменяются на число повторений и значение элемента, что существенно уменьшает избыточность данных.

Данный метод, как правило, достаточно эффективен для сжатия растровых графических изображений (BMP, PCX, TIFF), т.к. последние содержат достаточно длинные серии повторяющихся последовательностей байтов. Алгоритм в первую очередь рассчитан на изображения с большими областями повторяющегося цвета (деловая графика, рисунки и т.п.).
Наиболее активно используется алгоритм RLE при сжатии графических данных, имеющих последовательности повторяющихся байтов в виде сплошной заливки.

Принцип работы алгоритма LZW

В начале составляется таблица индексов всех цветов, имеющихся в исходном изображении. Таким образом, вместо значения цвета пиксела можно использовать индекс из таблицы. Наиболее часто встречающиеся цвета на изображении имеют меньшие коды, а редко встречающиеся цвета размещаются в конце таблицы. Таблица цветов (палитра) размешается между заголовком и собственно изображением.

Контрольные вопросы

1. Что такое сжатие данных?
2. Для чего применяется сжатие данных?
3. В каких случаях применяется сжатие без потерь?
4. Основные возможности программы WinZip.
5. Приведите пример алгоритма симметричного шифрования.
6. В чем заключается принцип работы алгоритма RLE?
7. Принцип построения оптимальных кодов.

Следующая лекция

Криптографические методы преобразования и защиты информации