Chương 3

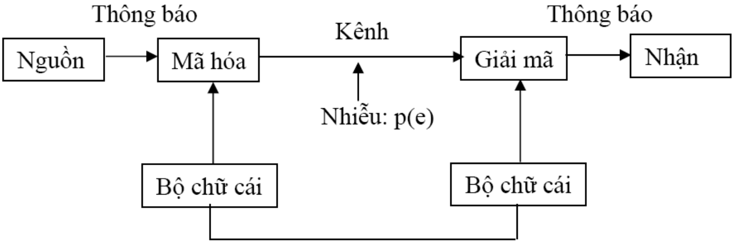
Sinh mã tách được

Nguyễn Thanh Bình

Khoa CNTT&TT - Đại học Cần Thơ

02 - 2008

Nội dung



Hệ thống liên lạc truyền tin

Tính chất

- Thông tin truyền qua kênh truyền có nhiễu cần phải được mã hóa cho phù hợp để chống nhiễu.

Bài toán sinh mã

Các ký hiệu quy ước

- $X = \{x_1, \ldots, x_M\}$: nguồn tin.
- $P = \{p_1, \ldots, p_M\}$: xác suất xuất hiện của nguồn tin X.
- $x_{i_1} x_{i_2} \ldots x_{i_n}$: thông báo / thông điệp (các $x_i \in X$ và có thể lặp lại).
- $A = \{a_1, \ldots, a_D\}$: tập hợp ký tự mã / bảng chữ cái sinh mã. D gọi là cơ số sinh mã.
- $W = \{w_1, \ldots, w_M\}$: tập hợp các từ mã. Từ mã w_i là một dãy hữu hạn các ký tự mã gán cho giá trị $x_i \in X$.
- $N = \{n_1, \ldots, n_M\}$: độ dài các từ mã.

Bài toán sinh mã

Mã tách được

Bộ mã được gọi là tách được nếu như ta luôn giải mã được với kết quả duy nhất từ một dãy liên tục các ký tự mã nhận được.

Sinh mã tối ưu

- Bài toán sinh mã tối ưu được đặt ra ở đây là: tìm phương pháp sinh mã sao cho độ dài trung bình các từ mã là nhỏ nhất.

$$\sum_{i=1}^{M} p_i n_i \rightarrow Min$$

- Huffman(1950) đã đưa ra quy trình xây dựng bảng mã tối ưu thỏa yêu cầu này.

Bảng mã không tách được

Định nghĩa

- Mã hóa thông báo **Msg** $\Rightarrow$ nhận được dãy các từ mã **ws**.
- Giải mã dãy các từ mã **ws** $\Rightarrow$ nhận được nhiều thông báo **Msg** khác nhau.

Ví dụ

- Cho $X = \{x_1, x_2, x_3, x_4\}$ có bảng mã $W = \{w_1 = 0, w_2 = 1, w_3 = 01, w_4 = 10\}$.
- Thông báo $Msg = x_1 x_2 x_3 x_4 x_3 x_2 x_1 \Rightarrow$ dãy mã $ws = 0101100110$.
- Giải mã có thể nhận được: $x_3 x_3 x_4 x_3 x_4$ (khác thông báo gốc).

Bảng mã tách được

Định nghĩa

- Mã hóa thông báo **Msg** $\Rightarrow$ nhận được dãy các từ mã **ws**.
- Giải mã dãy các từ mã **ws** $\Rightarrow$ nhận được duy nhất thông báo **Msg** ban đầu.

Ví dụ

- Cho $X = \{x_1, x_2\}$ có bảng mã $W = \{w_1 = 0, w_2 = 01\}$.
- Cần giải mã: $ws = 0010000101001$.
- Phương pháp giải mã: chỉ giải mã khi nào đã nhận được đoạn mã với độ dài bằng độ dài của từ mã dài nhất.
- Giải mã nhận được duy nhất: $Msg = x_1x_2x_1x_1x_1x_2x_2x_1x_2$ (thông báo gốc).

Bảng mã tức thời

Nhận xét về bảng mã tách được

- Không tồn tại từ mã này là mã khóa của từ mã khác.
- Có thể tồn tại từ mã này là tiền tố của từ mã khác.

Định nghĩa bảng mã tức thời

Là bảng mã tách được và không tồn tại từ mã này là tiền tố của từ mã khác.

Ví dụ

- Bảng mã $W = \{w_1 = 10, w_2 = 101, w_3 = 100\}$ không là bảng mã tức thời (w_1 là tiền tố của w_2).
- Bảng mã $W = \{w_1 = 0, w_2 = 110, w_3 = 101\}$ là tức thời.

Input: bảng mã W.
Output: kết luận bảng mã tách được hay không tách được.

- **Khởi tạo**: gán $S_0 = W$.
- **Bước 1**: xác định S_1 từ S_0

$$S_1 = \{A \mid \forall w_i, w_j \in S_0 : w_i = w_j A \text{ hoặc } w_j = w_i A\}$$

(tìm trong S_0 2 từ mã là tiền tố của nhau $\Rightarrow$ đưa hậu tố vào S_1)

- **Bước lặp k**: xác định $S_k (k \geq 2)$ từ S_0 và S_{k-1}

$$S_k = \{B \mid \forall w \in S_0 \text{ và } \forall C \in S_{k-1} : w = CB \text{ hoặc } C = wB\}$$

(tìm trong S_0 và S_{k-1} 2 từ mã tiền tố của nhau $\Rightarrow$ hậu tố vào S_k)

- **Điều kiện dừng**:
  - Nếu $S_k \cap S_0 \neq \emptyset$: kết luận bảng mã không tách được.
  - Nếu $S_k = \emptyset$ hoặc $S_k = S_{t<k}$: bảng mã tách được ($k \geq 1$).

Ví dụ minh họa: bài toán 1

- Kiểm tra tính tách được của bảng mã:

$$W = \{a, c, ad, abb, bad, deb, bbcde\}$$

Áp dụng thuật toán

- **Khởi tạo**: $S_0 = W = \{a, c, ad, abb, bad, deb, bbcde\}$
- **Bước 1**: Tính S_1. Khởi tạo $S_1 = \emptyset$.
  - a là tiền tố của $ad \Rightarrow$ đưa d vào $S_1 \Rightarrow S_1 = \{d\}$
  - a là tiền tố của $abb \Rightarrow$ đưa bb vào $S_1 \Rightarrow S_1 = \{d, bb\}$
  - Kiểm tra điều kiện dừng: không thỏa $\Rightarrow$ qua bước 2.
- **Bước 2**: Tính S_2. Khởi tạo $S_2 = \emptyset$.
  - $d \in S_1$ là tiền tố của $deb \in S_0 \Rightarrow$ đưa eb vào S_2.
  - $bb \in S_1$ là tiền tố của $bbcde \in S_0 \Rightarrow$ đưa cde vào S_2.
  - Kiểm tra điều kiện dừng: không thỏa $\Rightarrow$ qua bước 3.

Ví dụ minh họa: bài toán 1

Áp dụng thuật toán

- **Bước 3**: Tính S_3. Khởi tạo $S_3 = \emptyset$.
  - $c \in S_0$ là tiền tố của $cde \in S_2 \Rightarrow$ đưa de vào S_3.
  - Kiểm tra điều kiện dừng: không thỏa $\Rightarrow$ qua bước 4.
- **Bước 4**: Tính S_4. Khởi tạo $S_4 = \emptyset$.
  - $de \in S_3$ là tiền tố của $deb \in S_0 \Rightarrow$ đưa b vào S_4.
  - Kiểm tra điều kiện dừng: không thỏa $\Rightarrow$ qua bước 5.
- **Bước 5**: Tính S_5. Khởi tạo $S_5 = \emptyset$.
  - $b \in S_4$ là tiền tố của $bad \in S_0 \Rightarrow$ đưa ad vào S_5.
  - $b \in S_4$ là tiền tố của $bbced \in S_0 \Rightarrow$ đưa $bcde$ vào $S_5 \Rightarrow S_5 = \{ad, bcde\}$.
  - Kiểm tra điều kiện dừng: vì S_5 có chứa từ mã $ad \in S_0$ nên dừng lại và kết luận: bảng mã không tách được.

Ví dụ minh họa: bài toán 2

- Kiểm tra tính tách được của bảng mã:

  $W = 010, 0001, 0110, 1100, 00011, 00110, 11110, 101011\}$

- $S_0 = \{010, 0001, 0110, 1100, 00011, 00110, 11110, 101011\}$
- $S_1 = \{1\}$
- $S_2 = \{100, 1110, 01011\}$
- $S_3 = \{11\}$
- $S_4 = \{00, 110\}$
- $S_5 = \{01, 0, 011, 110\}$
- $S_6 = \{0, 10, 001, 110, 0011, 0110\}$
- Vì S_6 có chứa từ mã 0110 $\Rightarrow$ bảng mã không tách được.

Định lý Kraft (1949)

Định lý

Điều kiện cần và đủ để tồn tại bảng mã tức thời là:

$$\sum_{i=1}^{M} D^{-n_i} \leq 1$$

Ví dụ

- $W = \{w_1, w_2, w_3\}$ với $M = 3, n_1 = n_2 = 1, n_3 = 2, D = 2.$

$$\sum_{i=1}^{M} D^{-n_i} = \frac{1}{2^1} + \frac{1}{2^1} + \frac{1}{2^2} = \frac{5}{4} > 1$$

$\Rightarrow$ không tồn tại bảng mã tức thời.

Định nghĩa cây bậc D cỡ k

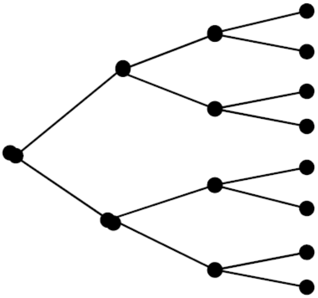
Định nghĩa

Cây bậc D cỡ k có hệ thống nút, cạnh thỏa điều kiện:

- Một nút không có quá D nút con.
- Nút lá cách nút gốc không vượt quá k cạnh.

Ví dụ

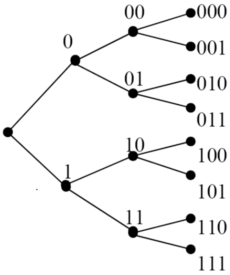
Cây bậc $D = 2$ cỡ $k = 3$



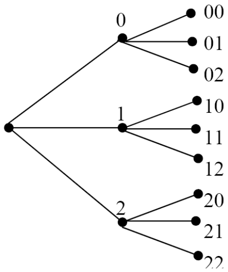
Phương pháp sinh mã

Mỗi nút (trừ nút gốc) được ký hiệu bởi dãy ký hiệu của nút cha làm tiền tố + một ký tự bổ sung lấy từ tập hợp $\{0, \ldots, D-1\}$.

Cây bậc $D = 2$ cỡ $k = 3$



Cây bậc $D = 3$ cỡ $k = 2$



Chứng minh định lý Kraft: điều kiện cần

Cho trước bảng mã tức thời $W = \{w_1, \dots, w_M\}$ với $N = \{n_1 \le \dots \le n_M\}$. Ta cần chứng minh: $\sum_{i=1}^{M} D^{-n_i} \le 1$.

- Xây dựng cây bậc D cỡ n_M và sinh mã cho cây.
- Mỗi nút (trừ nút gốc) đều có thể được chọn làm từ mã.
- Quy tắc: một nút được chọn thì tất cả các nút kề sau phải được xóa (tránh tiền tố).
- Chọn nút có độ dài mã n_1 gán cho $w_1 \Rightarrow$ xóa $D^{n_M - n_1}$ nút lá.
- ...
- Chọn nút có độ dài mã n_M gán cho $w_M \Rightarrow$ 1 nút lá gán mã.
- Tổng số nút bị xóa (hoặc gán từ mã):

$$\sum_{i=1}^{M} D^{n_M - n_i} \le D^{n_M} \text{ (tổng số nút lá)} \Rightarrow \sum_{i=1}^{M} D^{-n_i} \le 1$$

Chứng minh định lý Kraft: điều kiện đủ

Giả sử $\sum_{i=1}^{M} D^{-n_i} \leq 1$. Ta cần chỉ ra thủ tục xây dựng bảng mã tức thời.

Xét $N = \{n_1, \ldots, n_M\}$ và cơ số sinh mã D.

1. Xếp thứ tự $n_1 \leq \ldots \leq n_M$. Xây dựng cây bậc D cỡ $k = n_M$ và sinh mã cho các nút.
2. Chọn nút bất kỳ trên cây có độ dài n_1 gán cho từ mã w_1 và xóa tất cả các nút kề sau nó.
3. Lặp lại bước 2 đối với các từ mã còn lại $w_2, \ldots, w_M$ ứng với độ dài $n_2, \ldots, n_M$.

   $\Rightarrow$ Bảng mã $W = \{w_1, \ldots, w_M\}$ là tức thời.

Ví dụ minh họa

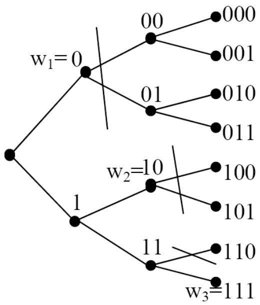
Ví dụ

- $W = \{w_1, w_2, w_3\}$ với $M = 3, n_1 = 1, n_2 = 2, n_3 = 3, D = 2.$

$$\sum_{i=1}^{M} D^{-n_i} = \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} = \frac{7}{8} \leq 1$$

$\Rightarrow$ có tồn tại bảng mã tức thời $W = \{w_1, w_2, w_3\}$.

- Chọn $w_1 = 0$, cắt bỏ các nút con của w_1.
- Chọn $w_2 = 10$, cắt bỏ các nút con của w_2.
- Chọn $w_3 = 111$



Định lý

- Đặt $\overline{n} = \sum_{i=1}^{M} p_i n_i$ là độ dài trung bình của bảng mã tách được. Khi đó:
$$\overline{n} \geq \frac{H(X)}{\log_2 D}$$
- Dấu đẳng thức xảy ra khi và chỉ khi: $p_i = D^{-n_i}$ hay $\sum_{i=1}^{M} D^{-n_i} = 1$.

Chú ý

$$H_D(X) = -\sum p_i \log_D p_i = \frac{-\sum p_i \log_2 p_i}{\log_2 D} = \frac{H(X)}{\log_2 D}$$

là Entropy với cơ số D của X.

Bảng mã tối ưu tuyệt đối và tương đối

Định lý bảng mã tối ưu tuyệt đối

Bảng mã được gọi là tối ưu tuyệt đối khi:

$$\overline{n} = \frac{H(X)}{\log_2 D} \text{ hay } p_i = D^{-n_i}$$

Định lý bảng mã tối ưu tương đối

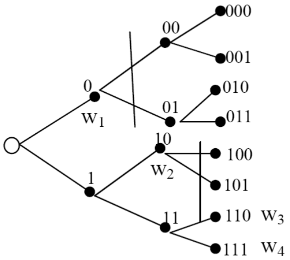
Bảng mã được gọi là tối ưu tương đối khi:

$$\frac{H(X)}{\log_2 D} \leq \overline{n} \leq \frac{H(X)}{\log_2 D} + 1$$

Ví dụ

- Biến ngẫu nhiên: $X = \{x_1, x_2, x_3, x_4\}$
- Phân phối: $P = \{1/2, 1/4, 1/8, 1/8\}$
- Bảng mã: $W = \{w_1 = 0, w_2 = 10, w3 = 110, w4 = 111\}$
- Độ dài trung bình từ mã: $\overline{n} = \sum_{i=1}^{M} p_i n_i = 1.75$
- Entropy $H(X) = -\sum_{i=1}^{M} p_i \log_2 p_i = 1.75$

$\Rightarrow W$ là tối ưu tuyệt đối



Điều kiện nhận biết một bảng mã tối ưu

Định lý

- Xác suất p_i càng lớn thì độ dài n_i của w_i càng nhỏ.
- Nếu $p_i \geq p_j$ thì $n_i \leq n_j$.
- 2 từ mã ứng với 2 giá trị có xác suất nhỏ nhất sẽ có độ dài mã bằng nhau $n_{M-1} = n_M$.
- Trong các từ mã có độ dài bằng nhau và bằng n_M (dài nhất) thì tồn tại ít nhất 2 từ mã w_{M-1} và w_M có $n_M - 1$ ký tự đầu giống nhau.

Ví dụ

Bảng mã $W = \{w_1 = 0, w_2 = 100, w_3 = 1101, w_4 = 1110\}$ không là bảng mã tối ưu vì 2 từ mã w_3 và w_4 có 3 ký tự đầu khác nhau.

Phương pháp sinh mã Huffman

Ghi chú

Ở đây ta chỉ xét phương pháp sinh mã Huffman với cơ số $D = 2$.

Thủ tục lùi

1. Sắp xếp các giá trị của X theo xác suất p giảm dần.
2. Xét 2 giá trị x_i, x_j có xác suất nhỏ nhất, gán mã 0 và 1 cho mỗi từ.
3. Tạo một giá trị mới x_{ij} có xác suất bằng $p_i + p_j$.
4. Lặp lại bước 1 cho đến khi chỉ còn 2 giá trị của X.

Thủ tục tiến

Lần vết ngược lại của thủ tục lùi để tìm ra các từ mã.

Ví dụ minh họa

Thủ tục lùi:

Bước 1

X	P
x_1	0.3
x_2	0.25
x_3	0.2
x_4	0.1
x_5	0.1
x_6	0.05

Bước 2

X	P
x_1	0.3
x_2	0.25
x_3	02
x_{56}	0.15
x_4	0.1

Bước 3

X	P
x_1	0.3
x_{564}	0.25
x_2	0,25
x_3	0.2

Bước 4

X	P
x_{23}	0.45
x_1	0.3
x_{564}	0.25

Bước 5

X	P
x_{1564}	0.55 $\leftarrow 0$
x_{23}	0.45 $\leftarrow 1$

Thủ tục tiến:

Bước 1

X	W
x_{1564}	0
x_{23}	1

Bước 2

X	W
x_{23}	1
x_1	00
x_{564}	01

Bước 3

X	W
x_1	00
x_{564}	01
x_2	10
x_3	11

Bước 4

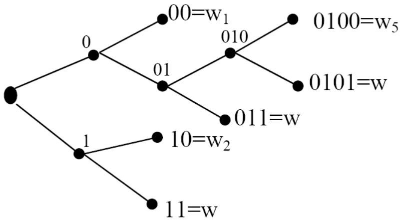
X	W
x_1	00
x_2	10
x_3	11
x_{56}	010
x_4	011

Bước 5

X	W	
x_1	00	$= w_1$
x_2	10	$= w_2$
x_3	11	$= w_3$
x_4	011	$= w_4$
x_5	0100	$= w_5$
x_6	0101	$= w_6$

Tính tối ưu của bảng mã Huffman

Cây Huffman của bảng mã vừa xây dựng



Nhận xét

- Độ dài trung bình từ mã: $\overline{n} = \sum_{i=1}^{M} p_i n_i = 2.4$
- Entropy: $H(X) = H(0.3, 0.25, 0.2, 0.1, 0.1, 0.05) = 2.4$
- Ta có $\overline{n} = H(X) \Rightarrow$ bảng mã tối ưu tuyệt đối.