Жим-жим данных - 3чт

Григоренко Виктор Михайлович - тел. (967) 1-7777-25

Изучаем сжатие данных и применение на конкретных примерах
Плюс будет обработка текста кажется

Возможны репрессии лмао

Модули

- Фундаментальные основы кодирования и теория информации (л 1-2)
- Статистические методы (л. 3)
- Семантические методы (л. 4)
- Алгоритмическая теория информации (л. 5)
- Онтология предметной области (л. 6)
- Прикладные задачи (л. 7)

Хихи прикольные книжки

- Павлов "Теория информации для бакалавров"

Основы Базы Фундаменты

Исходное сообщение - исходные данные
Алфавит источника - алфавит кодировки исходного сообщения
Алфавит кодирования - алфавит итогового сообщения

Кодовые слова	**Исходное сообщение**	**Кодовое слово**	**Исходное сообщение**	**Кодовое слово**
block-block	a	000	aa	0
block-var	b	001	bbb	1
var-block	c	010	cccc	10
var-var	d	011	ddddd	11
	e	100	eeeeee	100
	f	101	fffffff	101
	g	110	ggggggggg	110
	space	111	space	111
	block-block		var-var	
Ансамбль - последовательность иходных сообщений				
Различимые коды - неоднозачно декодируемые (11 = bbbbbb, dddddd)				
Префиксные коды - однозначно декодируемые				

Классификация методов сжатия

Статические - преобразование до передачи. Пример - код Хаффмана
Динамические - преобразование во время передачи (адаптивные коды, однопроходные коды)

Как измерить сжатие? Как сравнивать?

Сложность алгоритма и степень сжатия

- При передаче - скорость сжатия, скорость передачи;
- При хранении - степень сжатия;
  Статическая схема: 3 алгоритма для анализа
  Динамическая: 2 алгоритма

Метрики

- Избыточность (redundancy) - Shannon and Weaver 1949 et al
- Эффективность кода = средняя длина сообщение (average message length) - Huffman 1952 et al
- Степень сжатия (compression ratio) - Rubin 1976, Ruth and Kreutzer 1972 et al
- Вероятность (fixed probability assignment) p(a)
- Количество информации в исходном сообщении (энтропия)
  Важно использовать оптимальный код - с минимальной избыточностью

Где используется сжатие данных

Используется при:

- Хранении
- Передаче
- Поиске
- Обработке
  При сжатии испольуется хэширование - получение числа из строки по формуле
  В различных архитектурах широко используют код Хаффмана (степень сжатия ~50% и выше)