Семантические методы сжатия

Как наверняка очевидно из названия - методы, учитывающие смысл сообщений.
Сжатие смысла даёт одни из лучших показателей сжатия (как пример - всю Войну и Мир можно ужать в одно предложение, однако важно понимать, какой смысл нужно оставлять а какой ужимать)
Области применения:

- Обработка/представление изображений/видео
- Обработка/передача бизнес-информации
- Обработка текста
  // 1 печатный лист = 28 страниц. Прикольно
- Поисковые машины
- Big Data
  Алгоритмы:
- RLE (Run Length Encoding), CCITT
- Кодирование карты различий (Diff. Mapping)
- Словарная замена - EDIFx, IBM IMS
- Свободный разбор (free-parse) ZL78, ZLW, др.
- Онтологии - семантический разбор понятийной области
- Другие - по хорошему можно придумать другие новые методы

RLE

Пары (c, l), где c - яркость/цвет, l - длина последовательности пикселей одной яркости/цвета

Diff. Mapping

... - то же самое, что и RLE, но вместо c - Δc - разница яркости/цвета

Сжатие деловой информации

RLE - c - дорожка нулей/пробелов/одинаковых записей, l - длина

Словарная замена

Протоколы EDIFx, IBM IMS (42.1% сжатия)
Пример - 01890_ABCD_LMN
0, B, C, D, M, N - Letter code; 1890 - Number code
A, L - blank code

Алгоритмы сжатия текстовых файлов

- Словарные методы Hahn, 1974
- Словари окончаний слов, Tropper, 1982
- Программный выбор суффиксов/приставок
- Программный выбор пар словосочетаний
- Морфологические словари
- Лексические словари
  Характеристик слова - около 50. Neat

Связи между статистическими и семантическими алгоритмами

Чем однороднее строка - тем лушче сжатие (статистика? Статистика)
Алгоритмы по типу diff. mapping, Hahn, Tropper работают в специализированных обстановках в то время, когда коды Хаффмана, Зива-Лемпеля работают всегда

Алгоритмы с применением кодовых книг (codebooks)

Суть - составить словарь предметной области, где коду в соответствие ставится некий смысл

Следующая лекция - кусок из теории информации, через лекцию - лаба