CHƯƠNG 7-1: MÃ HÓA NGUỒN

TS. TRỊNH VĂN CHIẾN (SOICT-HUST)

ĐẶT VẤN ĐỀ

- Giả sử có một tệp dữ liệu với 100 000 ký tự muốn lưu trữ. Tệp dữ liệu chỉ bao gồm 06 ký tự với tần số xuất hiện như sau:

	a	b	c	d	e	f
Tần xuất x1000	45	13	12	16	9	5

- Mã nhị phân mã hóa từng ký tự dưới dạng chuỗi nhị phân hoặc từ mã
- Muốn tìm một mã nhị phân mã hóa tệp bằng cách sử dụng càng ít bit càng tốt; nén tệp càng nhiều càng tốt

ĐẶT VẤN ĐỀ

- Mã có độ dài cố định: mỗi từ mã có cùng độ dài
- Mã có độ dài thay đổi: các từ mã có thể có độ dài khác nhau
- Ví dụ:

	a	b	c	d	e	f
Tần xuất x1000	45	13	12	16	9	5
Fixed-length	000	001	010	011	100	101
Variable-length	0	101	100	111	1101	1100

- Mã có độ dài cố định cần 300000 bits để lưu trữ tệp
- Mã có độ dài thay đổi cần:

$$(45*1+13*3+12*3+16*3+9*4+5*4)*1000 = 224000\ [bits]$$

  - Giảm được rất nhiều bits
- Một mã bao gồm nhiều từ mã

VÍ DỤ

- Cho hai phương pháp mã hóa sau

a) Tính chiều dài mã trung bình của Code 1, Code 2

b) Tính lượng tin riêng trung bình của các phương pháp mã hóa

Đáp án:

a) Chiều dài trung bình mã 1 là 3, chiều dài mã 2 là 2

b) Lượng tin riêng của 2 mã là 2 bits

p_i	Code 1	Code 2
1/2	000	0
1/4	001	10
1/8	010	110
1/16	011	1110
1/64	100	111100
1/64	101	111101
1/64	110	111110
1/64	111	111111

Nên chọn mã nào?

MÃ HÓA

- Một nguồn mã (source code) C cho một biến ngẫu nhiên X

$$C(x): \tilde{X} \rightarrow D^*$$

- D^* là tập hợp các từ mã cấu thành từ một tập D
- Chiều dài của mã (code length): $l(x)$
- Ví dụ:

$$C(xanh) = 00, C(cam) = 11, \tilde{X} = \{xanh,\ cam\},\ D = \{0,1\}, D^* = \{00,11\}$$

ỨNG DỤNG MÃ HÓA NGUỒN

- Sử dụng trong ghi tín hiệu từ (magnetic recording): Đài, ổ đĩa, USB…
- Nén tiếng nói, nén ảnh
- Còn rất nhiều mảng nghiên cứu
  - Kết hợp mã hóa nguồn và mã hóa kênh trong truyền thông ngữ nghĩa (semantic communications)
  - Mã hóa nguồn không tập trung (distributed source coding) trong mạng IoT (Internet of things)

TIÊU CHÍ CỦA MỘT MÃ TỐT (1)

- Không suy biến (non-singular):

$$x \neq x' \Rightarrow C(x) \neq C(x')$$

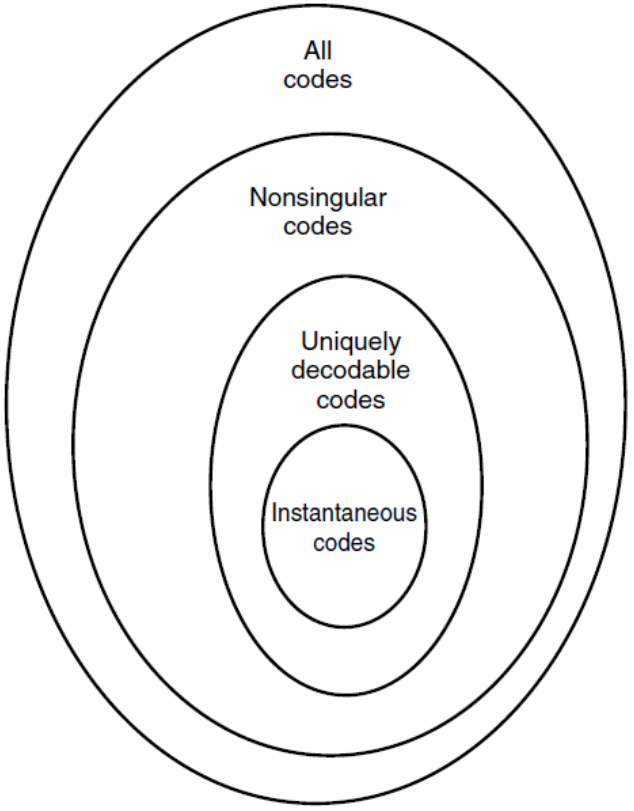
- Tính chất không suy biên được sử dụng để mô tả biến ngẫu nhiên *X*
- Khi bên phát gửi đi các chuỗi mã liên tiếp , bên thu vẫn giải mã được
- Thuộc tính giải mã duy nhất (uniquely decoable) nếu phiên bản mở rộng của mã hóa này cũng có thuộc tính không suy biến

$$C(x_1)C(x_2)...C(x_n)$$

TIÊU CHÍ CỦA MỘT MÃ TỐT (2)

- Có thể giải mã duy nhất (uniquely decoable) nếu chỉ có một chuỗi nguồn có thể tạo ra nó
  - Phải kiểm tra tất cả các chuỗi nguồn
- Mã tiền tố (prefix code)/ mã liên tục (instantaneous code): Không có một từ mã nào là tiền tố của một từ mã khác

X	Singular	Nonsingular not uniquely decodable	Uniquely decoable	Prefix
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111



CHIỀU DÀI TRUNG BÌNH TỪ MÃ

- Chiều dài trung bình từ mã (expected code length): Định nghĩa cho một mã nguồn (source code) C(x) có hàm mật độ xác suất p(x):

$$L(C) = \sum_{x \in \tilde{X}} p(x) l(x)$$

- Mong muốn xây dựng một mã tiền tố với chiều dài trung bình từ mã nhỏ nhất

BẤT ĐẲNG THỨC KRAFT (1)

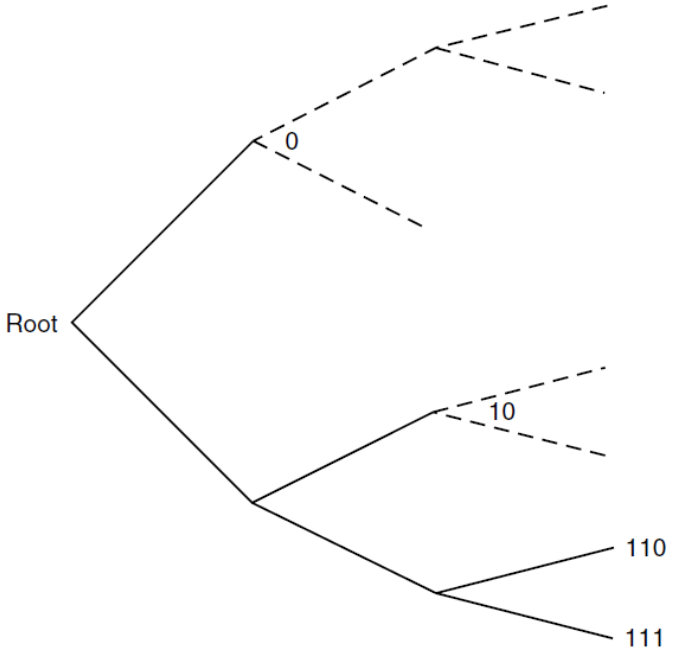
- Năm 1949, kích cỡ của code là $\tilde{D}$
- m từ mã với chiều dài $l_1,...,l_m$
- Chiều dài của tất cả các từ mã phải thỏa mãn bất đẳng thức Kraft

$$\sum_{i=1}^{m} \tilde{D}^{-l_i} \leq 1$$

- Nếu chiều dài các từ mã thỏa mãn bất đẳng thức Kraft ➔ Có thể xây dựng một mã tiền tố (prefix code)/mã liên tục

BẤT ĐẲNG THỨC KRAFT (2)

- Chứng minh:
  - Xây dựng một sơ đồ cây
  - Mỗi từ mã được đặt trên một nốt lá (leaf node)
  - Đường đi từ gốc sẽ xác định một ký hiệu
  - Mã tiền tố: Không có từ mã nào là từ mã nào là thế hệ trước (ancestor) của từ mã trên cây
  - Mỗi từ mã loại bỏ tất cả các hậu duệ (descendant) của nó



BẤT ĐẲNG THỨC KRAFT (3)

- Chiều dài từ mã lớn nhất là $l_{\max}$
- Một từ mã có chiều dài l_i sẽ có $\tilde{D}^{l_{\max}-l_i}$, tập hậu duệ và chúng là không liên quan

$$\sum \tilde{D}^{l_{\max}-l_i} \le \tilde{D}^{l_{\max}} \Rightarrow \sum \tilde{D}^{-l_i} \le 1$$

- Quy tắc:
  - Nếu $l_1,\ldots,l_m$ thỏa mãn bất đẳng thức Kraft, chúng ta có thể gán nhãn cho node đầu tiên là l_1 và loại bỏ toàn bộ hậu duệ của nó
  - Có thể xây dựng một mã tiền tố có kích cỡ lớn vô cùng nếu $l_{\max} \to \infty$

CHIỀU DÀI MÃ TỐI ƯU

- Chiều dài mã tối ưu được lựa chọn dựa vào bất đẳng thức Kraft
- Chiều dài mã được xác định bởi bất đẳng thức sau

$$L \geq H_D(X)$$

- Chiều dài mã là tối ưu (nhỏ nhất) nếu dấu bằng ở bất đẳng thức xảy ra

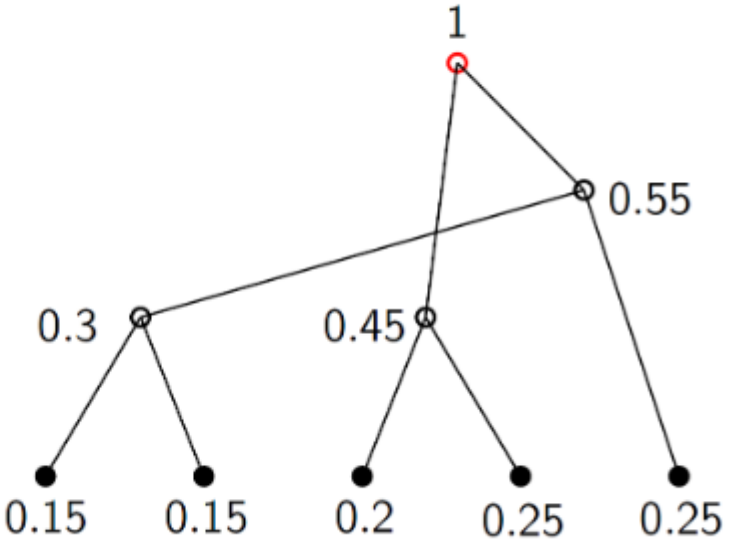
HUFFMAN CODES

Cho một nguồn tin và xác suất xuất hiện của mỗi ký tự, làm cách nào để thiết kế được một mã tối ưu (prefix code và ngắn nhất về mặt trung bình)

- **Huffman code**:
  - Bước 1: Hợp nhất D ký hiệu với xác suất nhỏ nhất để tạo ra một biểu tượng mới có xác suất là tổng của D ký hiệu trên.
  - Bước 2: Gán D ký hiệu trên với các số 0, 1,…,D-1 và quay lại Bước 1.
  - Lặp lại Bước 1, 2 đến khi tổng xác suất là 1.

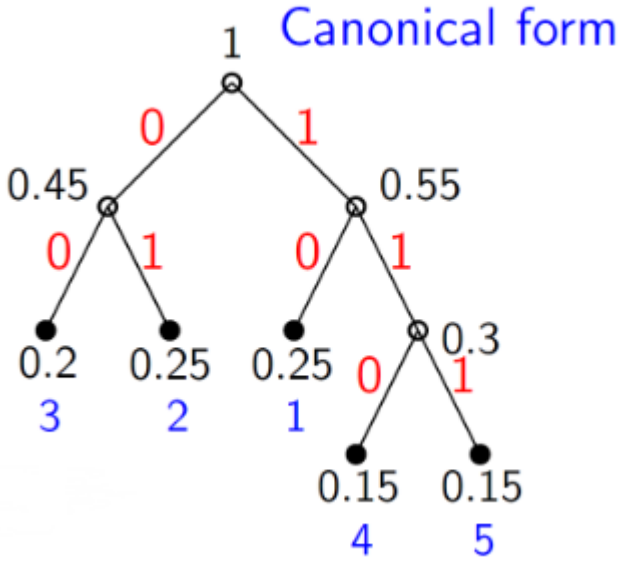
VÍ DỤ: HUFFMAN

x	p(x)
1	0.25
2	0.25
3	0.2
4	0.15
5	0.15



VÍ DỤ: HUFFMAN

x	p(x)	C(x)
1	0.25	10
2	0.25	01
3	0.2	00
4	0.15	110
5	0.15	111



$\ell(1) = \ell(2) = \ell(3) = 2, \ell(4) = \ell(5) = 3$

Ł $= \sum \ell(x)p(x) = 2.3$bits

$H_2(X) = -\sum p(x) \log_2 p(x) = 2.29$bits

$L \geq H_2(X)$

BÀI TẬP

- Xét biến ngẫu nhiên X

x	x1	x2	x3	x4	x5	x6	x7
P(x)	0.49	0.26	0.12	0.04	0.04	0.03	0.02

- Xây dựng mã Huffman cho X
- Tính chiều dài mã trung bình
- Tính lượng tin riêng
- Kiểm tra mối quan hệ của chiều dài từ mã và lượng tin riêng

BÀI TẬP

(a) The Huffman tree for this distribution is

Codeword								
1	x_1	0.49	0.49	0.49	0.49	0.49	0.51	1
00	x_2	0.26	0.26	0.26	0.26	0.26	0.49	
011	x_3	0.12	0.12	0.12	0.13	0.25		
01000	x_4	0.04	0.05	0.08	0.12			
01001	x_5	0.04	0.04	0.05				
01010	x_6	0.03	0.04					
01011	x_7	0.02						

(b) The expected length of the codewords for the binary Huffman code is 2.02 bits. ($H(X) = 2.01$ bits)

BÀI TẬP

- Mã hóa tên của bạn (không dấu) theo mã Huffman
  - Xây dựng mã Huffman
  - Tính chiều dài mã trung bình
  - Tính lượng tin riêng
  - Kiểm tra mối quan hệ của chiều dài từ mã và lượng tin riêng