Mã hóa nguồn - Nén dữ liệu

Lý thuyết thông tin

Biên soạn: Phạm Văn Sự

Bộ môn Xử lý tín hiệu và Truyền thông
Khoa Kỹ thuật Điện tử I
Học viện Công nghệ Bưu chính Viễn thông

20/08/2011

Tổng quan về mã hóa nguồn

Mục tiêu và phân loại

Mục tiêu của mã hóa nguồn

Thực hiện tìm kiếm các phương thức biểu diễn dữ liệu nhỏ gọn nhất có thể

Nguyên lý của mã hóa nguồn

Loại bỏ các thông tin dư thừa hoặc các thông tin dư thừa và các thông tin không cần thiết.

- Theo quan điểm bảo toàn thông tin:
  - Nén không tổn hao (lossless data compression)
  - Nén có tổn hao (lossy data compression)
- Theo đặc tính thay đổi:
  - Mã thích nghi (adaptive)
  - Mã không thích nghi (nonadaptive)
- Theo phương pháp:
  - RLE (run length encoding)
  - Mã hóa thống kê
  - Mã hóa từ điển
  - Mã hóa chuyển đổi
- Theo mô hình n-user:
  - Tập trung
  - Phân tán

Mã hóa

Cho nguồn rời rạc X với các tin x_k có xác suất phân bố $p(x_k)$. Một bộ dấu (chữ mã) M với các dấu (chữ mã) $\{m_1, m_2, \ldots, m_q\}$.

Định nghĩa (Mã hóa)

Mã hóa là một phép ánh xạ $1-1$ từ tập các tin rời rạc x_k lên tập các từ mã là tổ hợp có thể của các dấu (các chữ mã) m_k

$$f : x_k \longmapsto m_k^{l_k}$$

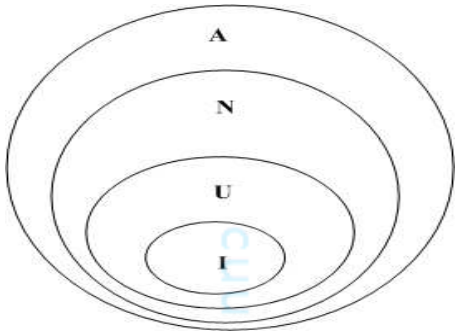
- l_k là độ dài từ mã thứ k.
- $m_k^{l_k}$ gọi là từ mã.

Các thông số cơ bản của bộ mã

- Độ dài từ mã: l_k là độ dài từ mã thứ k; $l_k = const\ \forall k$ gọi là mã đều, ngược lại gọi là mã không đều.
- Độ dài trung bình: là trung bình thống kê của độ dài các từ mã: $\bar{l} = \sum_{k=1}^{N} p(x_k) l_k$
- Cơ số mã: số các dấu (chữ mã) khác nhau được sử dụng trong bộ mã.
- Bộ mã mà tất cả các tổ hợp dấu mã là từ mã của tập tin tương ứng gọi là bộ mã đầy, ngược lại gọi là mã không đầy (mã vơi).
- Tính hiệu quả của phép mã hóa: $\eta = \frac{\bar{l}_{min}}{\bar{l}} = \frac{H(X)}{\bar{l}} \rightarrow \eta \leq 1$. Bộ mã hiệu quả khi $\eta \rightarrow 1$.
- Độ chậm giải mã: là số dấu (chữ mã) nhận được cần thiết trước khi có thể thực hiện được việc giải mã.
- Phương sai độ dài trung bình của bộ mã $\sigma_l^2 = \sum_{k=1}^{N} p(x_k)(l_k - \bar{l})^2$

Khái niệm các bộ mã (3)

Ví dụ về các bộ mã - Lược đồ Venn



Hình: Phân loại các lớp các mã (I) Mã giải mã tức thì (U) Mã có khả năng giải mã duy nhất (N) Mã không suy biến (A) Tất cả các mã

Khái niệm các bộ mã (1)

Định nghĩa (Mã không suy biến (không dị thường))

Một bộ mã được gọi là không suy biến (non-singular) nếu mọi tin x_k của nguồn X ánh xạ thành các từ mã khác nhau của bộ mã.

$$x_k \neq x_l \Rightarrow m_k^{l_k} \neq m_l^{l_l}$$

Định nghĩa (Từ mã mở rộng)

Một từ mã mở rộng là việc ánh xạ một chuỗi hữu hạn các tin thành các từ mã liên tiếp nhau.

$$x_1 x_2 \ldots \longmapsto m_1^{l_1} m_2^{l_2} \ldots$$

Nguyên tắc mã hóa tối ưu

Ví dụ

Ví dụ

Giả sử có bộ mã $\mathfrak{C} = \{0, 10, 110, 111\}$. Cho một đoạn văn bản sau: "aaaaabbbccd". Thực hiện việc mã hóa theo các phương án sau:

1. *Phương án 1 $a \leftrightarrow 111$, $b \leftrightarrow 110$, $c \leftrightarrow 10$ và $d \leftrightarrow 0$*
2. *Phương án 2 $d \leftrightarrow 111$, $c \leftrightarrow 110$, $b \leftrightarrow 10$ và $a \leftrightarrow 0$*

Tìm biểu diễn tương ứng của đoạn văn bản và so sánh các bản mã thu được.

Khái niệm các bộ mã (2)

Định nghĩa (Bộ mã có khả năng giải mã một cách duy nhất)

Một bộ mã được gọi là bộ mã có khả năng giải mã được một cách duy nhất nếu từ mã mở rộng của nó là một từ mã không suy biến.

Định nghĩa (Bộ mã có tính prefix)

Một bộ mã được gọi là bộ mã có tính prefix hay còn gọi mã có khả năng giải mã tức thời nếu không có bất cứ từ mã nào là phần mào đầu (prefix) của một từ mã khác trong bộ mã.

- Một bộ mã có khả năng giải mã được một cách duy nhất không phải là một bộ mã có tính prefix.
- Một bộ mã prefix là bộ mã có khả năng phân tách được.

Nguyên tắc mã hóa tối ưu

Bài toán mã hóa tối ưu

Bài toán mã hóa tối ưu

$$\min \bar{l} = \sum_k p(x_k) l_k$$

$$\text{sao cho} \sum_{k=1}^{N} q^{-l_k} \leq 1$$

$\Rightarrow l_k^* = -\log_q(p(x_k))$. Trường hợp tổng quát $l_k^* \notin Z^+$

Định lý

Gọi tập $l_1^*, l_2^*, ..., l_N^*$ là tập các độ dài từ mã tối ưu của phép mã hóa cơ số q cho nguồn rời rạc có phân bố p trên tập dấu mã M. Khi đó độ dài trung bình từ mã của bộ mã tối ưu $\bar{l}^*$ thỏa mãn bất đẳng thức kẹp:

$$H_q(X) \leq \bar{l}^* < H_q(X) + 1$$

Nguyên tắc mã hóa tối ưu

Nguyên tắc

Nguyên tắc

Gán các từ mã có độ dài ngắn cho các tin có xác suất xuất hiện lớn, và các từ mã có độ dài dài cho các từ mã có xác suất xuất hiện nhỏ.

Định nghĩa (Phép mã hóa tối ưu)

Một phép mã hóa được gọi là tiết kiệm (hay còn gọi là tối ưu) nếu nó đạt được độ dài trung bình từ mã cực tiểu $\bar{l}_{min}$

Nguyên tắc mã hóa tối ưu

Mã khối dữ liệu

- Dãy n ký hiệu (tin) từ nguồn rời rạc X, mỗi tin x_k được lấy với xác suất phân bố độc lập tương đồng (i.i.d) $p(x_k)$.
- Gọi $l(x_1, x_2, \ldots, x_n)$ là độ dài từ mã tương ứng với dãy $(x_1, x_2, \ldots, x_n)$.
- Định nghĩa L_n là độ dài trung bình từ mã với mỗi ký hiệu, nói cách khác:

$$L_n = \frac{1}{n} \sum p(x_1, x_2, \ldots, x_n) l(x_1, x_2, \ldots, x_n) = \frac{1}{n} E[l(X_1, X_2, \ldots, X_n)]$$

Định lý

Độ dài trung bình từ mã với mỗi ký hiệu khi thực hiện mã hóa khối đồng thời thỏa mãn bất đẳng thức

$$H(X) \leq L_n < H(X) + \frac{1}{n}$$

- $n \longrightarrow \infty \Rightarrow L_n \longrightarrow H(X)$

Nguyên tắc mã hóa tối ưu

Định lý (Bất đẳng thức Kraft)

Với bất cứ bộ mã prefix nào trên tập dấu (chữ mã) M có kích thước (cơ số) q thì tập độ dài các từ mã có thể $l_1, l_2, \ldots, l_N$ phải thỏa mãn bất đẳng thức:

$$\sum_{k=1}^{N} q^{-l_k} \leq 1$$

Ngược lại, với một tập các độ dài từ mã cho trước thỏa mãn bất đẳng thức này thì tồn tại một bộ mã prefix nhận tập độ dài này làm độ dài các từ mã.

Định lý

Độ dài trung bình từ mã L của bất cứ bộ mã có khả năng giải mã tức thì cơ số q nào biểu diễn một nguồn rời rạc X cũng lớn hơn hoặc bằng với entropy $H_q(X)$ của nguồn, nói cách khác:

$$\bar{l} \geq H_q(X)$$

xảy ra đẳng thức khi và chỉ khi $q^{-l_k} = p(x_k)$

Mã Shannon-Fano

Tổng quan

- Thuật toán đơn giản xây dựng bộ mã có tính prefix.
- Thuật toán tạo bộ mã không đều khá hiệu quả (tính toán đơn giản).
- Thuộc lớp thuật toán cận tối ưu (suboptimal).
  - Không luôn luôn tạo ra bộ mã tối ưu.
- Ít phổ biến.

Thuật toán Shannon-Fano

1. Sắp xếp các tin theo thứ tự xác suất (tần suất) từ cao đến thấp từ phía trái sang phía phải.
2. Chia dãy đó thành hai phần sao cho các phần có tổng xác suất xấp xỉ bằng nhau.
3. Gán nhãn cho phần nửa trái một bít 0, và nhóm bên phải bít 1.
4. Lặp lại các bước 3 và 4 cho mỗi nửa bằng cách chia nhóm nhỏ và gán nhãn bít cho đến tận khi các nhóm chỉ còn một nút tương ứng với lá của cây mã.
5. Từ mã thu được bằng cách duyệt từ gốc đến các nút lá tương ứng.

Nguyên tắc mã hóa tối ưu

Mã hóa với đặc trưng thống kê xấp xỉ

Định lý

Độ dài trung bình bộ mã biểu diễn một nguồn có hàm mật độ phân bố $p(x)$ với các độ dài từ mã được sử dụng $l_k = \lceil \log \frac{1}{q(x_k)} \rceil$ thỏa mãn

$$H(p) + D(p||q) \leq E[l_k]_p < H(p) + D(p||q) + 1$$

- $\rightarrow$ Nếu chúng ta sử dụng phân bố sai trong quá trình thiết kế mã, thì chúng ta phải trả giá $D(p||q)$ trong độ dài từ mã trung bình mô tả nguồn.

Mã Huffman

Tổng quan

- Thuộc lớp mã hóa Entropy, mã hóa nén dữ liệu không tổn hao (lossless data compression)
- Là lớp mã với độ dài từ mã thay đổi (variable-length code)
- Bộ mã thu được là bộ mã có tính prefix.
- Yêu cầu phân bố của nguồn phải biết trước.
- Thuộc dạng thuật toán "Greedy".
- Là thuật toán mã hóa tối ưu.

Định lý

Mã hóa Huffman là mã hóa tối ưu. Nói cách khác, gọi $\bar{l}_H$ là độ dài trung bình từ mã của bộ mã Huffman cho nguồn rời rạc X, $\bar{l}$ là độ dài trung bình từ mã của bộ mã tạo được bởi một phương pháp nào đó, khi đó chúng ta có:

$$\bar{l}_H \leq \bar{l}$$

Mã Shannon

Nguyên tắc và thuật toán

Nguyên tắc chọn độ dài từ mã

Với một tin x_k có $p(x_k)$ cho trước, mã Shannon có độ dài từ mã xác định bởi công thức:

$$l_k = \lceil \log_2 \frac{1}{p(x_k)} \rceil \quad (\forall x_k \in X)$$

Thuật toán

1. Sắp xếp các tin theo thứ tự xác suất phân bố giảm dần.
2. Chọn các từ mã có độ dài thích hợp theo thứ tự và tránh việc chọn các từ mã vi phạm tính prefix.

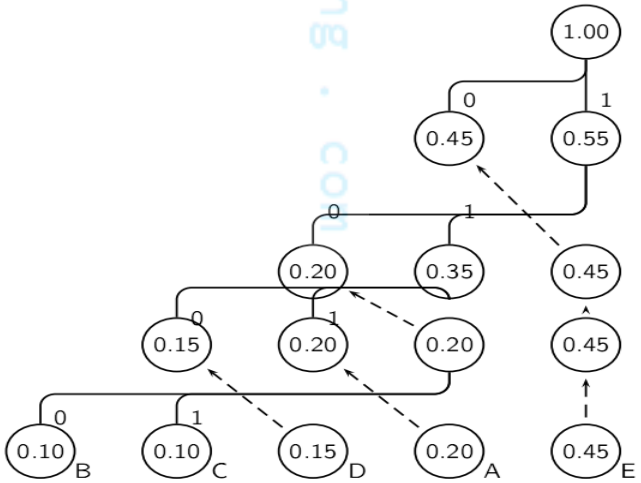
Ví dụ

Xét nguồn rời rạc X có các tin là các ký tự A,B,C,D và E có xác suất phân bố lần lượt là 0,2; 0,1; 0,1; 0,15 và 0,45. Sử dụng thuật toán mã hóa Huffman để mã hóa các tin. Tính toán độ dài trung bình của bộ mã đạt được. So sánh độ dài trung bình từ mã với H(X). Tính phương sai độ dài trung bình của bộ mã. Kiểm tra bất đẳng thức kẹp đối với độ dài trung bình từ mã.

Nhập vào: $X = \{x_k\}$ với các xác suất phân bố $p(x_k)$ tương ứng.

$$X = \{x_k\} = \begin{pmatrix} x_1 & x_2 & \dots & x_N \\ p(x_1) & p(x_2) & \dots & p(x_n) \end{pmatrix}$$

In ra: Các từ mã nhị phân $m_k^{l_k}$ tương ứng với tin x_k

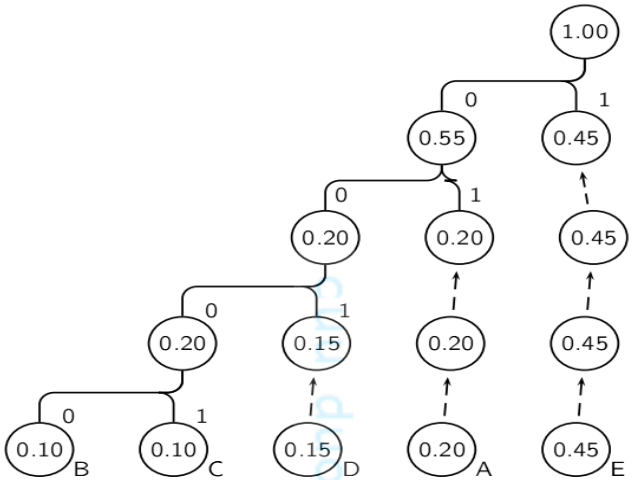


Hình: Sơ đồ cây mã Huffman theo nguyên lý Bubble

1. Khởi động danh sách cây nhị phân có một nút chứa các trọng lượng là xác suất phân bố tương ứng của các tin x_k, sắp xếp theo thứ tự tăng dần từ trái sang phải.
2. Thực hiện lặp các bước sau đến khi thu được một nút duy nhất.
   1. Tìm hai cây T' và T'' trong danh sách các nút gốc có trọng lượng tối thiểu p' và p''. Thay thế chúng bằng một cây có nút gốc có trọng bằng $p' + p''$ và các cây con là T' và T''.
   2. Đánh nhãn 0 và 1 trên các nhánh từ gốc mới đến các cây T' và T''.
   3. Sắp xếp các nút theo thứ tự tăng dần của trọng xác suất.
3. Duyệt từ gốc cuối cùng đến nút lá với các bít là các nhãn ta được từ mã tương ứng với các tin.

Mã Huffman

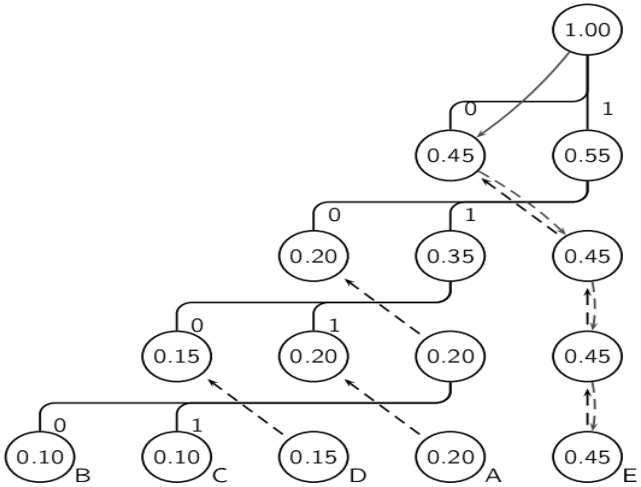
Thuật toán mã hóa - Ví dụ minh họa: Xây dựng cây bằng phương pháp khác



Hình: Một phương pháp xây dựng cây mã khác

Mã Huffman

Thuật toán mã hóa - Ví dụ minh họa: Duyệt cây



Hình: Duyệt cây mã xây dựng bộ mã

Mã Huffman

Thuật toán mã hóa - Ví dụ minh họa: Xây dựng cây bằng phương pháp khác - Kết quả

Kết quả: E(1), A(01), D(001), B(0000), C(0001)

Độ dài trung bình từ mã:

$$\begin{aligned}\bar{l} &= \sum_{k=1}^{4} p(x_k) l_k \\ &= 0.1 \times 4 + 0.1 \times 4 + 0.15 \times 3 + 0.2 \times 2 + 0.45 \times 1 = 2,1\end{aligned}$$

Entropy của nguồn: $H(X) = -\sum_{k=1}^{4} p(x_k) \log(p(x_k)) = 2,058$

$\Rightarrow H(X) \leq \bar{l} < H(X) + 1$

Tính hiệu quả của bộ mã: $\eta = \frac{H(X)}{\bar{l}} = 98\%$

Phương sai độ dài từ mã: $\sigma_l^2 = \sum_{k=1}^{N} p(x_k)(l_k - \bar{l})^2 = 1.39$

Mã Huffman

Thuật toán mã hóa- Ví dụ minh họa: Kết quả

Kết quả: E(0), A(111), D(110), B(100), C(101)

Độ dài trung bình từ mã:

$$\begin{aligned}\bar{l} &= \sum_{k=1}^{5} p(x_k) l_k \\ &= 0.1 \times 3 + 0.1 \times 3 + 0.15 \times 3 + 0.2 \times 3 + 0.45 \times 1 = 2,1\end{aligned}$$

Entropy của nguồn: $H(X) = -\sum_{k=1}^{5} p(x_k) \log(p(x_k)) = 2,058$

$\Rightarrow H(X) \leq \bar{l} < H(X) + 1$

Tính hiệu quả của bộ mã: $\eta = \frac{H(X)}{\bar{l}} = 98\%$

Phương sai độ dài từ mã: $\sigma_l^2 = \sum_{k=1}^{5} p(x_k)(l_k - \bar{l})^2 = 0.4455$

Mã Huffman

Thuật toán giải mã - Minh họa

Ví dụ

Với sơ đồ cây mã Huffman đã nhận được, giả sử nhận được chuỗi bít 010010101010101... Sử dụng thuật toán Huffman giải mã dãy tin đã phát

Kết quả: $0 : E$, $100 : B$, $101 : C$, $0 : E$, $101 : C$, $0 : E$, $101 : C$, ...

Mã Huffman

Nhận xét

- Phép mã hóa tối ưu Huffman: tập các từ mã cho bộ mã tối ưu là không duy nhất. Nói cách khác, có thể có nhiều hơn một tập các độ dài cho cùng độ dài trung bình:
  - Việc gán nhãn "0" và "1" là tùy ý.
  - Việc sắp xếp các phân bố xác suất hợp (cây thay thế) có thể thực hiện: xếp "trội" nhất, hoặc xếp "chìm" nhất
- Việc xếp xác suất phân bố "trội" nhất sẽ cho bộ mã có phương sai độ dài từ mã nhỏ nhất (gần bộ mã đều nhất)

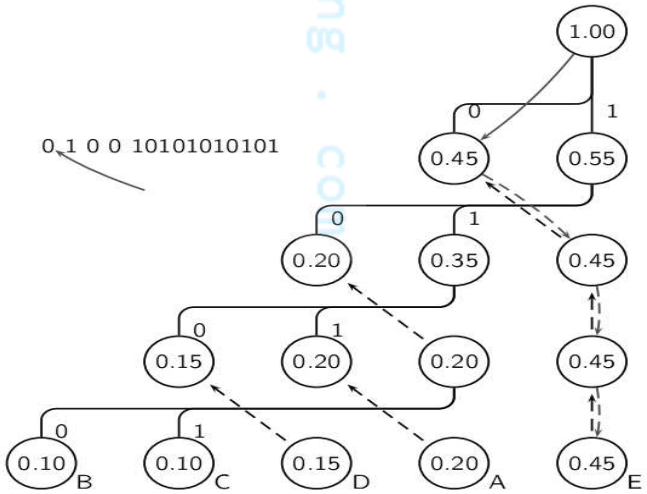
Mã Huffman thỏa mãn mã tối ưu:

1. Nếu $p(x_k) > p(x_l)$ thì $l_k < l_l$.
2. Hai từ mã có độ dài nhất có cùng độ dài.
3. Hai từ mã có độ dài nhất chỉ khác nhau một bít ở vị trí cuối cùng, và hai từ mã này tương ứng với hai tin (ký hiệu) có xác suất xuất hiện thấp nhất.

Mã Huffman cũng thỏa mãn giới hạn $\bar{l}_k \leq H(X) + 1$

Mã Huffman

Thuật toán giải mã - Minh họa: Duyệt cây mã



Hình: Minh họa quá trình giải mã

Mã Huffman

Thuật toán giải mã

Nhập vào: Chuỗi bít thông tin

In ra: Dãy tin tương ứng

1. Khởi động, đặt con trỏ P chỉ đến gốc (root) của cây mã hóa Huffman. Gán con trỏ bít b rỗng.
2. Lặp các bước sau đến khi kết thúc chuỗi bít thông tin
   1. Gán b bằng bít tiếp theo của chuỗi. Nếu $b = 0$ dịch con trỏ P theo nhánh có nhãn 0, nếu ngược lại, dịch con trỏ P theo nhánh có nhãn 1.
   2. Nếu P đã chỉ đến nút lá thì ghi ra tin tương ứng với từ mã. Khởi động lại con trỏ chỉ đến gốc

Thuật toán mã Lempel-Ziv

Tổng quan

- Thuộc lớp mã hóa không tổn hao.
- Thuộc lớp mã hóa thuật toán từ điển.
- Không yêu cầu phải biết trước phân bố của nguồn, thuật toán thích nghi.
- Ứng dụng rộng rãi trong thực tế, là cơ sở của nhiều trình tiện ích nén dữ liệu thương mại.

Mã hóa Huffman	Mã hóa LZ
Yêu cầu biết phân bố của nguồn	Không cần biết phân bố của nguồn
Bảng mã được chọn trước	Bảng mã được tạo trong quá trình
Phương thức mã độ dài cố định-thay đổi	Phương thức độ dài thay đổi-cố định

Bảng: So sánh giữa mã hóa Huffman và mã hóa LZ. © GIT

Thuật toán mã hóa Shannon-Fano-Elias

Tổng quan

1. Sử dụng hàm mật độ phân bố tích lũy để thực hiện mã hóa.
2. Định nghĩa hàm mật độ phân bố tích lũy cải tiến:

$$\bar{F}(x) = \sum_{a<x_k} p(a) + \frac{1}{2}p(x_k)$$

   - $\bar{F}(a) \neq \bar{F}(b)$ nếu $a \neq b$.
   - $\rightarrow$ có thể sử dụng $\bar{F}(x)$ như là một mã cho x_k.
3. Cắt $\bar{F}(x)$ còn l_k bít, ký hiệu là $\lfloor \bar{F} \rfloor_{l_k}$.
4. Nếu $l_k = \lceil \log_2 \frac{1}{p(x_k)} \rceil + 1$ thì:

$$\frac{1}{2^{l_k}} < \frac{p(x_k)}{2} = \bar{F}(x) - F(x-1)$$

   - $\rightarrow$ l_k bít là đủ để có thể mô tả x_k

Thuật toán mã Lempel-Ziv

Thuật toán

Thuật toán mã hóa Lempel-Ziv

1. Cho trước chuỗi $\mathcal{X} = x_1 x_2 \dots x_n$ (n rất lớn).
2. Khởi động bảng từ mã cơ bản khởi đầu.
3. Tìm kiếm trong chuỗi nguồn đã cho cụm mào đầu dài nhất có mặt trong bảng từ mã. Nói cách khác, tìm kiếm w dài nhất mà $\mathcal{X} = (w, \mathcal{X}')$.
4. Cập nhật bảng mã với từ mã mới được tạo thành từ (w, x_k), với x_k là ký hiệu tiếp theo trong chuỗi đầu vào.

Thuật toán mã hóa số học

Tổng quan

- Thuộc lớp mã hóa không đều.
- Thuộc lớp mã hóa Entropy.
- Thuộc lớp mã hóa không tổn hao.
- Được sử dụng rộng rãi trong thực tế và trong các trình tiện ích nén dữ liệu thương mại.
- Thực hiện việc mã hóa một nhóm dữ liệu.
- Là một mở rộng trực tiếp của phương pháp mã hóa Shannon-Fano-Elias.
- Ý tưởng quan trọng là tính toán và sử dụng hàm phân bố xác suất của X^n

Kết thúc phần mã hóa nguồn