Сравнение алгоритмов сжатия коротких текстовых сообщений в задаче информационного поиска

Дудин Виктор Дмитриевич, 545 группа

Научный руководитель:
ст. преп. Луцив Д. В.

Рецензент:
к.ф.-м.н. Кураленок И. Е.

СПбГУ, 2014 г.

Предметная область

Современные архиваторы хорошо сжимают длинный текст, но не умеют компактно сжимать короткие сообщения

Актуальность проблемы для поисковых систем:

- экономия дискового пространства
- сжатие интернет-трафика
- шифрование интернет-трафика
- оптимизация поисковых структур данных

Постановка задачи

- изучить существующие алгоритмы сжатия текста
- реализовать свой алгоритм, эффективно сжимающий короткие текстовые сообщения
- сравнить существующие алгоритмы с собственным алгоритмом, определить его применимость в инфраструктуре поисковой системы

Существующие решения

Лучшие алгоритмы для сжатия текста:

1. Prediction by Partial Matching (**PPM**)
2. Context mixing (**CM**)

9 из 10 топовых текстовых архиваторов основаны либо на **PPM**, либо на **CM**

Лучшие текстовые архиваторы:

1. Серия DURILCA (**PPM**)
2. Серия PAQ (**CM**)
3. CMIX (**CM**)

Собственный алгоритм (1)

Алгоритм Хаффмана с расширенным словарем

Предварительно создается **Словарь** - набор популярных комбинаций символов

Каждому элементу словаря определен код Хаффмана

Вхождения элементов словаря в текст заменяются
на соответствующие коды Хаффмана

Собственный алгоритм (2)

Главное достоинство алгоритма -

статический словарь

В сжимаемых файлах не нужно сохранять информацию для разархивирования

Поэтому данный алгоритм эффективно сжимает как длинные, так и короткие текстовые сообщения

Генерация словаря

Итеративное наполнение словаря

На каждой итерации добавляем самую хорошую комбинацию символов

Метод последовательных улучшений

На каждой итерации пытаемся улучшить тот словарь, который уже есть

Словарь "обучается" на подмножестве из тех данных, которые потом будет сжимать

Сравнение алгоритмов (1)

Критерии сравнения:

1. Скорость разархивирования
2. Коэффициент сжатия

Тестовые данные:

- small - 40 байт (1 URL)
- medium - 2 Кбайт (60 URL)
- large - 200 Кбайт (6000 URL)
- full - 27 Мбайт (750 000 URL)

Сравнение алгоритмов (2)

Program	small (40 B)		medium (2 KB)		large (200 KB)		full (27 MB)	
	%	time	%	time	%	time	%	time*
paq8hp12	159,08	22433	41,75	22657	25,15	33420	16,83	1618
cmix	116,80	22287	41,09	23600	24,97	112930	---	---
nanozip	297,20	13	53,04	17	32,06	50	21,08	2,28
gzip	180,39	1	51,53	1	37,51	3	37,34	0,13
bzip2	197,34	1	52,59	1	33,36	14	30,47	1,1
7z	420,52	4	57,30	4	33,19	11	23,93	0,42
Huffman	75,2	1	64,29	1	65,18	27	65,18	2,22
Huffman with vocabulary	**47,15**	1	**27,91**	1	**28,91**	16	29,11	1,21

% - процент от размера исходного файла
time - время для разархивации, мс (*сек)

Результаты

- изучены существующие алгоритмы сжатия текста
- реализован свой алгоритм, эффективно сжимающий короткие текстовые сообщения
- проведено сравнение различных архиваторов с собственным алгоритмом сжатия
- собственный алгоритм признан подходящим для применения в инфраструктуре поисковой системы