Chương 2. Mã Huffman

ntnhut@hcmus.edu.vn

Mã tối ưu

- Trong một đoạn văn bản, các ký tự có tần suất xuất hiện khác nhau.

→ *dùng mã tức thời để mã hoá ký tự có tần suất cao nhất thành từ mã có độ dài ngắn nhất.*

Bài toán: cho trước các tần suất xuất hiện của các ký tự, tìm mã tối ưu nhất.

Nguồn thông tin

Định nghĩa: ***Nguồn thông tin*** bao gồm bảng ký tự $\{a_1, a_2, \ldots, a_n\}$ cùng với phân phối xác suất của chúng $P(a_1), P(a_2), \ldots, P(a_n)$ thoả:

- $P(a_1) + P(a_2) + \ldots + P(a_n) = 1$.
- $0 \leq P(a_i) \leq 1$.

Ví dụ:

Symbol	*A*	*E*	*B*	*C*	*D*	*F*
Probability	0.4	0.2	0.1	0.1	0.1	0.1

Độ dài mã tối ưu

- ***Độ dài mã trung bình (average length)***

$$L = \sum_{i=1}^{n} d_i P(a_i).$$

- *Mã tối ưu nhất* là mã có độ dài mã trung bình nhỏ nhất (theo nghĩa chuỗi mã sẽ được nén ngắn nhất có thể được)

VD mã tối ưu

A	•—	N	—•
B	—•••	O	———
C	—•—•	P	•——•
D	—••	Q	——•—
E	•	R	•—•
F	••—•	S	•••
G	——•	T	—
H	••••	U	••—
I	••	V	•••—
J	•———	W	•——
K	—•—	X	—••—
L	•—••	Y	—•——
M	——	Z	——••

Figure 2: Morse code

A	$1(01)_8$	N	$1(16)_8$
B	$1(02)_8$	O	$1(17)_8$
C	$1(03)_8$	P	$1(20)_8$
D	$1(04)_8$	Q	$1(21)_8$
E	$1(05)_8$	R	$1(22)_8$
F	$1(06)_8$	S	$1(23)_8$
G	$1(07)_8$	T	$1(24)_8$
H	$1(10)_8$	U	$1(25)_8$
I	$1(11)_8$	V	$1(26)_8$
J	$1(12)_8$	W	$1(27)_8$
K	$1(13)_8$	X	$1(30)_8$
L	$1(14)_8$	Y	$1(31)_8$
M	$1(15)_8$	Z	$1(32)_8$

Mã ASCII

Mã Morse tối ưu hơn!

Mã Huffman

Định nghĩa: Cho trước nguồn thông tin S, ***mã Huffman*** là mã tức thời có độ dài mã trung bình nhỏ nhất $L_{min}(S)$.

Ví dụ: một mã Huffman cho nguồn thông tin sau

Symbol	*A*	*E*	*B*	*C*	*D*	*F*
Probability	0.4	0.2	0.1	0.1	0.1	0.1

là

A	*E*	*B*	*C*	*D*	*F*
0	10	1100	1101	1110	1111

có

$$L = 0.4 + 2\,(0.2) + 4.4\,(0.1) = 2.4.$$

Xây dựng mã Huffman nhị phân

- **2 ký tự nguồn $\{a_1, a_2\}$:**
  - Từ mã tương ứng là 0 và 1.
  - Độ dài các từ mã = 1.
- **3 ký tự nguồn $\{a_1, a_2, a_3\}$ trong đó $P(a_1)$ cao nhất:**
  - Rút về trường hợp 2 ký tự a_1 và $a_{2,3}$ với $P(a_{2,3}) = P(a_2) + P(a_3)$.

a_1	$a_{2,3}$
0	1

  - Tách từ mã '1' thành hai từ mã '10' và '11'

a_1	a_2	a_3
0	10	11

Tổng quát

- S là nguồn thông tin với bảng ký tự $\{a_1, a_2, \ldots, a_n\}$ và các phân phối xác suất $P(a_1) \geq P(a_2) \geq \ldots \geq P(a_n)$.
- Nguồn thông tin S* gồm n – 1 ký tự $\{a_1, a_2, \ldots, a_{n-2}$ và ký tự $a_{n-1,n}\}$ với các xác suất tương ứng là $P(a_1)$, $P(a_2)$, …, $P(a_{n-2})$ và $P(a_{n-1,n})=P(a_{n-1}) + P(a_n)$.

Định lý: *Giả sử K* là mã Huffman cho S*. Khi đó mã cho S có dạng*

a_1	a_2	…	a_{n-2}	a_{n-1}	a_n
$K^*(a_1)$	$K^*(a_2)$	…	$K^*(a_{n-2})$	$K^*(a_{n-1,n})0$	$K^*(a_{n-1,n})1$

Lưu ý: *sắp xếp ký tự $a_{n-1,n}$ tương ứng thứ tự của $P(a_{n-1,n})$ trong dãy xác suất được sắp xếp.*

Ví dụ

- Tìm mã Huffman cho nguồn thông tin sau

A	...	**0.4**
E	...	**0.2**
B	...	**0.1**
C	...	**0.1**
D	...	**0.1**
F	...	**0.1**

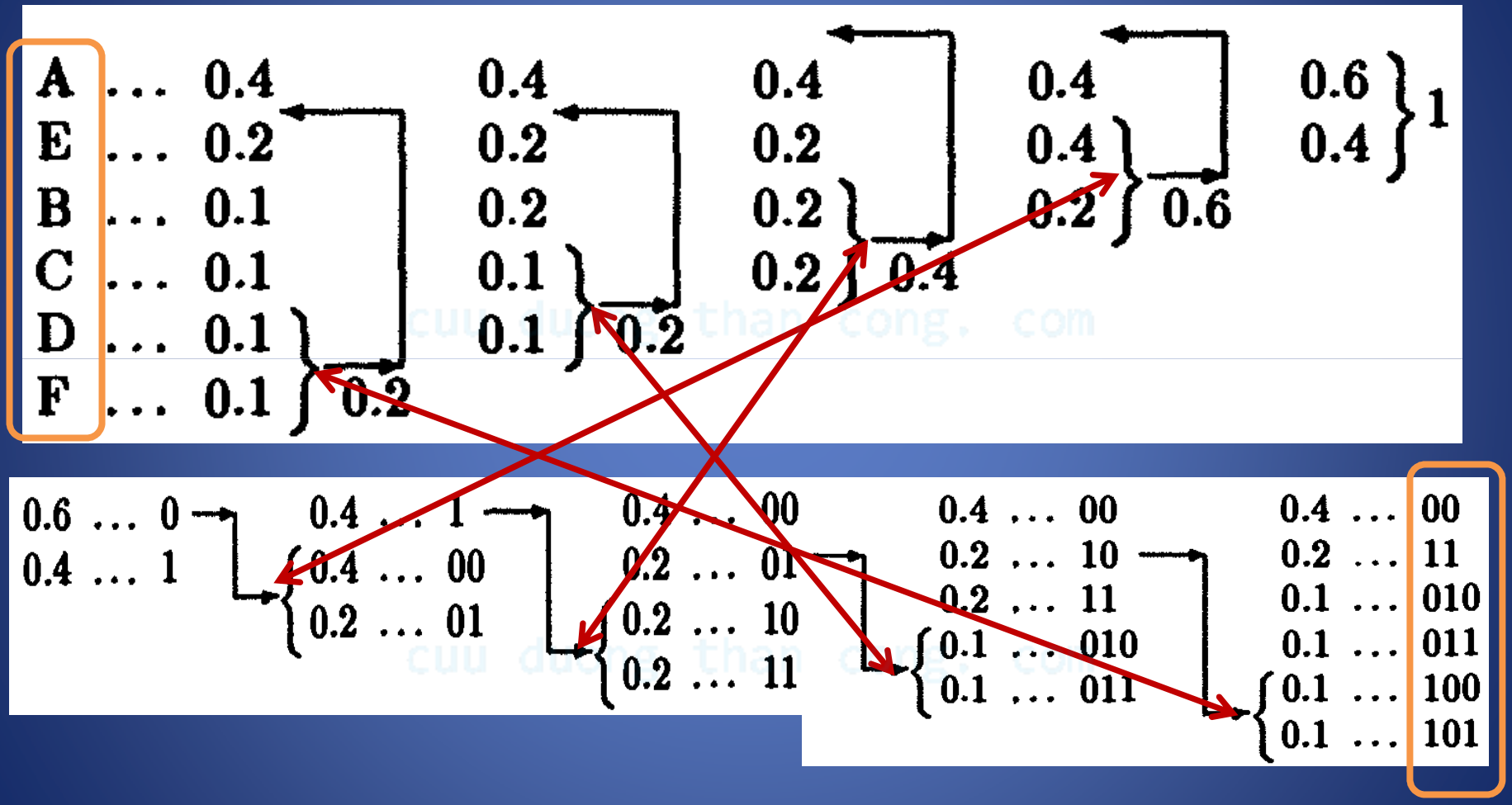
- Kết quả:

A	*E*	*B*	*C*	*D*	*F*
00	11	010	011	100	101

- Độ dài mã trung bình:

$$L_{\min}(S) = 2(0.4) + 2(0.2) + 3(0.4) = 2.4.$$

Các bước thực hiện



Mã Huffman mở rộng

- Bảng ký tự mã gồm k>2 ký tự (k>2).
- Ví dụ:

A	...	0.4
E	...	0.2
B	...	0.1
C	...	0.1
D	...	0.1
F	...	0.1

1
00
01
02
20
21

Tóm tắt

- Mã tối ưu
- Nguồn thông tin
- Độ dài mã tối ưu
- Mã Huffman

Đề tài nhóm

- Mã tự sửa [1] :
  1. Reed-Muller code
  2. Cyclic code
  3. BCH code
- Nén dữ liệu [2] :
  1. Arithmetic code
  2. Lempel-Ziv code

1. Jiri Adamek, ***Foundations of Coding***
2. David J. C. Mackay, ***Information Theory, Inference, and Learning Algorithms***.

Homework

- Đọc lại chương 2 [1] và làm các bài tập cuối chương.
- Đọc trước chương 3 [1]

Bài tập 1

- Tìm mã Huffman cho 3 trường hợp sau

Symbol	A	B	C	D	E	F	G	H
Prob. (1st source)	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
Prob. (2nd source)	0.1	0.2	0.1	0.3	0.05	0.1	0.05	0.1
Prob. (3rd source)	0.15	0.15	0.15	0.15	0.1	0.1	0.1	0.1

Bài tập 2

- Tìm số ký tự mã nhỏ nhất để mã tức thời cho các nguồn thông tin trong bài tập 1 sao cho độ dài mã trung bình không lớn hơn 1.5.

Symbol	A	B	C	D	E	F	G	H
Prob. (1st source)	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
Prob. (2nd source)	0.1	0.2	0.1	0.3	0.05	0.1	0.05	0.1
Prob. (3rd source)	0.15	0.15	0.15	0.15	0.1	0.1	0.1	0.1

Bài tập 3

- Tìm tất cả các mã Huffman nhị phân cho bảng ký tự {A, B, C, D}, biết rằng A xuất hiện nhiều gấp đôi B, còn B nhiều gấp đôi C và D.

Bài tập thực hành

1. Viết chương trình C tính tần suất xuất hiện của từng ký tự trong một file văn bản tiếng Anh.

2. Dùng Matlab viết hàm lập mã Huffman cho một nguồn thông tin cho trước.