

COMP9319 Web Data Compression and Search

BWT, MTE and Pattern Matching

1

1

BWT

- Burrows–Wheeler transform (BWT) is an algorithm used to prepare data for use with data compression techniques such as bzip2.
- It was invented by Michael Burrows and David Wheeler in 1994 at DEC SRC, Palo Alto, California.
- It is based on a previously unpublished transformation discovered by Wheeler in 1983.

2

2

Move to Front (MTF)

Reduce entropy based on local frequency correlation

Usually used for BWT encoding step

Author and detail:

Original paper at webcms3

<http://www.arturocampes.com/ac/mtf.html>

3

3

Example: abaaabccad

Symbol	Code	List
		abcde.....
		dacbe.....
		abdc.....
		abdc.....
	
	
	
d	3	dacbe.....

To transform a general file, the list has 256 ASCII symbols.

4

4

Example: abaaabbbccdddddcc

Symbols: abaaabbbccdddddcc

Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ...,

01100100, 01100011, 01100011

Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

5

5

Example: abaaabbbccdddddcc

Symbols: abaaabbbccdddddcc

Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ...,

01100100, 01100011, 01100011

Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Recall that Shannon's entropy reaches the max when there is max uncertainty, i.e., equal probability, like the example above (4 "97"s, 4 "98"s, 4 "99"s, 4 "100"s).

e.g., Entropy $H = 2.00$

6

6

Example: abaaabbbccdddddcc

Symbols: abaaabbbccdddddcc

Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ..., 01100100, 01100011, 01100011

Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value: a b c d e f

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97

7

Example: abaaabbbccdddddcc

Symbols: abaaabbbccdddddcc

Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ..., 01100100, 01100011, 01100011

Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value: a b c d e f

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value: a b c d e f

move to front

8

Example: abaaabbbccdddddcc

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value: a b c d e f

Codes (in ASCII): 97, 98, 97, 97, 99, 99

Codes (in MTF): 97, 98

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value: b a c d e f

9

Example: abaaabbbccdddddcc

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value:

7, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100,

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value:

10

Example: abaaabbbccdddddcc

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value: a b c d e f

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97, 98, 1, 0,

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value: a b c d e f

11

Example: abaaabbbccdddddcc

List

Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255

Value: a b c d e f

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0, 1, 0

12

7

8

9

10

11

12

13

15

17

14

16

18

15

17

15

16

18

16

17

15

17

15

18

16

18