

Сжатие данных

- Большинство типов данных хар-ся избыточностью
- степень избыточности видеоданных > степени избыточности графических данных > степень избыточности текстовых данных
- фактором, влияющим на степень избыточности является принятая система кодирования

Пример. кодирование текста русского языка дает в среднем избыточность на 20-25% большую, чем кодирование аналогичных данных в англ. языке

Избыточность - центральное понятие в теории сжатия информации. Любые данные с избыточной информацией можно сжать. Данные, в которых нет избыточности, сжать нельзя.

До появления работы **Шеннона** (энтропия) , кодирование символов алфавита при передаче сообщения по каналам связи осуществлялось одинаковым количеством бит, получаемым по формуле **Хартли** (1928 г.).

- Теорема **Шеннона об источнике шифрования** показывает, что невозможно сжать данные настолько, что оценка кода (среднее число бит на символ) станет меньше, чем энтропия Шеннона исходных данных, без потери точности информации.

Классификация методов сжатия

1. По степени соответствия вх (до) и вых (после) данных

- С потерями информации

- * JPEG - для графических данных;
- * MPG - для видеоданных;
- * MP3 - для аудиоданных.

- Без потерь информации

- * GIF, TIFF - для графических данных;
- * AVI - для видеоданных;
- * ZIP, ARJ, RAR, CAB, LH - для произвольных типов данных

2. По способу кодирования (преобразования)

- На основе выявления повторяющихся последовательностей данных и замены их более простой структурой (**интервалов, RLE, Б-У**) - **символориентированные**
- Сжатие по ключевым словам: кодирование лексических единиц группами байт фиксированной длины (**ЛЗ**)-**словарные**
- На основе использования частотных (вероятностных) свойств символов входного алфавита (**Хаффмана, Шеннона-Фано**)-**Вероятностные (статистические)**
- **Арифметические**