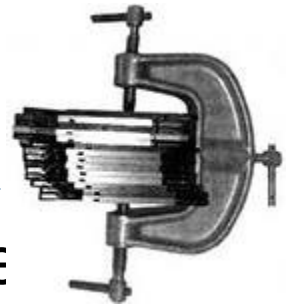
Технологии обработки информации

Лекция 5
Задачи анализа. Сжатие

Содержание

- Общие понятия
- Избыточность данных. Теорема Шеннона
- Классификации методов сжатия
- Перечень алгоритмов сжатия
- Описание отдельных методов и алгоритмов
  - RLE
  - LZW
  - Хаффмана
  - PPM

Сжатие данных



- (англ. *data compression*) — алгоритмическое преобразование данных, производимое с целью уменьшения занимаемого ими объёма. Применяется для более рационального использования устройств хранения и передачи данных.
- Синонимы — *упаковка данных, компрессия, сжимающее кодирование, кодирование источника*.
- Обратная процедура называется восстановлением данных (распаковкой, декомпрессией).

Википедия

Избыточность данных

Видео

Графика

Текст

Избыточности и кодирование

- Разные способы кодирования дают разную избыточность
- Пример: кодирование текста средствами русского языка избыточно на 20-25% по отношению к английскому

Всегда ли избыточность – это плохо?

Идея сжатия

- представлять часто используемые элементы длинными кодами
- редко используемые – короткими кодами
- Тогда: для хранения блока данных требуется меньший объем, чем для кодирования с одинаковыми длинами кодов
- Пример:

  азбука Морзе

Теорема Шеннона о кодировании источника

- элемент s_i, вероятность появления которого равняется $p(s_i)$, выгоднее всего представлять $-\log_2 p(s_i)$ битами
- Если распределение вероятностей неизменно и средняя длина кодов:

$$H = -\sum_i p(s_i) \cdot \log_2 p(s_i)$$

- энтропия Шеннона

- в подавляющем большинстве случаев истинная структура источника нам не известна, поэтому необходимо строить модель источника, которая позволила бы оценить вероятность $p(s_i)$

- Вывод: чтобы эффективно сжимать нужно знать

Классификации методов сжатия (1)

- Необратимое (с регулируемыми потерями) — методология, при которой для обеспечения максимальной степени сжатия исходного массива часть содержащихся в нем данных отбрасывается
- Обратимое (без потерь) — методология сжатия, при которой ранее закодированная порция данных восстанавливается после их распаковки полностью без внесения изменений

Классификации методов сжатия (2)

- Симметричное (*symmetric compression*) —время, затрачиваемое на сжатие и распаковку данных, соизмеримо
- Асимметричное (*asymmetric compression*) — методология, в соответствии с которой при выполнении работ «в одном направлении» времени затрачивается больше, чем при выполнении работ в другом направлении (сжатие изображений vs. резервное копирование)

Классификации методов сжатия (3)

- Адаптивное кодирование (adaptive encoding) — методология кодирования при сжатии данных, которая заранее не настраивается на определенный вид данных (двухпроходные алгоритмы)
- Неадаптивное кодирование (nonadaptive encoding) — методология кодирования, ориентированная на сжатие определенного типа или типов данных (словарные алгоритмы)
- Полуадаптивное кодирование (half-adaptive coding) — методология кодирования при сжатии данных, которая использует элементы адаптивного и неадаптивного кодирования

Классификации методов сжатия (4)

- Адаптивное кодирование (adaptive encoding) — методология кодирования при сжатии данных, которая заранее не настраивается на определенный вид данных (двухпроходные алгоритмы)
- Неадаптивное кодирование (nonadaptive encoding) — методология кодирования, ориентированная на сжатие определенного типа или типов данных (словарные алгоритмы)
- Полуадаптивное кодирование (half-adaptive coding) — методология кодирования при сжатии данных, которая использует элементы адаптивного и неадаптивного кодирования

Перечень алгоритмов сжатия. Без потерь (1)

- Преобразование Барроуза-Уилера (BWT)
- Преобразование Шиндлера (ST)
- Алгоритм DEFLATE
- Дельта-кодирование
- Инкрементное кодирование
- Семейство алгоритмов LZW
- Алгоритм сжатия PPM
- Кодирование длин серий (RLE)
- Алгоритм SEQUITUR
- EZW-кодирование

Википедия

Перечень алгоритмов сжатия. Без потерь (2)

- Энтропийное кодирование:
  - Алгоритм Шеннона-Фано
  - Алгоритм Хаффмана
    - Адаптивное кодирование Хаффмана
  - Усечённое двоичное кодирование
  - Арифметическое кодирование
    - Адаптивное арифметическое кодирование
  - Кодирование расстояний
- Энтропийное кодирование с известными характеристиками:
  - Унарное кодирование
  - дельта|гамма|омега-кодирование Элиаса
  - Кодирование Фибоначчи
  - Кодирование Голомба

Википедия

Перечень алгоритмов сжатия. С потерями

- Дискретно-косинусное преобразование
- Линейное предсказывающее кодирование
- А-закон
- Мю-закон
- Фрактальное сжатие
- Трансформирующее кодирование
- Векторное квантование
- Вейвлетное сжатие

Википедия

Группы методов сжатия без потерь

- алгоритм RLE (Run Length Encoding)
- алгоритмы группы KWE (KeyWord Encoding)
- вероятностные алгоритмы

Кодирование длин серий (Run-length encoding, RLE)

- простой алгоритм сжатия данных, который оперирует сериями данных, то есть последовательностями, в которых один и тот же символ встречается несколько раз подряд
- При кодировании строка одинаковых символов, составляющих серию, заменяется строкой, которая содержит сам повторяющийся символ и количество его повторов
  - WWWWWWWWWWWWBWWWWWWWWWWWWBBBWWWWWW WWWWWWWWWWWWWWWWWWWWBWWWWWWWWWWWWWW – 67 символов
  - 12W1B12W3B24W1B14W – 18 символов

Ограничения алгоритма RLE

1. Если строка состоит из большого количества неповторяющихся символов, её объем может вырасти

- Решение: использовать отрицательные числа для записи количества неодинаковых символов

ABCABCABCABCDDEFFFFFFFFF

-12ABCABCABCABC2D1E8F

2. Пределы длин численных данных

- Решение: разделение длинных последовательностей на группы

Алгоритмы группы KWE

- Принцип кодирования лексических единиц (повторяющихся последовательностей символов) группами байт фиксированной длины
- Словарные алгоритмы — разбиение данных на слова и замена их на индексы в словаре
- Наиболее известные – алгоритмы семейства LZ* (LZ77/78, LZW, LZO, DEFLATE, LZMA, LZX, ROLZ)

Алгоритм LZW (1)

- Авторы: Абрахам Лемпель (англ. Abraham Lempel), Якоб Зив (англ. Jacob Ziv) и Терри Велч (англ. Terry Welch)
- был опубликован Велчем в 1984 году, в качестве улучшенной реализации алгоритма LZ78, опубликованного Лемпелем и Зивом в 1978 году
- патент принадлежал Зиву
- Основные идеи:
  - Принцип скользящего окна
  - Механизм кодирования совпадений

Алгоритм LZW (2)

1. Инициализация словаря всеми возможными односимвольными фразами. Инициализация входной фразы W первым символом сообщения
2. Найти в словаре строку W наибольшей длины, которая совпадает с последними принятыми символами
3. Считать очередной символ K из кодируемого сообщения
4. Если КОНЕЦ_СООБЩЕНИЯ, то выдать код для W, иначе
5. Если фраза WK уже есть в словаре, присвоить входной фразе W значение WK и перейти к Шагу 3, иначе выдать код W, добавить WK в словарь, присвоить входной фразе W значение K и перейти к

Вероятностные методы

- Алгоритм Хаффмана
- Алгоритм PPM
- Алгоритм BWT

Алгоритм Хаффмана (1)

- адаптивный жадный алгоритм оптимального префиксного кодирования алфавита с минимальной избыточностью
- Был разработан в 1952 году аспирантом Массачусетского технологического института Дэвидом Хаффманом при написании им курсовой работы
- Идея: зная вероятности символов в сообщении, можно описать процедуру построения кодов переменной длины, состоящих из целого количества битов
- Символам с большей вероятностью ставятся в соответствие более короткие коды
- Коды Хаффмана обладают свойством префиксности (т.е. ни одно кодовое слово не является префиксом

Алгоритм Хаффмана (2)

- Состоит из двух основных этапов:
  1. Построение оптимального кодового дерева
  2. Построение отображения код-символ на основе построенного дерева
- Недостаток:
  - для восстановления содержимого сжатого сообщения декодер должен знать таблицу частот, которой пользовался кодер
- Выход: адаптивное сжатие Хаффмана

Алгоритм PPM

- PPM (prediction by partial matching) - это метод контекстно-ограниченного моделирования, позволяющий оценить вероятность символа в зависимости от предыдущих символов

Алгоритм BWT

- Преобразование Барроуза -Уилера (Burrows-Wheeler transform, BWT, также называется блочно-сортирующим сжатием)
- сравнительно новая и революционная техника для сжатия информации (в особенности-текстов), основанная на преобразовании, открытом в 1983 г. и описанная в 1994 г.
- Меняет порядок символов во входной строке таким образом, что повторяющиеся подстроки образуют на выходе идущие подряд последовательности одинаковых символов
- Используется последовательность BWT → MTF/RLE → Хаффман