Cơ sở Lý thuyết Truyền tin-2004

Hà Quốc Trung[1]

[1]Khoa Công nghệ thông tin
Đại học Bách khoa Hà nội

Chương 4: Mã hiệu

Giới thiệu

- Xử lí thông tin:
  - Có một nguồn tin nguyên thủy
  - Biến đổi nguồn tin nguyên thủy cho phù hợp với các quá trình xử lí thành các nguồn tin trung gian khác
  - Biến đổi ngược từ các nguồn tin thành nguồn tin có dạng ban đầu
  - Biểu diễn của các nguồn tin trung gian bằng *mã hiệu*
- Mã hiệu:
  - Các khái niệm liên quan
  - Điều kiện để sử dụng được mã hiệu
  - Cách biểu diễn mã hiệu

1. Mã hiệu, tham số, đặc tính

1.1.Khái niệm mã hiệu

- Mã hiệu là mã sử dụng tập ký hiệu số (các chữ số) để mã hóa thông tin
- Mã hóa
  - một song ánh giữa hai nguồn tin (một phép biến đổi 1-1 giữa các tin của hai nguồn tin)
  - Kết quả thu được là một nguồn tin có các thông số thống kê phù hợp:

    Entropy
    Độ chính xác
    Chiều dài các tin
  - Kết quả thu được này là mã hiệu
- Vậy mã hiệu là một nguồn tin với mô hình thống kê xác định trước, thỏa mãn yêu cầu nào đó, sử dụng các ký hiệu số

Các khái niệm liên quan của mã hiệu

- Mã hiệu gồm một tập hữu hạn các ký hiệu có phân bố xác suất nào đó, gọi là *dấu mã* hay *ký hiệu mã*
- Tập hợp một số nào đó các dấu mã gọi là *tổ hợp mã*
- Trong tập hợp tất cả các tổ hợp mã, một tập hợp các tổ hợp mã được xây dựng theo một luật nào đó, gọi là tổ hợp mã có thể (hợp lệ)
- Trong quá trình mã hóa, một tin của nguồn nguyên thủy được ánh xạ vào một tổ hợp mã. Một tổ hợp mã như vậy gọi là *từ mã*. Những tổ hợp có thể khác gọi là tổ hợp cấm (tổ hợp không sử dụng)
- Một dãy từ mã bất kỳ tạo thành một *từ thông tin*

Ví dụ: mã BCD Binary Coded Decimal đóng gói

- Nguồn tin nguyên thủy gồm các tin là các ký hiệu từ $0 - 9$
- Mã hóa thành các ký hiệu nhị phân $0 - 1$
- Các dấu (ký hiệu mã): $0, 1$
- Các tổ hợp mã có thể: 0000 đến 1111, gồm 16 tổ hợp mã
- Các tổ hợp mã được sử dụng (từ mã):

0	1	2	.....	9
0000	0001	0010	.....	1001

- Các tổ hợp mã bị cấm: 1010,1011,1100,1101,1110,1111
- Một từ thông tin:

$$2005 \rightarrow 0010000000000101$$

$$0010000000010100$$

Các khái niệm liên quan của mã hiệu

- Quá trình biến đổi nguồn tin ban đầu sử dụng mã hiệu gọi là *quá trình mã hóa*.
- Nguồn tin rời rạc gồm nhiều tin tạo thành bản tin. Các nguồn tin trong thực tế có số lượng các tin rất lớn. Ngược lại các mã hiệu thường có số lượng các ký hiệu tương đối nhỏ. Do đó một tin của nguồn ban đầu thường được mã hóa thành một chuỗi các ký hiệu mã:(từ mã)
- Quá trình biến đổi ngược lại từ một từ mã thành một tin ban đầu gọi là *quá trình giải mã*
- Ngoại lệ: mã hóa một chuỗi các tin của nguồn tin nguyên thủy thành một hoặc nhiều từ mã: *mã khối (mã theo từ)*

1.2.Các thông số cơ bản của mã hiệu

- Mã hiệu là một tập hợp các từ mã, thành lập từ một bảng ký hiệu
- Số lượng ký hiệu trong bảng ký hiệu gọi là *cơ số*
- *Độ dài của từ mã*: số lượng các ký hiệu của từ mã
- Độ dài trung bình của từ mã:

$$\overline{R} = \sum_{i=1}^{L} p(x_i) n_i$$

- L là tổng số từ mã: số tin được mã hóa, số từ mã, số tổ hợp mã có thể được sử dụng
  - Mã đầy: $L = M$, M là tổng số các từ mã có thể
  - Mã vơi: $L < m^n = M$
  - $R = M - L$: số các tổ hợp bị cấm (không sử dụng)
- Độ đo của từ mã

1.2.Các thông số cơ bản của mã hiệu *(Tiếp)*

- Để thuận tiện cho việc sử dụng mã hiệu, mỗi từ mã được gán cho một độ đo: trọng số
- Độ đo đơn giản nhất cho một từ của một bảng chữ cái: hệ đếm theo vị trí
- Số lượng ký hiệu gọi là cơ số của mã hiệu
- Mỗi ký hiệu được gán cho một giá trị gọi là *giá trị riêng* hay *trị* của ký hiệu. Ví dụ m ký hiệu có thể được gán các trị tương ứng là $0, 1, 2 \ldots m-1$
- *Chỉ số vị trí*: số thứ tự của mỗi ký hiệu trong từ mã. Ví dụ: đánh số từ 0, từ phải qua trái
- Trọng số vị trí w_k: hệ số nhân của từng vị trí ký hiệu k. Ví dụ: trong hệ đếm cơ số 10, trọng số của vị trí đầu tiên là 1, thứ 2 là 10,....
- Trọng số (giá trị) của từ mã:

$$b = \sum_{k=0}^{n-1} a_k w_k$$

Trong hệ đếm cơ số m $b = \sum_{k=0}^{n-1} a_k m^k$

1.2.Các thông số cơ bản của mã hiệu *(Tiếp)*

- Khoảng cách giữa hai từ mã có thể đo bằng
  - Hiệu giữa hai trọng số
  - Một độ đo định nghĩa riêng
- Hàm cấu trúc của mã hiệu
  - Cho biết phân bố của các từ mã theo độ dài
  - Hàm cấu trúc của mã đồng đều?

1.3.Đặc tính của mã hiệu

- Tính đồng đều: tất cả các từ mã có cùng một độ dài
- Tính đầy: Tất cả các từ mã có thể đều được sử dụng Ví dụ nếu chiều dài lớn nhất của từ mã là n_{max}, số lượng từ mã là $m^{n_{max}+1} - 1$
- Tính phân tách được: cho một từ thông tin, liệu có thể phân tách được *một cách duy nhất* từ thông tin đó ra một hoặc nhiều từ mã hay không?
- Ví dụ

tin nguyên thủy	mã hiệu 1	mã hiệu 2
a_1	00	0
a_2	01	00
a_3	10	10
a_4	11	11

- Từ thông tin 00010 với mã hiệu 2 có thể phân tách thành 0-0-0-10 hoặc 0-00-10. Vậy mã hiệu 2 không có tính phân tách được

1.3.Đặc tính của mã hiệu *(Tiếp)*

- Tính phân tách được quyết định việc giải mã
- Các điều kiện khác
  - Tối ưu về độ dài
  - Tối ưu về khả năng sửa sai
  - Tối ưu về thời gian giải mã

2. Điều kiện để mã phân tách được

2.1.Khả năng giải mã và độ chậm giải mã

- Bài toán giải mã
  - Nhận lần lượt từng dấu ký hiệu mã
  - Kiểm tra và tách chuỗi ký hiệu mã thu được thành các từ mã ???
  - Chuyển đổi các từ mã thành các ký hiệu của nguồn tin ban đầu
- Điều kiện giải mã
  - Chuyển đổi giữa các tin ban đầu thành các từ mã là 1-1
  - Có thể phân tách chuỗi ký hiệu mã nhận được thành các từ mã
  - Số lượng ký hiệu tối thiểu để có thể nhận dạng được một từ mã gọi là *độ chậm giải mã (độ trễ mã)*

Thuật toán giải mã

1. Nhận một ký hiệu vào bộ đệm: $B = B + a_i$
2. Kiểm tra nội dung bộ đệm (phân tách) xem có thể tách một cách duy nhất thành tổ hợp các từ mã hay không. Nếu không quay trở lại 1
3. Giải mã các từ mã. Xóa bộ đệm: $B = \emptyset$

Ví dụ

Các tiêu chuẩn (phương pháp phân tách)

- Căn cứ vào tính prefix của mã (?) tiền tố
  - Nhanh
  - Độ dài các từ mã khác nhau
  - Chống nhiễu kém
- Căn cứ vào dấu phân tách
  - Chống nhiễu tốt
  - Hiệu suất thấp
- Căn cứ vào chiều dài từ mã
  - Đơn giản
  - Chống nhiễu kém
- Chú ý: 2 và 3 thực chất là trường hợp riêng của 1

2.2.Điều kiện để mã phân tách được

- Điều kiện: bất cứ một dãy các từ mã nào không được trùng với một dãy các từ mã khác
- Vậy để xác định mã phân tách được hay không cần xác định : Tồn tại hay không một dãy từ mã trùng với một dãy từ mã khác
- Bảng thử mã
  - Liệt kê các từ mã ở cột 1 theo thứ tự chiều dài tăng dần
  - Kiểm tra theo thứ tự chiều dài tăng dần xem các từ mã có là phần đầu của một từ mã dài hơn hay không.
  - Nếu có, ghi phần còn lại của từ mã dài vào cột thứ 2
  - Với các từ mã thu được trong cột thứ hai, so sánh với các từ mã trong cột 1, nếu là phần đầu của một từ mã, ghi phần còn lại vào cột thứ 3
  - tiếp tục cho đến khi nào thu được cột trống
- Điều kiện cần và đủ để mã phân tách được: *không có một tổ hợp mã nào trong các cột từ thứ 2 trở đi là một từ mã trong cột 1*

Ví dụ 1

1	2	3
00		
01		
100		
1010		
1011		

- Trong trường hợp này, khi nhận hết các ký hiệu của một từ mã, có thể nhận dạng ngay từ mã. Vậy độ chậm giải mã bằng chiều dài từ mã

Ví dụ 2

1	2	3	4	5	6
10	0	1	0	1	...
100	1	11	00	11	...
01		0	1	0	...
011		00	11	00	...

- Bảng thử thỏa mãn yêu cầu định lý, nên mã này là mã phân tách được
- Độ chậm giải mã là vô hạn: chỉ khinào nhận hết bản tin, mới có thể phân tách được bản tin thành các từ mã
- Ví dụ dãy vô hạn 10010101010.... nếu không biết ký hiệu cuối cùng sẽ không phân tách được các từ mã
- Nếu chỉ xét dãy hữu hạn 10010101010, chỉ tồn tại một cách phân tách duy nhất 100-10-10-10-10
- Có thể đánh giá độ chậm giải mã

$$[\frac{j-1}{2}]n_{min} \leq T_{ch} \leq [\frac{j-1}{2}]n_{max}$$

2.3.Mã có tính prefix (tiền tố)

- Nếu bộ mã không có từ mã nào là phần đầu của một bộ mã khác, bộ mã là mã phân tách được
- Bộ mã như vậy gọi là *mã prefix*
- Biểu diễn mã prefix bằng cây: tất cả các từ mã đều biểu diễn bằng các nút lá, không có hai từ mã nào cùng nằm trên một đường tới gốc
- Mã đầy là mã prefix

Hàm cấu trúc của mã prefix

$$G(1) \leq m$$

$$G(2) \leq m^2 - mG(1)$$

$$\dots$$

$$G(n) \leq m^n - \sum_{j=1}^{n-1} m^{n-j} G(j)$$

$m^n \geq \sum_{j=1}^{n} m^{n-j} G(j)$ hay $1 \geq \sum_{j=1}^{n} m^{-j} G(j)$, dấu bằng xảy ra khi bộ mã là mã đầy

Ngược lại, nếu dãy số $n_j, 1 \leq j \leq k$ thỏa mãn

$$1 \geq \sum_{j=1}^{n} m^{-n_j}$$

Tồn tại bộ mã prefix với cơ số m với độ dài của các từ mã là n_j

Bất đẳng thức này còn gọi là bất đẳng thức Kraft(McMillan)

3. Phương pháp biểu diễn mã

3.1.Các bảng mã

- Bảng đối chiếu
  - Liệt kê tin và từ mã tương ứng bằng bảng
  - Ví dụ

Tin	a	b	c	d
Từ mã	00	01	10	11

- Mặt tọa độ
  - Trục hoành: độ dài từ mã, trục tung: trọng số của từ mã
  - Định lý: không tồn tại hai từ mã có cùng độ dài và cùng trọng số
  - Ví dụ 00,01,100,1010,1011

3.2.Cây mã

- Biểu diễn các từ mã sử dụng bằng một cây
- Gốc có *m* nhánh tương ứng với m khả năng của ký hiệu thứ nhất
- Các nút tiếp theo có các nhánh tương ứng với khả năng của ký hiệu tiếp theo
- Mỗi từ mã được biểu diễn bằng một nút, tương ứng với đường dẫn từ gốc đến nút đó
- Mỗi nút cuối tương ứng với một từ mã
- Căn cứ vào cây mã, ta có thể xác định được mã đầy, mã vơi, mã đồng đều hay mã không đồng đều, mã có tính prefix hay không

3.3.Đồ hình kết cấu

- Là các biểu diễn rút gọn của cây mã
- Mỗi cung biểu diễn một hoặc nhiều ký hiệu
- Mỗi từ mã biểu diễn bằng một vòng khép kín đi từ gốc qua các nút trung gian; các cung tương ứng với các ký hiệu rồi trở lại gốc
- Ví dụ: bộ mã 00, 01,100,1010,1011
- Đồ hình kết cấu thuận tiện cho việc tìm cách giải mã

3.4.Ví dụ về các phương pháp biểu diễn mã hiệu

- Cho bộ mã 00,10,110,1110,11110,11111
- Biểu diễn bằng bảng đối chiếu
- Biểu diễn bằng mặt tọa độ
- Biểu diễn bằng cây nhị phân
- Biểu diễn bằng đồ hình kết cấu
- Hàm cấu trúc của mã

3.5.Các phương pháp biểu diễn mã khác

- Biểu diễn hình học:
  - mỗi từ mã gồm n ký hiệu, mỗi ký hiệu có m giá trị
  - có thể biểu diễn mỗi từ mã như một điểm trong không gian n chiều
  - Bộ mã sẽlà một bộ điểm trong không gian n chiều
- Biểu diễn bằng một cấu trúc đại số.

4. Mã hệ thống

4.1.Mã hệ thống có tính prefix

- Mã hệ thống: từ mã được tạo thành từ một bộ các từ mã gốc
- Có thể coi là mã hiệu lập hai lần: các ký hiệu->mã gốc->mã hệ thống
- Mã hệ thống thường dùng
  - Các từ mã thuộc mã gốc chia làm 2 loại: từ mã sơ đẳng và từ mã kết thúc
  - Một từ mã hệ thống tạo thành bằng nhiều từ mã sơ đẳng và một từ mã kết thúc
- Biểu diễn
  - Sử dụng đồ hình kết cấu của mã gốc, biến đổi
  - Các từ mã sơ đẳng kết thúc tại gốc
  - Các từ mã kết thúc kết thúc tại một điểm đặc biệt (nút kết thúc)
  - Một từ mã hệ thống được biểu diễn bởi một đường có thể đi qua nút gốc nhiều lần, kết thúc ở nút kết thúc

4.1.Mã hệ thống có tính prefix *(Tiếp)*

- Giải mã
  - Cần qua hai bước: tách các từ mã gốc, sau đó xác định các từ mã kết thúc để tách các từ mã hệ thống
  - Có thể dùng đồ hình kết cấu để giải mã
- Khi mã gốc có tính prefix, thì mã hệ thống cũng có tính prefix, gọi là *mã hệ thống có tính prefix*

Ví dụ

- Mã gốc 1,00,010,011 làm gốc
- 1,00,010 là các từ mã sơ đẳng, 011 là từ mã kết thúc
- Các từ mã hệ thống sẽ là 100011, 1010011, 0100100100011 .....
- Giải mã
  - Tách các từ mã gốc
  - Tách các từ mã hệ thống
- Hàm cấu trúc của mã hệ thống

4.2.Mã có dấu phân cách

- Trong ví dụ trên, quá trình phân tách mã tương đối phức tạp
- Quá trình này phụ thuộc hoàn toàn vào các từ mã kết thúc
- Để đơn giản hóa, ta có thể dùng một từ mã kết thúc gọi là dấu phân cách để tách các từ mã hệ thống
- Tổng quát hơn, chúng ta có thể dung một ký hiệu, một chuỗi ký hiệu đặc biệt để phân tách các từ mã. Chuỗi này không được trùng với bất cứ một từ mã nào trong bộ mã
- Dấu phân cách thường được thiết kế để có khả năng chống nhiễu rất lớn. Khi đó quá trình truyền (xử lí) tin được chia thành nhiều công đoạn độc lập lẫn nhau bằng các dấu phân cách (đồng bộ hóa)
- Khi giải mã, các ký hiệu nhận được được ghi vào một bộ đệm rồi so sánh với dấu phân cách (ví dụ: Bộ lọc tuyến tính điều chỉnh theo dấu phân cách)

Bộ lọc tuyến tính sử dụng dấu phân cách

- Điều kiện của dấu phân cách và các tổ hợp mã khác
- Cho chuỗi dấu mã $x_1, x_2 \ldots x_k$
- Tổ hợp mã $a_1, a_2 \ldots a_l$ là một từ mã nếu k và chỉ k ký hiệu cuối cùng của dãy $x_2 \ldots x_k a_1 a_2 \ldots a_l$ trùng với dấu phân cách

Chương 4: Mã hiệu