Словарные методы сжатия

Лекция 3

План

- Общие сведения
- LZ77
- LZSS
- LZ78
- LZW

Общие сведения

- Последовательности символов сохраняются в словаре и кодируются в виде меток
- В ходе кодирования ищется слово в словаре и в выходной файл записывается его метка
- Если встречается новое слово, которого нет в словаре, то оно записывается в выходной файл без сжатия
- Для отличия слов от меток вводится дополнительный бит, который указывает, что за ним идет – слово или метка
- Статический словарь составляется заранее и имеет определенное число слов
- Динамический словарь начинается с минимального количества слов и модифицируется по мере поступления информации из входного потока

LZ77 (скользящее окно)

Буфер поиска
(словарь)
(та, часть , которая уже закодирована)

Упреждающий буфер
(текст, который нужно закодировать)

Кодер ищет совпадение символа из упреждающего буфера в буфере поиска. Найдя, записывает смещение от правого края буфера поиска, длину совпадения и первый из не совпавших символов

Декодер строит такой же буфер поиска, как и кодера и по нему находит совпадения

Процесс кодирования

Буфер поиска	Упреждающий буфер	Выходная строка
	sir_sid_eastman_r	0,0,s
s	ir_sid_eastman_r	0,0,i
si	r_sid_eastman_r	0,0,r
sir	_sid_eastman_r	0,0,_
sir_	sid_eastman_r	4,2,d
sir_sid	_eastman_r	4,1,e
sir_sid_e	astman_r	0,0,a

LZSS

- Упреждающий буфер сохраняется в виде циклической очереди
- Словарь (буфер поиска) записывается в виде двоичного дерева
- Метки имеют 2 поля, а не три. Если не найдено совпадений, то кодер просто подает на выход несжатый код следующего символа. Для различения меток и несжатых кодов используется флаговый бит.

Пример. Построим дерево с окном 5

Метка кодера 16,2

Строки буфера поиска	Смещение	Строки буфера поиска	Смещение
sid_e	16	stman	10
id_ea	15	tman_	09
d_eas	14	man_c	08
_east	13	an_cl	07
eastm	12	n_clu	06

Пример. Перестроим дерево

si

Строки буфера поиска	Смещение	Строки буфера поиска	Смещение
d_eas	16	man_c	10
_east	15	an_cl	09
eastm	14	n_clu	08
astma	13	_clum	07
stman	12	clums	06

LZ78

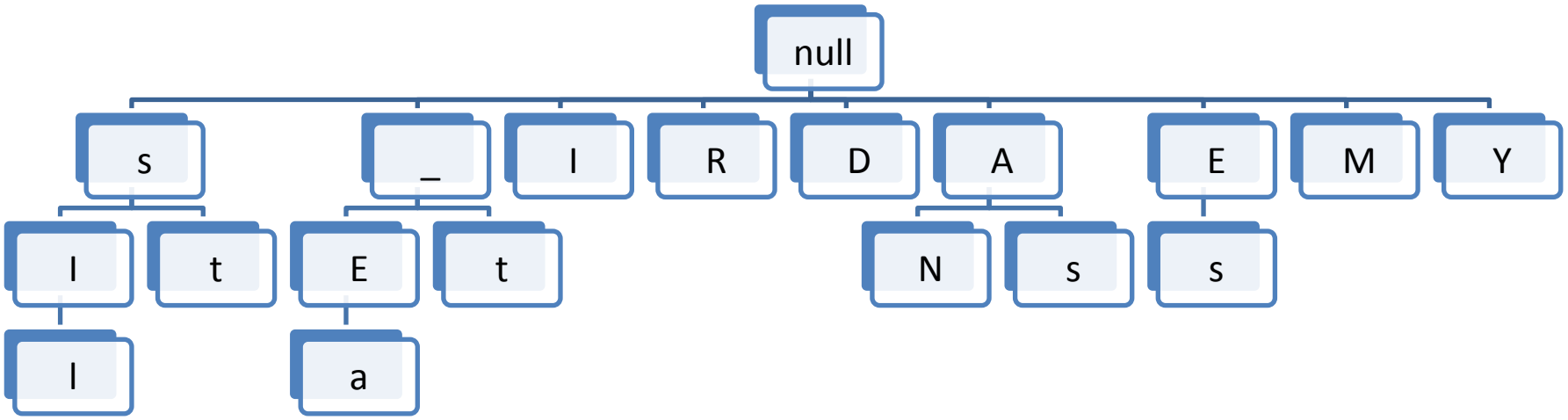
- Использует словарь встретившихся ранее слов
- На первом шаге он почти пуст
- По мере поступления новые строки получают метки 1,2,3…
- По мере чтения входного файла ищется позиция символа во словаре, если он там есть, то читается следующий символ и ищется вхождение 2 символов в словарь и так далее пока не поступит символ строки, которого нет в словаре.
- Как только нашелся новый символ, кодер добавляет его в словарь и строит метку.
- Метка содержит 2 поля. 1- указатель на найденную строку в словаре, 2- символ, на котором произошел обрыв

Пример. Кодирование

sir_sid_eastman_easily_teases

Словарь	Метка	Словарь	Метка
0 null		9 ‘st’	(1,’t’)
1 ‘s’	(0,’s’)	10 ‘m’	(0,’m’)
2 ‘i’	(0,’i’)	11 ‘an’	(8,’n’)
3 ‘r’	(0,’r’)	12 ‘_ea’	(7,’a’)
4 ‘_’	(0,’_’)	13 ‘sil’	(5,’l’)
5 ‘si’	(1,’i’)	14 ‘y’	(0,’y’)
6 ‘d’	(0,’d’)	15 ‘_t’	(4,’t’)
7 ‘_e’	(4,’e’)	16 ‘e’	(0,’e’)
8 ‘a’	(0,’a’)	17 ‘as’	(8,’s’)
		18 ‘es’	(16,’s’)

Словарное дерево



LZW

- Инициализация словаря всеми символами исходного алфавита
- Каждый поступающий символ записывается во входную строку I и ищется в словаре, если очередной символ не найден, то в выходной файл записывается указатель на найденную часть строки
- В словарь записывается строка + новый символ
- Строка I инициализируется новым символом

Пример. Кодирование sir_sid_eastman

Символ	Код	Символ	Код
S	115	t	116
i	105	m	109
R	114	e	101
_	32	a	97
d	100		

Пример. Кодирование sir_sid_eastman

Входная строка I	Есть в словаре?	Новая запись	Выход	Входная строка I	Есть в словаре?	Новая запись	Выход
S si	Да нет	256-si	115 (s)	E ea	Да нет	263-ea	101 (e)
i ir	Да нет	257-ir	105 (i)	A as	Да нет	264 -as	97 (a)
R R_	Да нет	258-r_	114 (r)	S st	Да нет	265-st	115 (s)
_ _s	Да нет	259-_s	32 (_)	T tm	Да нет	266-tm	116 (t)
S Si sid	Да Да нет	260-sid	256 (si)	M ma	Да нет	267-ma	109 (m)
D D_	Да нет	261-d_	100 (d)	A an	Да нет	268-an	97 (a)
_ _e	Да нет	262-_e	32 (_)				

Декодирование

- Заполнение словаря первыми символами алфавита (256)
- По указателям из входного файла восстанавливаем несжатые символы и записываем их в выходной файл

Пример.

- Входной словарь из прошлого примера
- Входной код – 115 105 114 32 256 100 32

Входной код	Выход	Входной код	Выход
115	s	256	si
105	i	100	d
114	r	32	_
32	_		