

Сжатие с учётом контекста. Словарные методы, где словарём является несжатый текст — семейство LZ77

Александра Игоревна Кононова

МИЭТ

27 ноября 2023 г. — актуальную версию можно найти на
<https://gitlab.com/illinc/otik>



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

О терминах

Размер байта — 3 бита

Источник с памятью

Семейство RLE (Run Length Encoding) — концепция

О терминах

Символ — элемент качественной информации $a \in A$ (множество A — алфавит).

Текст — последовательность $m \in A^+$ таких элементов.

На практике для всех алгоритмов, где алфавит может быть произвольным,
символ кодирования = байт (так как в большинстве ЭВМ байт 8-битен — это 00...FF),
исходный текст = любой бинарный файл, сжатый текст — тоже бинарный файл:

- использование в программе для ЭВМ символов меньших, чем байт — неудобно;
- использование символов фиксированной разрядности бóльших, чем байт \Rightarrow слишком большой алфавит \Rightarrow объёмные структуры данных для восстановления.
- использование в качестве символа кодирования печатного символа ASCII или koi8r/cp1251/dos/iso/maccyrillic не позволяет рассматривать в качестве исходного текста произвольный файл и приводит к труднодиагностируемым ошибкам;
- использование печатного символа UTF-8 (144 697 символов Unicode в 2023 г.) — то же самое + проигрыш в объёме.

В книгах для наглядности используются обозначения A, B, \dots и т. п. (маленький алфавит + визуальное отличие символа от индекса или частоты), но в программе это всё равно байты!

RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

О терминах

Размер байта — 3 бита

Источник с памятью

Семейство RLE (Run Length Encoding) — концепция

Размер байта – 3 бита

Для компактной иллюстрации ограничений алгоритмов примем, что для устройства «доска» байт (символ кодирования) составляет не 8 бит — октет (как для Intel x86/amd64), и не 6 бит (как для IBM 7030 Stretch), а 3 бита — триаду, или одну 8-ричную цифру:

0 = 000

1 = 001

2 = 010

3 = 011

4 = 100

5 = 101

6 = 110

7 = 111

Источник с памятью

Вероятность порождения элемента определяется предысторией источника
→ при сжатии необходимо учитывать контекст символа.

источник Маркова (N -го порядка) — состояние на i -м шаге зависит от состояний на N предыдущих шагах: $i - 1, i - 2, \dots, i - N \rightarrow$ сжимаются не отдельные символы, а устойчивые сочетания — слова: коды Зива—Лемпеля (LZ77, LZ78);

аналоговый сигнал — источник количественных данных Маркова 1-го порядка → кодирование длин повторений (Run Length Encoding, RLE).

Для задания конкретного источника Маркова N -го порядка: вероятности зависят от предыдущих N символов \implies нужно $T^{N+1} + T$ условных + безусловных вероятностей, где T — размер алфавита ($T = 2^k$).

Ни один код сжатия с контекстом не хранит вероятности \implies длина кода не оценивается по частотам сочетаний в файле.



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

О терминах

Размер байта — 3 бита

Источник с памятью

Семейство RLE (Run Length Encoding) — концепция

Семейство RLE (Run Length Encoding) — концепция

Модель источника данных — Маркова первого порядка (аналоговый сигнал), при этом:

$$\forall a \neq b: \begin{cases} p(a|a) = p(b|b) = r, \\ p(a|b) = p(b|a) = s, \end{cases} \quad r \gg (T-1)s, \quad \text{где } T — \text{размер алфавита.}$$

Run Length Encoding (RLE): AAAAAAAAAABCCCCC \rightarrow 8 \times A, 1 \times B, 4 \times C

Повторение символа c подряд L раз ($L \times c$) — цепочку длины L , $L_{\min} \leq L \leq L_{\max}$ — будем записывать как пару $\begin{Bmatrix} L \\ c \end{Bmatrix}$ (сжатая цепочка):

- цепочки длины более L_{\max} символов — делятся на несколько;
- последовательности символов, где ни один не повторяется L_{\min} раз подряд — несжатый текст.

RLE — не код, а **семейство кодов**, основанных на одном принципе сжатия и похожих моделях источника.

Способы отделения сжатых цепочек от несжатого текста; RLE-н-Δ

❶ Несжатого текста нет: $L_{\min} = 1$ («наивная» реализация RLE). Для наивного RLE:

- порядок записи L и c может быть любым: $\begin{Bmatrix} L \\ c \end{Bmatrix} \sim (L - \Delta, c)$ или $(c, L - \Delta)$;
- так как $L \geq L_{\min} = 1$, можно записывать L кодом со смещением:
 $\begin{Bmatrix} L \\ c \end{Bmatrix} \sim (L - 1, c), (c, L - 1)$ [макс. смещ.: $\Delta = 1$]; (L, c) или (c, L) [без смещения: $\Delta = 0$].

Рассматриваем $|L| = |c| = k$ бит, код с максимальным смещением и порядок $(L - 1, c)$:

$0 \leq L - 1 \leq 2^k - 1 \implies 1 \leq L \leq 2^k \implies L_{\max} = 2^k$; для трёхбитного байта ($k = 3$) $L_{\max} = 8$.

$m = 7777\ 7000\ 0000\ 0000\ 0123\ 4567\ 0011\ 2233\ 4455\ 6677$ (40 трёхбитных байтов)

$\rightarrow \begin{Bmatrix} L=5 \\ c=7 \end{Bmatrix} \begin{Bmatrix} L=8 \\ c=0 \end{Bmatrix} \begin{Bmatrix} L=4 \\ c=0 \end{Bmatrix} \begin{Bmatrix} L=1 \\ c=1 \end{Bmatrix} \begin{Bmatrix} L=1 \\ c=2 \end{Bmatrix} \dots \begin{Bmatrix} L=1 \\ c=7 \end{Bmatrix} \begin{Bmatrix} L=2 \\ c=0 \end{Bmatrix} \begin{Bmatrix} L=2 \\ c=1 \end{Bmatrix} \begin{Bmatrix} L=2 \\ c=2 \end{Bmatrix} \dots \begin{Bmatrix} L=2 \\ c=7 \end{Bmatrix}$

$code(m) = 4770\ 3001\ 0203\ 0405\ 0607\ 1011\ 1213\ 1415\ 1617$ (36 трёхбитных байтов)

❷ Несжатый текст тоже группируется в цепочки, различие — флаг-биты (RLE с флаг-битом).

❸ Несжатый текст записывается как есть, сжатые цепочки предваряются односимвольным префиксом (RLE с префиксом).



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Способы отделения сжатых цепочек от несжатого текста; RLE-н-Δ

RLE с флаг-битом (несжатый текст тоже группируется в цепочки), тезисы

RLE с флаг-битом — варианты RLE-фб $L_{\min}^{сж} - \Delta_{сж} \Delta_{несж}$

RLE с флаг-битом — варианты (продолжение) и выбор RLE-фб3-31

RLE с флаг-битом, кодирование с выбранными параметрами

RLE с флаг-битом (несжатый текст тоже группируется в цепочки), тезисы

Наивный RLE особенно плох на фрагментах типа 0123...: один байт невыгодно записывать двумя.

Зададимся $L_{\min}^{\text{сж}}$ таким, что $L \geq L_{\min}^{\text{сж}}$ повторяющихся байтов выгодно записывать двумя байтами;

такие последовательности байтов, где ни один не повторяется хотя бы $L_{\min}^{\text{сж}}$ подряд, тоже дополняем служебным байтом с общей длиной L — несжатая цепочка $\left\{ \begin{smallmatrix} L \\ c_1 \dots c_L \end{smallmatrix} \right\}$, где $L \geq 1$ включительно ($L_{\min}^{\text{несж}} = 1$).

- Для различения сжатых и несжатых цепочек (два вида) достаточно одного бита θ — флаг-бита. Флаг-бит θ и длина цепочки L упаковываются в один байт (k бит):

① 1 бит на θ (значения 0 и 1) и $k - 1$ бит на L ($0 \leq L - \Delta \leq 2^{k-1} - 1$);

② длина кода сжатой цепочки $\left\{ \begin{smallmatrix} L \\ c \end{smallmatrix} \right\}$: $|L \cup \theta| + |c| = 2$ байта,

длина кода несжатой $\left\{ \begin{smallmatrix} L \\ c_1 \dots c_L \end{smallmatrix} \right\}$: $|L \cup \theta| + |c_1 \dots c_L| = 1 + L$ байтов.

- Длина кода несжатой цепочки непостоянна \implies байт $L \cup \theta$ в начале.



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Способы отделения сжатых цепочек от несжатого текста; RLE-н- Δ
 RLE с флаг-битом (несжатый текст тоже группируется в цепочки), тезисы
 RLE с флаг-битом — варианты RLE-фб $L_{\min}^{\text{сж}} - \Delta_{\text{сж}} \Delta_{\text{несж}}$
 RLE с флаг-битом — варианты (продолжение) и выбор RLE-фб3-31
 RLE с флаг-битом, кодирование с выбранными параметрами

RLE с флаг-битом — варианты RLE-фб $L_{\min}^{\text{сж}} - \Delta_{\text{сж}} \Delta_{\text{несж}}$

- 1 Флаг-бит θ сж/несж может быть 0/1 или 1/0 — не влияет на длину кода.
- 2 Положение бита θ в байте $L \cup \theta$ — не влияет на длину кода; обычно θ — старший: удобнее читать дампы.
- 3 Выбор $L_{\min}^{\text{сж}}$ — между 2 и 3:

- однократное вхождение символа всегда невыгодно записывать двумя символами $\Rightarrow L_{\min}^{\text{сж}} > 1$;
- трёхкратное — всегда выгодно $\Rightarrow L_{\min}^{\text{сж}} \leq 3$.

Иногда выгоднее 2 (файл 001122), иногда — 3 (файл 0112); узнать без кодирования невозможно.

$L_{\max}^{\text{сж}}$ и $L_{\max}^{\text{несж}}$ не выбираются, а **рассчитываются** на основе разрядности $k - 1$ смещённого L и $\Delta_{\text{сж}}$ и $\Delta_{\text{несж}}$:
 $L_{\max} = 2^{k-1} - 1 + \Delta$ ($\Delta_{\text{сж}} \neq \Delta_{\text{несж}} \Rightarrow L_{\max}^{\text{сж}} \neq L_{\max}^{\text{несж}}$).

- 4 Код со смещением для L :
 - без смещ. ($\Delta_{\text{сж}} = \Delta_{\text{несж}} = 0$) \Rightarrow меньше $L_{\max} \Rightarrow$ самый длинный код для выбранного $L_{\min}^{\text{сж}}$;
 - минимальное <ненулевое> смещение: $L - 1$ как для сжатых, так и для несжатых ($\Delta_{\text{сж}} = \Delta_{\text{несж}} = 1$);
 - максимальное смещение: записывать $L - L_{\min}$ (разные для сжатых и несжатых цепочек:
 $\Delta_{\text{сж}} = L_{\min}^{\text{сж}}, \Delta_{\text{несж}} = 1$) \Rightarrow самый короткий код для выбранного $L_{\min}^{\text{сж}}$;
 - смещения между макс. и мин. допустимы, но не дают выигрыша ни в длине, ни в читаемости дампа.

Относительная величина проигрыша $\Delta_{\text{сж}} = \Delta_{\text{несж}} = 0$ относительно $\Delta_{\text{сж}} = L_{\min}^{\text{сж}}$ и $\Delta_{\text{несж}} = L_{\min}^{\text{несж}} = 1$ меньше для $k = 8$ (октета), чем для $k = 3$. Но проигрыш *есть*; и найдутся файлы, где он проявится.

⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↺ ↻

RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Способы отделения сжатых цепочек от несжатого текста; RLE-н- Δ
 RLE с флаг-битом (несжатый текст тоже группируется в цепочки), тезисы
 RLE с флаг-битом — варианты RLE-фб $L_{\min}^{\text{сж}} - \Delta_{\text{сж}} \Delta_{\text{несж}}$
 RLE с флаг-битом — варианты (продолжение) и выбор RLE-фб3-31
 RLE с флаг-битом, кодирование с выбранными параметрами

RLE с флаг-битом — варианты (продолжение) и выбор RLE-фб3-31

Экзотические варианты кодирования: выбор $L_{\min}^{\text{сж}} > 3$ (в частности, 4), или запись в $k - 1$ бит длины не $L - \Delta$, а, например, $\frac{L}{2}$ (рассматривать только чётные L , а «лишний» байт добавлять к несжатой цепочке — технически **возможны**).

Почти для любого «безумного» кода найдётся файл, для которого именно этот код будет лучшим!

И только код без смещения $L = \text{код с неиспользуемыми кодовыми позициями} = \text{код с заведомой избыточностью}$ всегда не лучше (и на некоторых файлах хуже) аналога без избыточности.

Но все варианты RLE с флаг-битом рассмотреть в рамках лекций невозможно — слишком много.

Ниже используются:

- ❶ флаг-бит θ сж/несж — 1/0;
- ❷ в k -битном байте $|L \cup \theta|$ флаг θ — старший бит байта $L \cup \theta$, $L - L_{\min}$ в $(k - 1)$ младших; $k = 3$;
- ❸ $L_{\min}^{\text{сж}} = 3$;

$$\text{❹ Макс. смещение: } \begin{cases} \left\{ \begin{matrix} L^{\text{сж}} \\ c \end{matrix} \right\} \rightarrow ((\theta = 1, L^{\text{сж}} - 3), c), & 3 \leq L^{\text{сж}} \leq 2^{k-1} + 2 \\ \left\{ \begin{matrix} L^{\text{несж}} \\ c_1 \dots c_L \end{matrix} \right\} \rightarrow ((\theta = 0, L^{\text{несж}} - 1), c_1 \dots c_L), & 1 \leq L^{\text{несж}} \leq 2^{k-1} \end{cases}$$



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Способы отделения сжатых цепочек от несжатого текста; RLE-н- Δ
 RLE с флаг-битом (несжатый текст тоже группируется в цепочки), тезисы
 RLE с флаг-битом — варианты RLE-фб $L_{\min}^{\text{сж}} - \Delta_{\text{сж}} \Delta_{\text{несж}}$
 RLE с флаг-битом — варианты (продолжение) и выбор RLE-фб3-31
 RLE с флаг-битом, кодирование с выбранными параметрами

RLE с флаг-битом, кодирование с выбранными параметрами

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} L^{\text{сж}} \\ c \end{array} \right\} \\ \left\{ \begin{array}{l} L^{\text{несж}} \\ c_1 \dots c_L \end{array} \right\} \end{array} \right\} \rightarrow \left((\theta = 1, L^{\text{сж}} - 3), c \right), \quad 3 \leq L^{\text{сж}} \leq 6$$

$$\rightarrow \left((\theta = 0, L^{\text{несж}} - 1), c_1 \dots c_L \right), \quad 1 \leq L^{\text{несж}} \leq 4$$

$m = 7777\ 7000\ 0000\ 0000\ 0123\ 4567\ 0011\ 2233\ 4455\ 6677$ (40 трёхбитных байтов)

$$\left\{ \begin{array}{l} 1 \\ 5 \\ 7 \end{array} \right\} \left\{ \begin{array}{l} 1 \\ 6 \\ 0 \end{array} \right\} \left\{ \begin{array}{l} 1 \\ 6 \\ 0 \end{array} \right\} \left\{ \begin{array}{l} 0 \\ 4 \\ 1234 \end{array} \right\} \left\{ \begin{array}{l} 0 \\ 4 \\ 5670 \end{array} \right\} \left\{ \begin{array}{l} 0 \\ 4 \\ 0112 \end{array} \right\} \left\{ \begin{array}{l} 0 \\ 4 \\ 2334 \end{array} \right\} \left\{ \begin{array}{l} 0 \\ 4 \\ 4556 \end{array} \right\} \left\{ \begin{array}{l} 0 \\ 3 \\ 677 \end{array} \right\}$$

$$\left((1, 5 - 3 = 2), 7 \right) \left((1, 6 - 3 = 3), 0 \right) \left((1, 6 - 3 = 3), 0 \right) \left((0, 4 - 1 = 3), 1234 \right) \left((0, 4 - 1 = 3), 5670 \right)$$

$$\left((0, 4 - 1 = 3), 0112 \right) \left((0, 4 - 1 = 3), 2334 \right) \left((0, 4 - 1 = 3), 4556 \right) \left((0, 3 - 1 = 2), 677 \right)$$

$(110, 7) (111, 0) (111, 0) (011, 1234) (011, 5670) (011, 0112) (011, 2334) (011, 4556) (010, 677)$

$code(m) = 67707031234356703011232334345562677$ (35 трёхбитных байтов)

RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Способы отделения сжатых цепочек от несжатого текста; RLE-н-Δ
RLE с флаг-битом (несжатый текст тоже группируется в цепочки), тезисы
RLE с флаг-битом — варианты RLE-фб $L_{\min}^{\text{сж}} - \Delta_{\text{сж}} \Delta_{\text{несж}}$
RLE с флаг-битом — варианты (продолжение) и выбор RLE-фб3-31
RLE с флаг-битом, кодирование с выбранными параметрами

RLE с односимвольным префиксом, тезисы

Текст, который не имеет смысла записывать как сжатые цепочки (несжатый текст) — пишется как есть, без дополнительных служебных байтов.

Чтобы отделить сжатые цепочки $\left\{ \begin{smallmatrix} L \\ c \end{smallmatrix} \right\}$ от несжатого текста, они предваряются *префиксом* — выбираемым для каждого файла индивидуально байтом p ; то есть цепочка записывается тремя байтами.

- длина кода цепочки $\left\{ \begin{smallmatrix} L \\ c \end{smallmatrix} \right\}$ 3 байта \implies имеет смысл только при $L \geq 4$;
- в несжатом тексте может встретиться любой байт, в том числе $p \implies$ нужно экранировать p , pp и ppp (четырёхкратное $pppp$ в любом случае цепочка $\left\{ \begin{smallmatrix} L=4 \\ p \end{smallmatrix} \right\}$);
- любой способ экранирования одного байта p в несжатом тексте длиннее байта $\implies p$ выбирается для конкретного файла как самый редкий байт (в идеале — отсутствующий); следовательно, $\text{count}(p) \leq \frac{n}{2^k}$;
- кодирование в два прохода по файлу: 1) поиск p , 2) сжатие RLE (наивный и флаг-биты — в один проход);
- значение p сохраняется в заголовке файла — нужно для декодирования.



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

RLE с односимвольным префиксом, тезисы

RLE с префиксом — варианты RLE- $p1p/p0p0p0/p0p2p/p0p0p3p-\Delta_{c \neq p} \Delta_p$

RLE с префиксом, одиночный p как цепочка [RLE-p1p-41]

RLE с префиксом, одиночный p как $p0$, сдвоенный как цепочка [RLE-p0p2p-3

RLE с префиксом — варианты RLE- $p1p/p0p0p0/p0p2p/p0p0p3p-\Delta_{c \neq p} \Delta_p$

❶ Способ экранирования в несжатом тексте p , pp и ppp :

❶ p как цепочка $\left\{ \begin{smallmatrix} 1 \\ p \end{smallmatrix} \right\}$, pp как $\left\{ \begin{smallmatrix} 2 \\ p \end{smallmatrix} \right\}$, ppp как $\left\{ \begin{smallmatrix} 3 \\ p \end{smallmatrix} \right\} \Rightarrow$ а) $L_{\min}^p = 1$; $L_{\min}^{c \neq p} = 4$; б) порядок любой:
 $(p, L - \Delta, c)$ или $(p, c, L - \Delta)$, причём $\Delta \leq L_{\min}$;

❷ p как $p0$, pp как $p0p0$, ppp как $p0p0p0 \Rightarrow$ а) $L_{\min}^p = L_{\min}^{c \neq p} = 4$; б) код $\left\{ \begin{smallmatrix} L \\ c \end{smallmatrix} \right\}$ не должен начинаться
с $p0 \Rightarrow$ следующий после p байт ненулевой \Rightarrow порядок $(p, \underbrace{L - \Delta}_{\neq 0}, c)$, причём $\Delta \leq L_{\min} - 1 = 3$;

❸ p как $p0$, но pp как цепочка $\left\{ \begin{smallmatrix} 2 \\ p \end{smallmatrix} \right\} \Rightarrow$ а) $L_{\min}^p = 2$; $L_{\min}^{c \neq p} = 4$; б) порядок $(p, \underbrace{L - \Delta}_{\neq 0}, c)$, $\Delta \leq L_{\min} - 1$;

❹ p как $p0$, pp как $p0p0$, ppp как $\left\{ \begin{smallmatrix} 3 \\ p \end{smallmatrix} \right\} \Rightarrow$ а) $L_{\min}^p = 3 \dots$

❷ Код со смещением для L — аналогично предыдущим; макс. Δ_p и $\Delta_{c \neq p}$ — в общем случае разные.

При макс. смещении L для любого из способов экранирования ❶–❹ найдётся файл, где этот способ будет лучше трёх остальных (где хуже код p , pp и ppp — там лучше L_{\max}).



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флагом-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

RLE с односимвольным префиксом, тезисы

RLE с префиксом — варианты RLE- $p1p/p0p0p0/p0p2p/p0p0p3p-\Delta_{c \neq p} \Delta_p$

RLE с префиксом, одиночный p как цепочка [RLE-p1p-41]

RLE с префиксом, одиночный p как $p0$, вдвоенный как цепочка [RLE-p0p2p-3]

RLE с префиксом, одиночный p как цепочка [RLE-p1p-41]

Выбираем порядок (p, L, c) и код с максимальным смещением:

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} L \\ c \neq p \end{array} \right\} \rightarrow (p, L-4, c), \quad 0 \leq L^{c \neq p} - 4 \leq 2^k - 1 \iff 4 \leq L^{c \neq p} \leq 2^k + 3 \\ \left\{ \begin{array}{l} L \\ p \end{array} \right\} \rightarrow (p, L-1, p), \quad 0 \leq L^p - 1 \leq 2^k - 1 \iff 1 \leq L^p \leq 2^k \end{array} \right.$$

$$k = 3: \quad 4 \leq L^{c \neq p} \leq 11, \quad 1 \leq L^p \leq 8$$

$m = 7777\ 7000\ 0000\ 0000\ 0123\ 4567\ 0011\ 2233\ 4455\ 6677$ (40 трёхбитных байтов)

самые редкие $\{1, 2, 3, 4, 5, 6\}$ — выбираем префикс $p = 1$

$$\left\{ \begin{array}{l} 5 \\ 7 \neq p \end{array} \right\} \left\{ \begin{array}{l} 11 \\ 0 \neq p \end{array} \right\} 0 \left\{ \begin{array}{l} 1 \\ 1 = p \end{array} \right\} 23456700 \left\{ \begin{array}{l} 2 \\ 1 = p \end{array} \right\} 223344556677$$

$$\left\{ \begin{array}{l} 5-4=1 \\ 7 \neq p \end{array} \right\} \left\{ \begin{array}{l} 11-4=7 \\ 0 \neq p \end{array} \right\} 0 \left\{ \begin{array}{l} 1-1=0 \\ 1 = p \end{array} \right\} 23456700 \left\{ \begin{array}{l} 2-1=1 \\ 1 = p \end{array} \right\} 223344556677$$

$code(m) = 117170010123456700111223344556677$ (33 трёхбитных байта)

RLE с префиксом, одиночный p как $p0$, сдвоенный как цепочка [RLE-p0p2p-31]

Выбираем код с максимальным смещением:

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} L \\ c \neq p \end{array} \right\} \rightarrow (p, L-3, c), \quad 1 \leq L^{c \neq p} - 3 \leq 2^k - 1 \iff 4 \leq L^{c \neq p} \leq 2^k + 2 \\ \left\{ \begin{array}{l} 1 \\ p \end{array} \right\} \rightarrow (p, 0), \\ \left\{ \begin{array}{l} L \geq 2 \\ p \end{array} \right\} \rightarrow (p, L-1, p), \quad 1 \leq L^p - 1 \leq 2^k - 1 \iff 2 \leq L^p \leq 2^k \end{array} \right.$$

$$k = 3: \quad 4 \leq L^{c \neq p} \leq 10, \quad 2 \leq L^p \leq 8$$

$m = 7777\ 7000\ 0000\ 0000\ 0123\ 4567\ 0011\ 2233\ 4455\ 6677$ (40 трёхбитных байтов), выбираем $p = 1$

$$\left\{ \begin{array}{l} 5 \\ 7 \neq p \end{array} \right\} \left\{ \begin{array}{l} 10 \\ 0 \neq p \end{array} \right\} 00 \left\{ \begin{array}{l} 1 \\ 1 = p \end{array} \right\} 23456700 \left\{ \begin{array}{l} 2 \\ 1 = p \end{array} \right\} 223344556677$$

$$\left\{ \begin{array}{l} 5-3=2 \\ 7 \neq p \end{array} \right\} \left\{ \begin{array}{l} 10-3=7 \\ 0 \neq p \end{array} \right\} 00(p, 0) 23456700 \left\{ \begin{array}{l} 2-1=1 \\ 1 = p \end{array} \right\} 223344556677$$

$$code(m) = 127170001023456700111223344556677 \text{ (33 трёхбитных байта)}$$

Лучший и худший случаи RLE (для октетов: $k = 8$)

Лучший исходный файл длины n для всех RLE — $m_n = \underbrace{0000 \dots 0000}_n$;

худший — $m_x = \underbrace{0123 \dots 0123 \dots}_n$ (все байты от 0 до $2^k - 1$ [октеты от 00 до FF]; нет ни одного повторения).

Считаем $n \gg L_{\max}$ («хвостом» можно пренебречь);

L_{\max} оцениваем по порядку величины (без смещения будет на 1 меньше, с макс. — немного больше).

Все длины $|code|$ оцениваются приблизительно (кроме $|code(m_x)|$ для RLE-н).

$code$	L_{\max}	$L_{\max} _{k=8}$	$ code(m_n) $	$ code(m_n) _{k=8}$	$ code(m_x) $	$ code(m_x) _{k=8}$
RLE-н	2^k	256	$2 \cdot \frac{n}{L_{\max}} \approx \frac{n}{2^{k-1}}$	$\frac{n}{128} = 2 \cdot \frac{n}{256}$	$= 2 \cdot n$	$2 \cdot n$
RLE-фб	2^{k-1}	128	$2 \cdot \frac{n}{L_{\max}} \approx \frac{n}{2^{k-2}}$	$\frac{n}{64} = 4 \cdot \frac{n}{256}$	$(1 + \frac{1}{L_{\max}}) \cdot n$	$(1 + \frac{1}{128}) \cdot n$
RLE- $p0^*$	2^k	256	$3 \cdot \frac{n}{L_{\max}} \approx \frac{3n}{2^k}$	$3 \cdot \frac{n}{256}$	$n + count(p) \leq (1 + \frac{1}{2^k}) \cdot n$	$(1 + \frac{1}{256}) \cdot n$
RLE- $p1p$					$n + 2count(p) \leq (1 + \frac{1}{2^{k-1}}) \cdot n$	$(1 + \frac{1}{128}) \cdot n$

RLE- p^* лучше RLE-фб в наилучшем и наихудшем случаях; но есть файлы, где RLE-фб лучше.

RLE-н лучше всех прочих в наилучшем случае (все байты входят сериями по L_{\max} раз).



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флагом-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Лучший и худший случаи RLE (для октетов: $k = 8$)

Модель источника наивного RLE — построение

Модель источника наивного RLE — расчёт

Задачи: модель источника RLE

Модель источника наивного RLE — построение

Существует ли модель аналогового сигнала, которая позволяет **точно** оценить снизу размер кода? Попробуем составить для k -битного байта (алфавит — из $T = 2^k$ байтов) и RLE-н-1 ($L - 1, c$).

Из симметрии RLE-н $\forall a \neq b: \begin{cases} p(a|a) = p(b|b) = r, \\ p(a|b) = p(b|a) = s. \end{cases} (*)$ Из предположений RLE $r \gg (T - 1)s$.

- 1 Модель X_{62} построим из соображений баланса: длина кода RLE-н двух цепочек из L_{max} одинаковых символов a равна длине кода двух разных символов ab ; если RLE-н соответствует источнику X_1 , то вероятности их тоже равны:

$$p(\underbrace{aaa..aaa}_{L_{max}} \underbrace{aaa..aaa}_{L_{max}}) = p(ab) \implies p(a) \cdot r^{2L_{max}-1} = p(a) \cdot s \implies r^{2L_{max}-1} = s$$

- 2 $X_{л\alpha}$ — по лучшему случаю: длина кода $\underbrace{aaa..aaa}_{\alpha \cdot L_{max}} — 2\alpha$ байтов ($2\alpha k$ бит):

$$p(\underbrace{aaa..aaa}_{\alpha \cdot L_{max}}) = \frac{1}{2^{2\alpha k}} \implies \frac{1}{T} \cdot r^{\alpha \cdot L_{max}-1} = \frac{1}{2^{2\alpha k}}$$

- 3 $X_{х\alpha}$ — по худшему случаю: длина кода $\underbrace{abc...}_{\alpha} — 2\alpha$ байтов ($2\alpha k$ бит):

$$p(\underbrace{abc...}_{\alpha}) = \frac{1}{2^{2\alpha k}} \implies \frac{1}{T} \cdot s^{\alpha-1} = \frac{1}{2^{2\alpha k}}$$

В $X_{л\alpha}$ и $X_{х\alpha}$ используется $p(a) = \frac{1}{T}$, что верно только для начала сообщения.

RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Лучший и худший случаи RLE (для октетов: $k = 8$)

Модель источника наивного RLE — построение

Модель источника наивного RLE — расчёт

Задачи: модель источника RLE

Модель источника наивного RLE — расчёт

Найдём r, s с учётом $L_{max} = 2^k = T$ и $r + (T - 1)s = 1$, а также $(T - 1)s \ll r < 1$:

$$\textcircled{1} X_{62}: \quad r^{2L_{max}-1} = s \quad \Longleftrightarrow \quad r^{2T-1} = s \quad \Longleftrightarrow \quad \left(1 - (T - 1)s\right)^{2T-1} = s$$

Решаем численно: шаг 1 — из приближения $1 - (T - 1) \cdot (2T - 1)s \approx s$ начальное $s_0 = \frac{1}{(T-1) \cdot (2T-1) + 1}$;

шаг 2 — уточняем численно решение $\left(1 - (T - 1)s\right)^{2T-1} = s$ от s_0 .

Если аналогично $X_{л\alpha}$ и $X_{х\alpha}$ рассмотреть $X_{6\alpha}$ как « $\alpha \cdot L_{max}$ одинаковых символов равновероятны α разным», r с ростом α стремится к 1, а s к 0.

$$\textcircled{2} X_{л\alpha}: \quad \frac{1}{T} \cdot r^{\alpha \cdot L_{max}-1} = \frac{1}{2^{2\alpha k}} \quad \Longleftrightarrow \quad \frac{1}{T} \cdot r^{\alpha T-1} = \frac{1}{T^{2\alpha}} \quad \Longleftrightarrow \quad r = T^{-\frac{2\alpha-1}{\alpha T-1}} = T^{-\frac{2-\frac{1}{\alpha}}{T-\frac{1}{\alpha}}}$$

$$\textcircled{3} X_{х\alpha}: \quad \frac{1}{T} \cdot s^{\alpha-1} = \frac{1}{2^{2\alpha k}} \quad \Longleftrightarrow \quad \frac{1}{T} \cdot s^{\alpha-1} = \frac{1}{T^{2\alpha}} \quad \Longleftrightarrow \quad s = T^{-\frac{2\alpha-1}{\alpha-1}} = T^{-\frac{2-\frac{1}{\alpha}}{1-\frac{1}{\alpha}}}$$

Для октетов: $k = 8$, $T = 256$

	$X_{л\infty}$	$X_{л1}$	X_{62}	$X_{х\infty}$	$X_{х2}$	$X_{6\infty}$
r	0,9576	0,9785	0,9815	0,9961	$1 - 1,5 \cdot 10^{-5}$	1
s	$1,7 \cdot 10^{-4}$	$8 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$1,5 \cdot 10^{-5}$	$6 \cdot 10^{-8}$	0

RLE-н-1 имеет смысл при $p(a|a) \approx r_{л\infty}$, при этом избыточен для такого источника.



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Лучший и худший случаи RLE (для октетов: $k = 8$)

Модель источника наивного RLE — построение

Модель источника наивного RLE — расчёт

Задачи: модель источника RLE

Задачи: модель источника RLE

- 1 Оценить $I(m)$ по моделям 1, 2, 3 для сообщений длины $n \rightarrow \infty$
 $m_1 = 0000...0000$ (наилучший случай) и $m_2 = 0123...4567$ (наихудший
 случай); сравнить с длиной наивного RLE ($L - 1, c$) этих сообщений
 (рассчитать $\frac{code(m)}{I_i(m)}$).
- 2 При каком r имеет смысл наивный RLE с двухоктетным L ?
- 3 Построить модель/модели источника для оценки снизу длины кода
 RLE с флаг-битом (не обязательно с максимальным смещением L).
 Существует ли простая (марковская первого порядка) модель для
 точной оценки?
 Как отличаются вероятности от модели наивного RLE?

Концепция LZ77, код Зива–Лемпеля LZ77/LZ1

1977 г., Якоб Зив (Jacob Ziv) и Абрахам Лемпель (Abraham Lempel)

J. Ziv and A. Lempel, «A universal algorithm for sequential data compression», in IEEE Transactions on Information Theory, vol. 23, no. 3, pp. 337-343, May 1977, doi: 10.1109/TIT.1977.1055714.

— идея замены слова ссылкой $\begin{Bmatrix} S \\ L \end{Bmatrix}$ и концепт $(S, L, c)/(0, 0, c)$:

если цепочка символов (не обязательно одинаковых) длины $L_{\min} \leq L \leq L_{\max}$ (слово) встречается более одного раза, то каждое **следующее вхождение** слова **заменяется ссылкой на предыдущее**.

Ссылка $\begin{Bmatrix} S \\ L \end{Bmatrix}$ состоит из:

- относительного смещения $S \geq 1$ предыдущего вхождения слова относительно текущей позиции;
- длины L слова.

Скользящее окно: область перед текущей позицией кодирования, в которой можем искать и адресовать ссылки ($S_{\min} \leq S \leq S_{\max}$).

Поиск выполняется по несжатому тексту (при декодировании — по **уже разжатой** части)!

В окне поиска нет ни ссылок $\begin{Bmatrix} S \\ L \end{Bmatrix}$, ни флаг-байтов, ни экранирующих нулей, ни иных служебных структур.



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепция LZ77, код Зива–Лемпеля LZ77/LZ1

Неоднозначность кодирования LZ77

Семейство LZ77 — варианты

Выбор для примеров ниже (реализовать иначе — можно, но не всё стоит)

Неоднозначность кодирования LZ77

В окне $1 \leq S \leq S_{\max}$ может быть несколько подходящих совпадений $L_{\min} \leq L \leq L_{\max}$:

$S_{\min} = 1$ всегда. Пусть $S_{\max} = 16$, $L_{\min} = 4$, $L_{max} = 6$.

16 15 14 13 12 11 10 09 08 07 06 04 03 02 01 ↓
 ... 7 6 5 1 1 2 3 1 2 5 1 2 3 1 6 1 2 3 1 2 3 1 1 ...

① $\begin{cases} S_1 = 15 \\ L_1 = 5 \end{cases}$, ② $\begin{cases} S_2 = 9 \\ L_2 = 4 \end{cases}$, ③ $\begin{cases} S_3 = 3 \\ L_3 = 6 \end{cases}$ (при большем L_{\max} было бы 7).

Любая одна из этих пар будет допустимым кодом слова в текущей (↓) позиции.

- полный перебор всех смещений ($S_{\min} \dots S_{\max}$) или ускоренный, по вспомогательным структурам данных (как в LZSS)
- поиск первого совпадения длины L_{\min} (слабее сжатие) или поиск наилучшего (медленно);
- по возрастанию абсолютного адреса (от S_{\max} к S_{\min}) или по убыванию и т. д.

Поиск предыдущего вхождения — основная сложность семейства LZ77 и неоднозначность кодирования.

Декодирование любого кода семейства LZ77 однозначно.

Семейство LZ77 — варианты

LZ77, как и RLE — **семейство кодов**, основанных на одной идее; так как семейство LZ77 чаще применяется на практике — многие коды семейства получили собственные имена. Отличаются:

- 1 способом отделения ссылок $\left\{ \begin{smallmatrix} L \\ S \end{smallmatrix} \right\}$ от несжатого текста (концепт Зива–Лемпеля / флаг-байт ссылка/символ / флаг-бит ссылка/цепочка / префикс и т. д.);
- 2 порядком $(L, S)/(S, L)$ для тех случаев, где порядок не определён способом отделения;
- 3 выбором L_{\min} : значение $\inf(L_{\min})$, то есть наименьшее L такое, что ссылка $\left\{ \begin{smallmatrix} L \\ S \end{smallmatrix} \right\}$ занимает меньший объём, чем несжатые L символов, определяется способом кодирования; но можно задать $L_{\min} > \inf(L_{\min})$ (в частности, $L_{\min} = 4$ вместо 3 — хуже сжатие, но быстрее поиск).
- 4 разрядностью L и S : для сильно избыточных файлов лучше $|L| = |S| = k$ бит (больше L_{\max}), для малоизбыточных $|L| = k - \delta$ бит, $|S| = k + \delta$ бит (больше окно).
Но слишком большое окно замедляет кодирование — 16 разрядов много.
- 5 (не)использованием для L и S кода со смещением и другими деталями реализации.

Алгоритм семейства LZ77 = код + алгоритм поиска оригинала в окне. Так, алгоритм LZSS (Сторер, Сжимански) — флаг-бит ссылка/символ и битовый выходной поток + дерево для ускорения поиска.



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепция LZ77, код Зива–Лемпеля LZ77/LZ1

Неоднозначность кодирования LZ77

Семейство LZ77 — варианты

Выбор для примеров ниже (реализовать иначе — можно, но не всё стоит)

Выбор для примеров ниже (реализовать иначе — можно, но не всё стоит)

- 1 Символ=байт, $|c| = 1 \text{ байт} = k \text{ бит}$ (примеры для $k = 3$). Выходной поток байтовый. $S_{\min} = 1$.
- 2 Принимаем $L_{\min} = \inf(L_{\min})$, не увеличиваем (если способ кодирования не даёт однозначного $\inf(L_{\min})$ — берём наибольший из вариантов).
- 3 Рассматриваем на примере малоизбыточных файлов \implies окно S_{\max} желательно вчетверо больше алфавита, минимум вдвое \implies разрядность S от $k + 2$, минимум от $k + 1$.

Считаем, что L и S в сумме занимают **ровно два байта**. 3-битный байт, примеры:

$|L| = 2 \text{ бита}$, $|S| = 4 \text{ бита}$ ($|L| : |S| = 2 : 4$) — двухбитное L дополняется старшим битом S до байта; Октет: есть реализация $|L| : |S| = 6 : 10$ (LZJB).

Флаг-биты ссылка/цепочка рассматриваем на примере $|S| = 1 \text{ байт} = 3 \text{ бита}$.

- 4 Везде, где можно, **для L и S используется код со смещением**.
Обозначим смещённое S как $\tilde{S} = S - \Delta_S$ и смещённое L как $\tilde{L} = L - \Delta_L$.

Концепт Зива—Лемпеля [LZ77-к]

Ссылки чередуются с несжатыми символами: ссылка $\begin{Bmatrix} S \\ L \end{Bmatrix}$, за которой следует c — тройка (S, L, c) ; если в окне не на что дать ссылку — тройка $(0, 0, c)$.

При необходимости исходный текст дополняется (обычно нулями).

$L_{\min} = S_{\min} = 1$, ноль используется как специальное \Rightarrow код со смещением невозможен.

Примем $k = 3$, $|L| : |S| = 2 : 4 \Rightarrow S_{\max} = 2^4 - 1 = 15$, $L_{\max} = 2^2 - 1 = 3$

$m = 7777\ 7000\ 0000\ 0000\ 0123\ 4567\ 0123\ 4343\ 7012\ 1077$ (40 трёхбитных байтов)

$$\begin{Bmatrix} 0 \\ 0 \\ - \end{Bmatrix} 7 \begin{Bmatrix} S=1 \\ L=3 \end{Bmatrix} 7 \begin{Bmatrix} 0 \\ 0 \\ - \end{Bmatrix} 0 \begin{Bmatrix} S=1 \\ L=3 \end{Bmatrix} 0 \begin{Bmatrix} S=1 \\ L=3 \end{Bmatrix} 0 \begin{Bmatrix} S=1 \\ L=3 \end{Bmatrix} 1 \begin{Bmatrix} 0 \\ 0 \\ - \end{Bmatrix} 2 \begin{Bmatrix} 0 \\ 0 \\ - \end{Bmatrix} 3 \begin{Bmatrix} 0 \\ 0 \\ - \end{Bmatrix} 4 \begin{Bmatrix} 0 \\ 0 \\ - \end{Bmatrix} 5 \begin{Bmatrix} 0 \\ 0 \\ - \end{Bmatrix} 6 \begin{Bmatrix} 0 \\ 0 \\ - \end{Bmatrix} 7$$

$$\begin{Bmatrix} S=8 \\ L=3 \end{Bmatrix} 3 \begin{Bmatrix} S=8 \\ L=1 \end{Bmatrix} 3 \begin{Bmatrix} S=2 \\ L=2 \end{Bmatrix} 7 \begin{Bmatrix} S=9 \\ L=3 \end{Bmatrix} 1 \begin{Bmatrix} S=4 \\ L=1 \end{Bmatrix} 7 \begin{Bmatrix} S=1 \\ L=1 \end{Bmatrix} 0$$

— 18 троек $\begin{Bmatrix} S \\ L \end{Bmatrix} + c$, каждая записывается тремя байтами \Rightarrow 54 байта в коде.

$code(m) = 007137000130...$

LZ77 с флаг-байтами ссылка/символ [LZ77-cc]

Символы и ссылки группируются по k штук, каждая группа предваряется байтом, где каждый бит показывает тип соответствующего объекта (флаг-бит). Без группировки — битовый выходной поток.

При необходимости исходный текст дополняется (обычно нулями).

- флаг-бит θ ссылка/символ может быть 0/1 или 1/0; ниже рассматривается 1/0;
- порядок S и L в ссылке любой; примем (S, L) ;
- длина кода ссылки $\begin{Bmatrix} S \\ L \end{Bmatrix}$ с учётом флаг-бита $2 + \frac{1}{k}$ байта, длина кода несжатого символа — $1 + \frac{1}{k}$
 $\Rightarrow L_{\min} = 2$ (ссылка на двухбайтовое слово занимает $2 + \frac{1}{k}$ байта, два несжатых байта $2 + \frac{2}{k}$).
- возможен код со смещением: $\tilde{S} = S - S_{\min} = S - 1$, $\tilde{L} = L - L_{\min} = L - 2$.

В наихудшем случае к файлу из n байтов добавляется $\frac{n}{k}$ флаг-байтов.

$$\text{Примем } k = 3, |L| : |S| = 2 : 4 \Rightarrow \begin{cases} 0 \leq S - 1 \leq 2^4 - 1 \\ 0 \leq L - 2 \leq 2^2 - 1 \end{cases} \Rightarrow \begin{cases} 1 \leq S \leq 2^4 \\ 2 \leq L \leq 2^2 + 1 \end{cases} \Rightarrow \begin{cases} S_{\max} = 16 \\ L_{\max} = 5 \end{cases}$$

$m = 7777\ 7000\ 0000\ 0000\ 0123\ 4567\ 0123\ 4343\ 7012\ 1077$ (40 трёхбитных байтов)

$$7 \begin{Bmatrix} S=1 \\ L=4 \end{Bmatrix} 0 \begin{Bmatrix} S=1 \\ L=5 \end{Bmatrix} \begin{Bmatrix} S=1 \\ L=5 \end{Bmatrix} 01234567 \begin{Bmatrix} S=8 \\ L=5 \end{Bmatrix} \begin{Bmatrix} S=2 \\ L=3 \end{Bmatrix} \begin{Bmatrix} S=9 \\ L=4 \end{Bmatrix} 1077$$



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепт Зива—Лемпеля [LZ77-к]

LZ77 с флаг-байтами ссылка/символ [LZ77-cc]

LZ77 с флаг-байтами ссылка/цепочка [LZ77-cc, предложен студентом МИЭТ]

LZ77 с односимвольным префиксом [LZ77-p]

Лучший и худший случаи LZ77 (для октетов: $k = 8$)

Сравнение с RLE

Модель источника для LZ77

LZ77 с флаг-байтами ссылка/символ [LZ77-сс] (продолжение): ссылка/символ 1/0

$$7 \left\{ \begin{matrix} S=1 \\ L=4 \end{matrix} \right\} 0 \left\{ \begin{matrix} S=1 \\ L=5 \end{matrix} \right\} \left\{ \begin{matrix} S=1 \\ L=5 \end{matrix} \right\} 01234567 \left\{ \begin{matrix} S=8 \\ L=5 \end{matrix} \right\} \left\{ \begin{matrix} S=2 \\ L=3 \end{matrix} \right\} \left\{ \begin{matrix} S=9 \\ L=4 \end{matrix} \right\} 1077$$

группировка по $k = 3$ объекта (дополняем до $3x$ нулём); код со смещением $\tilde{S} = S - 1$, $\tilde{L} = L - 2$:

$$(010)_7 \left\{ \begin{matrix} 1-1 \\ 4-2 \end{matrix} \right\} 0(110) \left\{ \begin{matrix} 1-1 \\ 5-2 \end{matrix} \right\} \left\{ \begin{matrix} 1-1 \\ 5-2 \end{matrix} \right\} 0(000)123(000)456(001)_7 \left\{ \begin{matrix} 8-1 \\ 5-2 \end{matrix} \right\} \left\{ \begin{matrix} 2-1 \\ 3-2 \end{matrix} \right\} (100) \left\{ \begin{matrix} 9-1 \\ 4-2 \end{matrix} \right\} 10(000)770$$

7 флаг-байтов + 6 двухбайтовых ссылок + 15 несжатых символов — всего $7 + 6 \cdot 2 + 15 = 34$ байта

$$27 \left\{ \begin{matrix} 0 \\ 2 \end{matrix} \right\} 06 \left\{ \begin{matrix} 0 \\ 3 \end{matrix} \right\} \left\{ \begin{matrix} 0 \\ 3 \end{matrix} \right\} 00123045617 \left\{ \begin{matrix} 7 \\ 3 \end{matrix} \right\} \left\{ \begin{matrix} 1 \\ 1 \end{matrix} \right\} 4 \left\{ \begin{matrix} 8 \\ 2 \end{matrix} \right\} 100770$$

$$27 \left\{ \begin{matrix} 0000 \\ 10 \end{matrix} \right\} 06 \left\{ \begin{matrix} 0000 \\ 11 \end{matrix} \right\} \left\{ \begin{matrix} 0000 \\ 11 \end{matrix} \right\} 00123045617 \left\{ \begin{matrix} 0111 \\ 11 \end{matrix} \right\} \left\{ \begin{matrix} 0001 \\ 01 \end{matrix} \right\} 4 \left\{ \begin{matrix} 1000 \\ 10 \end{matrix} \right\} 100770$$

старший бит \tilde{S} записывается в байт, хранящий \tilde{L} :

$$27 \left\{ \begin{matrix} 000 \\ 010 \end{matrix} \right\} 06 \left\{ \begin{matrix} 000 \\ 011 \end{matrix} \right\} \left\{ \begin{matrix} 000 \\ 011 \end{matrix} \right\} 00123045617 \left\{ \begin{matrix} 111 \\ 011 \end{matrix} \right\} \left\{ \begin{matrix} 001 \\ 001 \end{matrix} \right\} 4 \left\{ \begin{matrix} 000 \\ 110 \end{matrix} \right\} 100770$$

$$\left\{ \begin{matrix} S \\ L \end{matrix} \right\} \text{ записывается в порядке } (S, L): \text{code}(m) = 2702060303001230456177311406100770 \text{ (34 трёхбитных байта)}$$

RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепт Зива—Лемпеля [LZ77-к]

LZ77 с флаг-байтами ссылка/символ [LZ77-сс]

LZ77 с флаг-байтами ссылка/цепочка [LZ77-сц, предложен студентом МИЭТ]

LZ77 с односимвольным префиксом [LZ77-р]

Лучший и худший случаи LZ77 (для октетов: $k = 8$)

Сравнение с RLE

Модель источника для LZ77

LZ77 с флаг-битами ссылка/цепочка [LZ77-сц, предложен студентом МИЭТ]

Несжатые символы объединяются в цепочки аналогично RLE с флаг-битом.

- флаг-бит θ ссылка/цепочка может быть 0/1 или 1/0; ниже рассматривается 1/0;
- θ помещается в байт L ; L или S на один бит короче (разрядность L цепочки может отличаться от L ссылки);
- порядок (L, S);
- для ссылки: аналогично RLE $\inf(L_{\min}) \in \{2, 3\}$ — возьмём $L_{\min} = 3$;
- для цепочки несжатых символов: $L_{\min}^{\text{несж}} = 1$ по определению;
- возможен код со смещением: $\tilde{S} = S - S_{\min} = S - 1$, $\tilde{L} = L - L_{\min} = L - 3$;
 $\widetilde{L^{\text{несж}}} = L^{\text{несж}} - 1 \implies L_{\max}^{\text{несж}} = 2^{k-1}$.

В наихудшем случае к файлу из n байтов добавляется $\frac{n}{L_{\max}^{\text{несж}}} = \frac{n}{2^{k-1}}$ дополнительных байтов (при $k > 3$ — меньше, чем для флаг-байтов ссылка/символ).

Примем $k = 3$:

$$\text{для ссылок } \left\{ \begin{matrix} S \\ L \end{matrix} \right\} \mid \theta : |L| : |S| = 1 : 2 : 3, \begin{cases} 0 \leq S - 1 \leq 2^3 - 1 \\ 0 \leq L - 3 \leq 2^2 - 1 \end{cases} \implies \begin{cases} 1 \leq S \leq 2^3 \\ 3 \leq L \leq 2^2 + 2 \end{cases} \implies \begin{cases} S_{\max} = 8 \\ L_{\max} = 6 \end{cases}$$

$$\text{для цепочек несжатых символов } |\theta| : |L| = 1 : 2, \quad L_{\max}^{\text{несж}} = 4.$$



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепт Зива—Лемпеля [LZ77-к]

LZ77 с флаг-байтами ссылка/символ [LZ77-сс]

LZ77 с флаг-битами ссылка/цепочка [LZ77-сц, предложен студентом МИЭТ]

LZ77 с односимвольным префиксом [LZ77-р]

Лучший и худший случаи LZ77 (для октетов: $k = 8$)

Сравнение с RLE

Модель источника для LZ77

LZ77 с флаг-битами ссылка/цепочка [LZ77-сц, предложен студентом МИЭТ] (продолжение): ссылка/цепочка 1/0

$m = 7777\ 7000\ 0000\ 0000\ 0123\ 4567\ 0123\ 4343\ 7012\ 1077$ (40 трёхбитных байтов)

ссылки ($\theta = 1$): $1 \leq S \leq 8$, $3 \leq L \leq 6$; цепочки несжатых символов ($\theta = 0$): $1 \leq L \leq 4$:

$$\left\{ \begin{matrix} 0 \\ L=1 \\ 7 \end{matrix} \right\} \left\{ \begin{matrix} 1 \\ S=1 \\ L=4 \\ 7777 \end{matrix} \right\} \left\{ \begin{matrix} 0 \\ L=1 \\ 0 \end{matrix} \right\} \left\{ \begin{matrix} 1 \\ S=1 \\ L=6 \\ 000000 \end{matrix} \right\} \left\{ \begin{matrix} 1 \\ S=1 \\ L=5 \\ 00000 \end{matrix} \right\} \left\{ \begin{matrix} 0 \\ L=4 \\ 1234 \end{matrix} \right\} \left\{ \begin{matrix} 0 \\ L=3 \\ 567 \end{matrix} \right\} \left\{ \begin{matrix} 1 \\ S=8 \\ L=5 \\ 01234 \end{matrix} \right\} \left\{ \begin{matrix} 1 \\ S=2 \\ L=3 \\ 343 \end{matrix} \right\} \left\{ \begin{matrix} 0 \\ L=4 \\ 7012 \end{matrix} \right\} \left\{ \begin{matrix} 0 \\ L=4 \\ 1077 \end{matrix} \right\}$$

смещение для ссылки: $\tilde{S} = S - 1$, $\tilde{L} = L - 3$; для цепочки $\tilde{L} = L - 1$:

$$\begin{aligned} &((0, 1 - 1 = 0), 7) ((1, 4 - 3 = 1), 1 - 1 = 0) ((0, 1 - 1 = 0), 0) ((1, 6 - 3 = 3), 1 - 1 = 0) ((1, 5 - 3 = 2), 1 - 1 = 0) \\ &((0, 4 - 1 = 3), 1234) ((0, 3 - 1 = 2), 567) ((1, 5 - 3 = 2), 8 - 1 = 7) ((1, 3 - 3 = 0), 2 - 1 = 1) \\ &((0, 4 - 1 = 3), 7012) ((0, 4 - 1 = 3), 1077) \end{aligned}$$

$(000, 7)(101, 0)(000, 0)(111, 0)(110, 0)(011, 1234)(010, 567)(110, 7)(100, 1)(011, 7012)(011, 1077)$

$code(m) = 075000706031234256767413701231077$ (33 трёхбитных байта)



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепт Зива—Лемпеля [LZ77-к]

LZ77 с флаг-байтами ссылка/символ [LZ77-сс]

LZ77 с флаг-битами ссылка/цепочка [LZ77-сц, предложен студентом МИЭТ]

LZ77 с односимвольным префиксом [LZ77-р]

Лучший и худший случаи LZ77 (для октетов: $k = 8$)

Сравнение с RLE

Модель источника для LZ77

LZ77 с односимвольным префиксом [LZ77- p]

- $\left\{ \begin{smallmatrix} S \\ L \end{smallmatrix} \right\}$ записывается как $\left(p, \left\{ \begin{smallmatrix} S \\ L \end{smallmatrix} \right\} \right)$; байт p самый редкий, $p(p) \leq \frac{1}{2^k}$;
значение p сохраняется в заголовке файла;
- $L_{\min} = 4$;
- байт p в несжатом тексте экранируется только как $(p, 0) \implies$
 \implies в $\left(p, \left\{ \begin{smallmatrix} S \\ L \end{smallmatrix} \right\} \right)$ второй байт $\neq 0 \implies$ порядок $\left(p, \tilde{L}, \tilde{S} \right)$ с $\tilde{L} \neq 0 \implies$
 \implies смещение не на $L_{\min} = 4$, а на $L_{\min} - 1 = 3$: $(p, L - 3, S - 1)$

В наихудшем случае к файлу из n байтов добавляется столько дополнительных байтов, сколько было символов p — не более $\frac{n}{2^k}$ (меньше, чем для флаг-битов).

LZ77 с односимвольным префиксом [LZ77- p]

$$\text{Примем } k = 3, |L| : |S| = 2 : 4 \implies \begin{cases} 0 \leq S - 1 \leq 2^4 - 1 \\ 1 \leq L - 3 \leq 2^2 - 1 \end{cases} \implies \begin{cases} 1 \leq S \leq 2^4 \\ 4 \leq L \leq 2^2 + 2 \end{cases} \implies \begin{cases} S_{\max} = 16 \\ L_{\max} = 6 \end{cases}$$

выбираем префикс $p = 5$

$m = 7777\ 7000\ 0000\ 0000\ 0123\ 4567\ 0123\ 4343\ 7012\ 1077$ (40 трёхбитных байтов)

$$7 \begin{Bmatrix} S=1 \\ L=4 \\ 7777 \end{Bmatrix} 0 \begin{Bmatrix} S=1 \\ L=6 \\ 000000 \end{Bmatrix} \begin{Bmatrix} S=1 \\ L=5 \\ 00000 \end{Bmatrix} 1234 \begin{Bmatrix} S=8 \\ L=5 \\ 01234 \end{Bmatrix} 343 \begin{Bmatrix} S=9 \\ L=4 \\ 7012 \end{Bmatrix} 1077$$

$$7 \begin{Bmatrix} 1-1=0 \\ 4-3=1 \end{Bmatrix} 0 \begin{Bmatrix} 1-1=0 \\ 6-3=3 \end{Bmatrix} \begin{Bmatrix} 1-1=0 \\ 5-3=2 \end{Bmatrix} 1234 \begin{Bmatrix} 5,0 \end{Bmatrix} 67 \begin{Bmatrix} 8-1=7 \\ 5-3=2 \end{Bmatrix} 343 \begin{Bmatrix} 9-1=8 \\ 4-3=1 \end{Bmatrix} 1077$$

$$7 \begin{Bmatrix} 0000 \\ 01 \end{Bmatrix} 0 \begin{Bmatrix} 0000 \\ 11 \end{Bmatrix} \begin{Bmatrix} 0000 \\ 10 \end{Bmatrix} 1234 \begin{Bmatrix} 5,0 \end{Bmatrix} 67 \begin{Bmatrix} 0111 \\ 10 \end{Bmatrix} 343 \begin{Bmatrix} 1000 \\ 01 \end{Bmatrix} 1077$$

$$7 \begin{Bmatrix} 000 \\ 001 \end{Bmatrix} 0 \begin{Bmatrix} 000 \\ 011 \end{Bmatrix} \begin{Bmatrix} 000 \\ 010 \end{Bmatrix} 1234 \begin{Bmatrix} 5,0 \end{Bmatrix} 67 \begin{Bmatrix} 111 \\ 010 \end{Bmatrix} 343 \begin{Bmatrix} 000 \\ 101 \end{Bmatrix} 1077$$

$code(m) = 75100530520123450675273435401077$ (33 трёхбитных байта)



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепт Зива—Лемпеля [LZ77-к]

LZ77 с флаг-байтами ссылка/символ [LZ77-сс]

LZ77 с флаг-байтами ссылка/цепочка [LZ77-сц, предложен студентом МИЭТ]

LZ77 с односимвольным префиксом [LZ77- p]

Лучший и худший случаи LZ77 (для октетов: $k = 8$)

Сравнение с RLE

Модель источника для LZ77

Лучший и худший случаи LZ77 (для октетов: $k = 8$)

Лучший исходный файл — $m_n = \underbrace{0000\dots 0000}_n$; худший — ни в одном в окне нет ни одного совпадения $L \geq L_{\min}$

Считаем $n \gg L_{\max}$; всё оцениваем приблизительно; в частности, для LZ77-сц $L_{\max}^{\text{несж}} \approx L_{\max}$.

Первые две колонки $|code(m_n)|$ и $|code(m_x)|$ — для любого L_{\max} (то есть любых $|L|$ и $|S|$), далее $|L| = |S| = k$.

$code$	$ code(m_n) $	$ code(m_x) $	$L_{\max}^{ L = S =k}$	$L_{\max}^{ L = S =k} _{k=8}$	$ code(m_n) $	$ code(m_x) $
LZ77-к	$3 \cdot \frac{n}{L_{\max}+1}$	$3 \cdot n$	$2^k - 1$	255	$3 \cdot \frac{n}{256}$	$3 \cdot n$
LZ77-сс	$(2 + \frac{1}{k}) \cdot \frac{n}{L_{\max}}$	$(1 + \frac{1}{k}) \cdot n$	$2^k + 1$	257	$(2 + \frac{1}{8}) \cdot \frac{n}{257}$	$(1 + \frac{1}{8}) \cdot n$
LZ77-сц	$2 \cdot \frac{n}{L_{\max}}$	$(1 + \frac{1}{L_{\max}}) \cdot n$	$\approx 2^{k-1}$	130 (128 несж)	$\frac{2n}{130} = 4 \cdot \frac{n}{260}$	$(1 + \frac{1}{128}) \cdot n$
LZ77- p	$3 \cdot \frac{n}{L_{\max}}$	$n + count(p) \leq n \cdot (1 + \frac{1}{2^k})$	$2^k + 2$	258	$3 \cdot \frac{n}{258}$	$(1 + \frac{1}{256}) \cdot n$

При $|L| = k - 1$ и $|S| = k + 1$: а) L_{\max} примерно вдвое меньше; б) худший случай для LZ77-к не реализуется (окно больше алфавита и $L_{\min} = 1$). При $|L| = k - 2$ и $|S| = k + 2$: L_{\max} вчетверо меньше, а окно больше...

LZ77-сс лучший на m_n ; используется в LZJB с $|L| = 6$ и $|S| = 10$ бит и изб. данными — худший случай редок.

LZ77- p лучший на m_x . Вопрос: всегда ли LZ77- p лучше или равноценен LZ77-к?

⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↺

RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепт Зива—Лемпеля [LZ77-к]

LZ77 с флаг-байтами ссылка/символ [LZ77-сс]

LZ77 с флаг-байтами ссылка/цепочка [LZ77-сц, предложен студентом МИЭТ]

LZ77 с односимвольным префиксом [LZ77- p]

Лучший и худший случаи LZ77 (для октетов: $k = 8$)

Сравнение с RLE

Модель источника для LZ77

Сравнение с RLE

Как RLE, так и LZ77 не требуют отдельного словаря.

Сравниваем реализации с префиксом для $k = 3$: $p = 5$, $RLE : \begin{cases} 4 \leq L^{c \neq p} \leq 11, \\ 1 \leq L^p \leq 8; \end{cases} \quad LZ77 : \begin{cases} 1 \leq S \leq 16, \\ 4 \leq L \leq 6. \end{cases}$

- ❶ Одиночная цепочка повторений длины $L \leq L_{\max}^{LZ77}$:

RLE: $L \rightarrow 3 \quad 77777 \rightarrow 527$

LZ77: $L \rightarrow 4 \quad 77777 \rightarrow 7510$

- ❷ Есть цепочки повторений длины порядка L_{\max}^{RLE} :

RLE: 777770000000000000012345670123434370121077 $\rightarrow 52757000123450670123434370121077 \quad 40 \rightarrow 33$

LZ77: 777770000000000000012345670123434370121077 $\rightarrow 75100530520123450675273435401077 \quad 40 \rightarrow 33$

- ❸ Длина повторений не более L_{\max}^{LZ77} :

RLE: 77777000000000012345670123434370121077 $\rightarrow 52753000123450670123434370121077 \quad 36 \rightarrow 32$

LZ77: 77777000000000012345670123434370121077 $\rightarrow 751005300123450675273435401077 \quad 36 \rightarrow 31$

Для большинства файлов LZ77 эффективнее RLE аналогичной реализации; существуют файлы, для которых лучшим будет RLE; для некоторых (модель наивного RLE) лучшим будет наивный RLE.



RLE: подсемейства «наивный RLE» [RLE-н] и «RLE с флаг-битом» [RLE-фб]

RLE: подсемейство «RLE с односимвольным префиксом»

RLE: лучший и худший случаи + модель источника

Семейство LZ77 (идея предложена Зивом с участием Лемпеля в 1977 году)

LZ77: основные варианты отделения ссылок от несжатого текста

Семинар RLE/LZ77

Концепт Зива—Лемпеля [LZ77-к]

LZ77 с флаг-байтами ссылка/символ [LZ77-сс]

LZ77 с флаг-байтами ссылка/цепочка [LZ77-сц, предложен студентом МИЭТ]

LZ77 с односимвольным префиксом [LZ77-р]

Лучший и худший случаи LZ77 (для октетов: $k = 8$)

Сравнение с RLE

Модель источника для LZ77

Модель источника для LZ77

LZ77 строится из предположения, что модель марковская.

Задача: каковы характеристики модели для оценки снизу длины кода алгоритмов семейства LZ77:

- стационарна ли она?
- какова глубина памяти такой модели?

Задача №1

Закодируйте различными реализациями RLE/LZ77 сообщение

$m = 7770\ 0000\ 0123\ 4567\ 7770\ 7770\ 0000\ 0000\ 0011\ 2233\ 4455\ 6677$

исходная длина m в символах $n = 4 \cdot 12 = 48$ символов

- 1 в байте $k = 3$ бита;
- 2 символ кодирования — 3-битный байт ($0 - 7$); сгруппированы по 4 и 16 штук для удобства.

Задача №2

- 1 Рассчитать количество информации в сообщении
 $C = \underbrace{\text{«ля-ля-ля-...-ля»}}_{80 \text{ раз}}$ (кодировка koï8-r)
- 2 Закодировать сообщение C алгоритмом LZ77

Спасибо за внимание!

МИЭТ

www.miet.ru

Александра Игоревна Кононова

illinc@mail.ru

gitlab.com/illinc/raspisanie

Информация и модели источника

Заменяя вероятности символов на их оценки по модели X , получаем **оценку** количества информации:

- ❶ Без памяти, $A_1 = \{\text{л}, \text{я}, -\}$:

$$I_1 = 2 \cdot 80 \cdot \left(-\log_2 \left(\frac{80}{239} \right) \right) + 79 \cdot \left(-\log_2 \left(\frac{79}{239} \right) \right) = 254,2 \text{ бит} = 31,8 \text{ байт}$$

- ❷ Без памяти, $A_1 = \{\text{ля-}, \text{ля}\}$:

$$I_2 = 79 \cdot \left(-\log_2 \left(\frac{79}{80} \right) \right) + 1 \cdot \left(-\log_2 \left(\frac{1}{80} \right) \right) = 7,6 \text{ бит} = 0,9 \text{ байт}$$

- ❸ Маркова, $A_1 = \text{коі8-г}, N = 3$ со следующими оценками вероятностей:

$$\begin{aligned} p(c_1 = \text{л}) = p(c_2 = \text{я}) = p(c_3 = -) &= \frac{1}{256}, & p(-|\text{-ля}) &= \frac{79}{80}, \\ p(\text{л}|\text{ля-}) = p(\text{я}|\text{я-л}) &= 1, & p(\text{eof}|\text{-ля}) &= \frac{1}{80}: \end{aligned}$$

$$I_3 = 3 \cdot \log_2 256 + 79 \cdot \left(-\log_2 \left(\frac{79}{80} \right) \right) + 1 \cdot \left(-\log_2 \left(\frac{1}{80} \right) \right) = 31,6 \text{ бит} = 3,9 \text{ байт}$$