Символьные модели сжатия

Дискретный анализ 2012/13

Андрей Калинин, Татьяна Романова

11 мая 2013 г.

Литература

- Witten, Moffat, Bell, Managing gigabytes: compressing and indexing documents and images.
- Ватолин Д., Ратушняк А., Смирнов М., Юкин В., Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео.

Раздел

Предсказание по частичному совпадению (PPM)

Блочное сжатие

Преобразование Барроуза-Уилера (BWT)

Использование BWT

"I never heerd a skilful old married feller of twenty years' standing pipe "my wife" in a more used note than 'a did,"said Jacob Smallbury. "It might have been a little more true to nater if't had been spoke a little chillie

- Следующий символ r.
- Проверяется 5-и символьный контекст $illie$ — никогда не встречался в тексте, модель переходит к 4-х символьному контексту.
- Контекст $llie$ встречался один раз, за ним следовал s. Посылается ESC и модель переходит к 3-х символьному контексту (допустим, ESC это код с частотой 1).
- lie встречался 201 раз и 19 раз за ним следовал символ r, поэтому r может быть упакован с вероятностью 19/202, т.е. 3.4 битами (1 — для ESC).
- При достижении 0-го порядка символ кодируется равновероятно.
- PPMn — начинается с модели порядка n, PPM* — модель бесконечного порядка.

Исключения

- При расчёте вероятности следования r за lie не учитывался тот факт, что 22 раза из 201 за lie следовала буква s, обрабатываемая моделью старшего порядка.
- Можно вычесть 22 из 201 и улучшить оценку для r: 19/180 против 19/202.

Оценка ESC

n — количество символов в текущем контексте, r — количество разных символов, c_i — количество i-х символов, t_1 — количество символов с $c_i = 1$.

- PPMA: $P(esc) = 1/(n+1)$; $P(i) = c_i/(n+1)$.
- PPMC: $P(esc) = r/(n+r)$; $P(i) = c_i/(n+r)$.
- PPMD: $P(esc) = r/(2n)$; $P(i) = (2c_i - 1)/(2n)$.
- PPMX: $P(esc) = t_1/(n+t_1)$; $P(i) = c_i/(n+t_1)$. Если $t_1 = 0$, то $P(esc) = (t_1+1)/(n+t_1+1)$ и $P(i) = c_i/(n+t_1+1)$.

Раздел

Предсказание по частичному совпадению (PPM)

Блочное сжатие

- Преобразование Барроуза-Уилера (BWT)
- Использование BWT

Раздел

Идея преобразования

- Построить для строки все её циклические перестановки.
- Остортировать в лексикографическом порядке реверсов перестановок.
- Каждый столбец полученной в результате матрицы символов — перестановка символов исходной строки.
- Первый столбец и есть BWT.
- Можно получить из суффиксного массива A, построенного для реверса S' исходной строки: $S'[A[i] - 1]$.
- Символы с похожими контекстами в BWT окажутся рядом.

Пример прямого преобразования

mississippi$		m	ississippi$
ississippi$m		s	sissippi$mi
ssissippi$mi		$	mississippi
sissippi$mis		s	sippi$missi
issippi$miss		p	pi$mississi
ssippi$missi	$\Longrightarrow$	i	ssissippi$m
sippi$missis		p	i$mississip
ippi$mississ		i	$mississipp
ppi$mississi		s	issippi$mis
pi$mississip		s	ippi$missis
i$mississipp		i	ssippi$miss
$mississippi		i	ppi$mississ

Восстановление строки

1	m
2	s
3	$
4	s
5	p
6	i
7	p
8	i
9	s
10	s
11	i
12	i

Восстановление строки

1	m	$
2	s	i
3	$	i
4	s	i
5	p	i
6	i	m
7	p	p
8	i	p
9	s	s
10	s	s
11	i	s
12	i	s

Восстановление строки

1	m	$	$m
2	s	i	is
3	$	i	i$
4	s	i	is
5	p	i	ip
6	i	m	mi
7	p	p	pp
8	i	p	pi
9	s	s	ss
10	s	s	ss
11	i	s	si
12	i	s	si

Восстановление строки

1	m	$	$m	i$
2	s	i	is	mi
3	$	i	i$	pi
4	s	i	is	si
5	p	i	ip	si
6	i	m	mi	$m
7	p	p	pp	ip
8	i	p	pi	pp
9	s	s	ss	is
10	s	s	ss	is
11	i	s	si	ss
12	i	s	si	ss

Восстановление строки

1	m	$	$m	i$	i$m
2	s	i	is	mi	mis
3	$	i	i$	pi	pi$
4	s	i	is	si	sis
5	p	i	ip	si	sip
6	i	m	mi	$m	$mi
7	p	p	pp	ip	ipp
8	i	p	pi	pp	ppi
9	s	s	ss	is	iss
10	s	s	ss	is	iss
11	i	s	si	ss	ssi
12	i	s	si	ss	ssi

Восстановление строки

1	m	$	$m	i$	i$m	pi$
2	s	i	is	mi	mis	$mi
3	$	i	i$	pi	pi$	ppi
4	s	i	is	si	sis	ssi
5	p	i	ip	si	sip	ssi
6	i	m	mi	$m	$mi	i$m
7	p	p	pp	ip	ipp	sip
8	i	p	pi	pp	ppi	ipp
9	s	s	ss	is	iss	mis
10	s	s	ss	is	iss	sis
11	i	s	si	ss	ssi	iss
12	i	s	si	ss	ssi	iss

Восстановление строки

1	m	$	$m	i$	i$m	pi$	pi$m
2	s	i	is	mi	mis	$mi	$mis
3	$	i	i$	pi	pi$	ppi	ppi$
4	s	i	is	si	sis	ssi	ssis
5	p	i	ip	si	sip	ssi	ssip
6	i	m	mi	$m	$mi	i$m	i$mi
7	p	p	pp	ip	ipp	sip	sipp
8	i	p	pi	pp	ppi	ipp	ippi
9	s	s	ss	is	iss	mis	miss
10	s	s	ss	is	iss	sis	siss
11	i	s	si	ss	ssi	iss	issi
12	i	s	si	ss	ssi	iss	issi

Восстановление строки

1	m	$	$m	i$	i$m	pi$	pi$m	ppi$
2	s	i	is	mi	mis	$mi	$mis	i$mi
3	$	i	i$	pi	pi$	ppi	ppi$	ippi
4	s	i	is	si	sis	ssi	ssis	issi
5	p	i	ip	si	sip	ssi	ssip	issi
6	i	m	mi	$m	$mi	i$m	i$mi	pi$m
7	p	p	pp	ip	ipp	sip	sipp	ssip
8	i	p	pi	pp	ppi	ipp	ippi	sipp
9	s	s	ss	is	iss	mis	miss	$mis
10	s	s	ss	is	iss	sis	siss	ssis
11	i	s	si	ss	ssi	iss	issi	miss
12	i	s	si	ss	ssi	iss	issi	siss

Восстановление строки

1	m	$	$m	i$	i$m	pi$	pi$m	ppi$	ppi$m
2	s	i	is	mi	mis	$mi	$mis	i$mi	i$mis
3	$	i	i$	pi	pi$	ppi	ppi$	ippi	ippi$
4	s	i	is	si	sis	ssi	ssis	issi	issis
5	p	i	ip	si	sip	ssi	ssip	issi	issip
6	i	m	mi	$m	$mi	i$m	i$mi	pi$m	pi$mi
7	p	p	pp	ip	ipp	sip	sipp	ssip	ssipp
8	i	p	pi	pp	ppi	ipp	ippi	sipp	sippi
9	s	s	ss	is	iss	mis	miss	$mis	$miss
10	s	s	ss	is	iss	sis	siss	ssis	ssiss
11	i	s	si	ss	ssi	iss	issi	miss	missi
12	i	s	si	ss	ssi	iss	issi	siss	sissi

Восстановление строки

1	m	$	$m	i$	i$m	pi$	pi$m	ppi$	ppi$m	ippi$
2	s	i	is	mi	mis	$mi	$mis	i$mi	i$mis	pi$mi
3	$	i	i$	pi	pi$	ppi	ppi$	ippi	ippi$	sippi
4	s	i	is	si	sis	ssi	ssis	issi	issis	missi
5	p	i	ip	si	sip	ssi	ssip	issi	issip	sissi
6	i	m	mi	$m	$mi	i$m	i$mi	pi$m	pi$mi	ppi$m
7	p	p	pp	ip	ipp	sip	sipp	ssip	ssipp	issip
8	i	p	pi	pp	ppi	ipp	ippi	sipp	sippi	ssipp
9	s	s	ss	is	iss	mis	miss	$mis	$miss	i$mis
10	s	s	ss	is	iss	sis	siss	ssis	ssiss	issis
11	i	s	si	ss	ssi	iss	issi	miss	missi	$miss
12	i	s	si	ss	ssi	iss	issi	siss	sissi	ssiss

Восстановление строки в один проход

1	$m	
2	is	
3	i$	
4	is	
5	ip	
6	mi	m
7	pp	
8	pi	
9	ss	
10	ss	
11	si	
12	si	

Восстановление строки в один проход

1	$m
2	is
3	i$
4	is
5	ip
6	mi
7	pp
8	pi
9	ss
10	ss
11	si
12	si

mi

Восстановление строки в один проход

1		$m
2	m	is
3	p	i$
4	s	is
5	s	ip
6		mi
7		pp
8		pi
9		ss
10		ss
11		si
12		si

mis

Восстановление строки в один проход

1		$m
2		is
3		i$
4		is
5		ip
6		mi
7		pp
8		pi
9	i	ss
10	i	ss
11	s	si
12	s	si

miss

Восстановление строки в один проход

1		$m
2		is
3		i$
4		is
5		ip
6		mi
7		pp
8		pi
9	i	ss
10	i	ss
11	s	si
12	s	si

missi

Восстановление строки в один проход

1		$m
2	m	is
3	p	i$
4	s	is
5	s	ip
6		mi
7		pp
8		pi
9		ss
10		ss
11		si
12		si

missis

Восстановление строки в один проход

1		$m
2		is
3		i$
4		is
5		ip
6		mi
7		pp
8		pi
9	i	ss
10	i	ss
11	s	si
12	s	si

mississ

Восстановление строки в один проход

1		$m
2		is
3		i$
4		is
5		ip
6		mi
7		pp
8		pi
9	i	ss
10	i	ss
11	s	si
12	s	si

mississi

Восстановление строки в один проход

1		$m
2	m	is
3	p	i$
4	s	is
5	s	ip
6		mi
7		pp
8		pi
9		ss
10		ss
11		si
12		si

mississip

Восстановление строки в один проход

1		$m
2		is
3		i$
4		is
5		ip
6		mi
7	i	pp
8	p	pi
9		ss
10		ss
11		si
12		si

mississipp

Восстановление строки в один проход

1		$m
2		is
3		i$
4		is
5		ip
6		mi
7	i	pp
8	p	pi
9		ss
10		ss
11		si
12		si

mississippi

Восстановление строки в один проход

1		$m
2	m	is
3	p	i$
4	s	is
5	s	ip
6		mi
7		pp
8		pi
9		ss
10		ss
11		si
12		si

mississippi$

Алгоритм восстановления строки

1 $P[1..N]$ содержит преобразованную строку.
2 $p \leftarrow$ номер первого символа в P.
3 $K[s] \leftarrow$ количество вхождений символа s в P.
4 $M[\min\{\Sigma\}] \leftarrow 1$.
5 **for** $s \in \Sigma - \min\{\Sigma\}$
6 $\quad M[s] \leftarrow M[s-1] + K[s-1]$
7 **for** $i \leftarrow 1$ **to** N
8 $\quad s \leftarrow P[i]$, $L[i] \leftarrow M[s]$, $M[s] \leftarrow M[s] + 1$
// Теперь в массиве L содержатся связи, позволяющие обойти P
9 $i \leftarrow p$
10 **for** $k \leftarrow 1$ **to** N
11 $\quad$ Вывести $P[i]$
12 $\quad i \leftarrow L[i]$

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L												

	K	M
$	1	1
i	4	2
m	1	6
p	2	7
s	4	9

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6											

	K	M
$	1	1
i	4	2
m	1	7
p	2	7
s	4	9

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9										

	K	M
$	1	1
i	4	2
m	1	7
p	2	7
s	4	10

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1									

	K	M
$	1	2
i	4	2
m	1	7
p	2	7
s	4	10

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10								

	K	M
$	1	2
i	4	2
m	1	7
p	2	7
s	4	11

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7							

	K	M
$	1	2
i	4	2
m	1	7
p	2	8
s	4	11

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2						

	K	M
$	1	2
i	4	3
m	1	7
p	2	8
s	4	11

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8					

	K	M
$	1	2
i	4	3
m	1	7
p	2	9
s	4	11

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3				

	K	M
$	1	2
i	4	4
m	1	7
p	2	9
s	4	11

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11			

	K	M
$	1	2
i	4	4
m	1	7
p	2	9
s	4	12

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12		

	K	M
$	1	2
i	4	4
m	1	7
p	2	9
s	4	13

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	

	K	M
$	1	2
i	4	5
m	1	7
p	2	9
s	4	13

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

m

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

mi

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

mis

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

miss

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

missi

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

missis

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

mississ

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

mississi

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

mississip

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

mississipp

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

mississippi

Работа алгоритма

	1	2	3	4	5	6	7	8	9	10	11	12
P	m	s	$	s	p	i	p	i	s	s	i	i
L	6	9	1	10	7	2	8	3	11	12	4	5

	K	M
$	1	2
i	4	6
m	1	7
p	2	9
s	4	13

mississippi$

Раздел

```
 nly thrown into greater relie  f
n. Nevertheless, he↵was relie  v
 eba, feeling a nameless relie  f
 rise, experienced great relie  f
 thsheba was momentarily relie  v
P 398>↵foreheads, quite↵relie  v
t such times is a great↵relie  f
e droning of↵blue-bottle flie  s
 and the reasonable probabilie  s
  her head and feet,↵the lilie  s
eads, all↵about their familie  s
d been spoke a little↵chillie  r
 no absurd sides to the follie  s
lways be your↵friend,' replie  d
 s I've got no chance,' replie  d
↵'O no -- not at all,' replie  d
  'tis my only doctor,' replie  d
```

Идея блочного сжатия

- Текст разбивается на блоки.
- Для каждого блока считается BWT.
- Для текста одинаковые символы будут объединяться в кластеры, например, в предыдущем примере часть блока будет содержать в себе $fvffvvfssssssrsdddd$.
- BWT можно закодировать, например, при помощи move-to-front кодировщика (MTF) или комбинацией RLE с кодами Хаффмана.
- Качество сжатия близко к PPM*.

Move-To-Front

- Ведётся список последних закодированных символов (Last Recently Used List).
- Следующий символ из потока кодируется с вероятностью, соответствующей его позиции в LRU. После кодирования этот символ поднимается наверх.
- Вероятности могут назначаются статически (коды Хаффмана), могут оцениваться по ходу работы алгоритма (арифметические коды).