

Кодирование и сжатие данных

Алексей Владыкин

СПбГУ ИТМО

3 мая 2010

Определения

- Кодирование — преобразование $f : A^* \rightarrow B^*$, где $A = \{a_1, \dots, a_n\}$ и $B = \{b_1, \dots, b_m\}$ — алфавиты, A^* и B^* — множества слов в этих алфавитах.
- Декодирование — обратное преобразование $f^{-1} : B^* \rightarrow A^*$.
- Коэффициент сжатия сообщения α

$$c(\alpha) = \frac{|f(\alpha)|}{|\alpha|}.$$

Мотивация

- Представление данных — чисел, текста, графики, звука — в памяти компьютера.
- Защита данных от искажений при передаче по каналам связи.
- Сжатие данных.
- Шифрование данных.

Примеры способов сжатия

- Сжатие 7-битных ASCII-текстов
- Сжатие Unicode-текстов в UTF-8
- Сжатие текстов с использованием сокращений
- Сжатие числовых последовательностей с низкой вариацией
- RLE-сжатие
- Сжатие до порождающей программы

Алфавитное кодирование

- Каждый символ сообщения кодируется независимо:

$$f(\alpha) = f(a_{i_1} \dots a_{i_k}) = \sigma(a_{i_1}) \dots \sigma(a_{i_k}) = \beta_{i_1} \dots \beta_{i_k}$$

- $\sigma : A \rightarrow B^*$ — схема, или таблица кодов.
- Коды постоянной или переменной длины.
- Схема разделима, если любую кодовую последовательность $b_{i_1} \dots b_{i_j}$ можно однозначно разделить на кодовые последовательности отдельных символов $\beta_{i_1} \dots \beta_{i_k}$

Префиксный код

- Префиксная схема, префиксный код, условие Фано: ни одна кодовая последовательность β_i не является префиксом никакой другой кодовой последовательности β_j .
- Кодовые последовательности β_i префиксной схемы можно развесить на кодовом дереве. При этом кодовые последовательности будут заканчиваться только в листах дерева.

Оптимальный префиксный код

- Дано распределение вероятностей символов кодируемого сообщения: $p_1 \geq p_2 \geq \dots p_n$
- Задача: минимизировать $\sum_{i=1}^n p_i |\beta_i|$
- Если $p_i > p_j$, то $|\beta_i| \leq |\beta_j|$.
- Среди кодов β_i максимальной длины имеются два, которые различаются только в последнем разряде.
- Если $p_j = q' + q''$ и $p_n \geq q' \geq q''$, то схема

$$\beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_n, \beta_j 0, \beta_j 1$$

оптимальна для распределения вероятностей

$$p_1, \dots, p_{j-1}, p_{j+1}, \dots, p_n, q', q''.$$

Код Хаффмена

- 1 Сгенерировать набор одиночных узлов: по узлу на каждый символ кодируемого сообщения.
- 2 Отсортировать узлы по весу (частоте соответствующего символа).
- 3 Взять два узла с минимальными весами, объединить их общим родительским узлом, сложив веса.
- 4 Повторять п. 3, пока узлов больше одного.
- 5 В полученном дереве пометить пути нулями и единицами.