Vấn đề số hóa thông tin

Bởi:

unknown

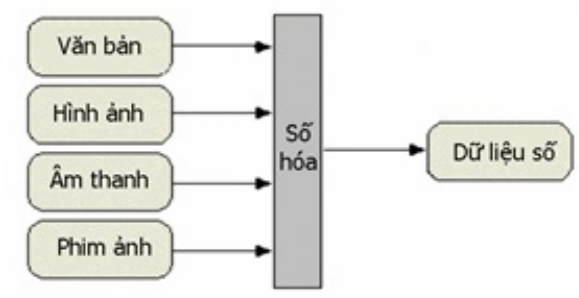
Vấn đề số hóa thông tin

Thông tin tồn tại dưới nhiều hình thức khác nhau. Để xử lý, mà đặc biệt để truyền tải thông tin ta cần phải mã hóa chúng.

Lời nói : Hệ thống : điện thoại Bộ mã hóa : micro Bộ giải mã : Loa Truyền tải : tín hiệu tuần tự hay tín hiệu số	***Ảnh tĩnh :*** Hệ thống: fax Bộ mã hóa : scanner Bộ giải mã : Bộ thông dịch tập tin Truyền tải : Tín hiệu tuần tự hoặc tín hiệu số.
Dữ liệu tin học : Hệ thống : mạng truyền tin. Bộ mã hóa : Bộ điều khiển truyền thông. Bộ giải mã:Bộ điều khiển truyền thông Truyền tải : Tín hiệu tuần tự hoặc tín hiệu số.	***Truyền hình :*** Hệ thống : truyền quảng bá Bộ mã hóa : caméra Bộ giải mã : bộ thu TV + antenne Truyền tải : Tín hiệu tuần tự hoặc tín hiệu số.

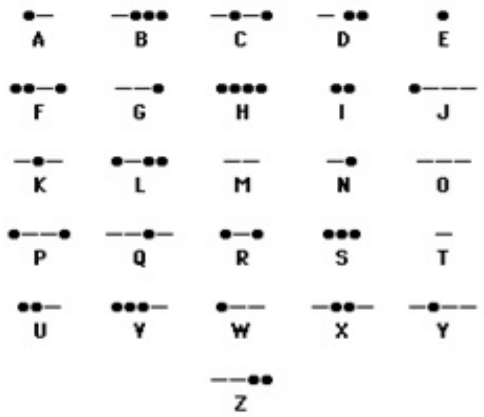
Trong thời đại chúng ta, thông tin thường được thể hiện dưới dạng các trang tài liệu hỗn hợp, như các trang web, mà ở đó đồng thời có thể thể hiện văn bản, hình ảnh tĩnh, hình ảnh động, phim ảnh,.... Thông tin thực tế được thể hiện dưới dạng đa phương tiện. Mỗi một loại thông tin sở hữu hệ thống mã hóa riêng, nhưng kết quả thì giống nhau: một chuỗi các số 0 và 1. Việc truyền tải thông tin bao gồm việc truyền tải các bit này.

Mô hình mã hóa như sau:



Sơ đồ số hóa dữ liệu

Số hóa văn bản



Mã Morse

Hệ thống mã hóa đầu tiên liên quan đến văn bản là hệ thống mã Morse, được sử dụng rộng rãi trước khi có máy tính. Đây là một bộ mã nhị phân sử dụng 2 ký tự chấm (**.**) và gạch (**-**) để số hóa văn bản (có thể xem tương đương với các bit 0 và 1).

Tuy nhiên nó có nhiều điểm bất lợi sau:

- Nghèo nàn: ít các ký tự được mã hóa;
- Nó sử dụng sự phối hợp của các dấu gạch và dấu chấm với độ dài khác nhau, điều này không được tiện lợi đặc biệt cho các ký tự có tần suất xuất hiện giống nhau.

Chính vì thế nó không được dùng để số hóa thông tin.

Nếu chúng ta qui định rằng số bit dùng để mã hóa cho một ký tự phải bằng nhau thì với p bit ta có thể mã hóa cho 2^p ký tự. Hệ thống mã hóa như thế đã được dùng trong quá khứ.

Ví dụ :

5 bit: dùng trong hệ thống ATI (Alphabet Télégraphique International)

7 bit : gọi là mã ASCII (American Standard Code for Informatics Interchange) được dùng rộng rãi trong máy tính.

poids forts

poids faibles	000	001	010	011	100	101	110	111
0000	NUL	DLE	SP	0	@	P	\	p
0001	SOH	DCI	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	,	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	'	<	L	Ç	l	ù
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	↑	n	≈
1111	SI	US	/	?	O	<--	o	DEL

code ASCII

Mã ASCII chuẩn

Bảng mã này có cả các ký tự không in được gọi là các ký tự điều khiển được dùng để tạo ra các tác vụ trên các thiết bị tin học hay dùng để điều khiển thông tin truyền tải.

Bảng mã 8 bits: có mã ASCII mở rộng và mã EBCDIC

Vì máy tính lưu thông tin dưới dạng các byte 8 bit nên khi sử dụng mã ASCII 7 bít thì bit có trọng số lớn nhất (vị trí thứ 7) luôn có giá trị là 0. Chúng ta có thể sử dụng bit này để định nghĩa các ký tự đặc biệt bằng cách đặt nó giá trị 1. Và như thế chúng ta có một bảng mã ASCII mở rộng. Tuy nhiên, điều này sẽ dẫn đến việc tồn tại nhiều bảng mã ASCII mở rộng khác nhau làm khó khăn trong việc trao đổi thông tin trên phạm vi toàn thế giới.

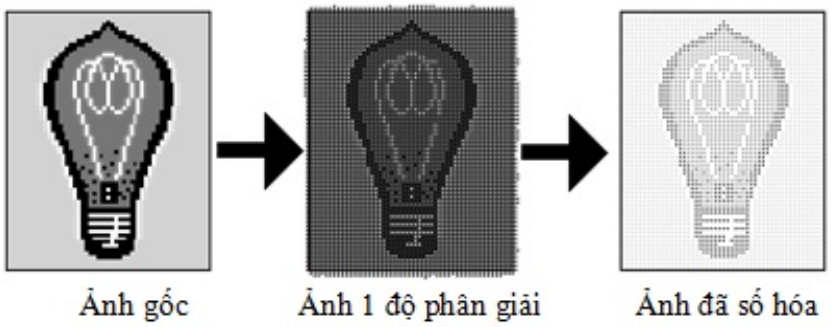
Mã EBCDIC dùng 8 bits để mã hóa nhờ đó có thể thể hiện được 256 ký tự. Nó được sử dụng trong các máy tính IBM. Tuy nhiên nó không thông dụng như mã ASCII.

Mã 16 bits : Mã Unicode

Mã này được phát triển gần đây để thỏa mãn nhu cầu trao đổi thông tin giữa những người dùng Web. Nó mã hóa hầu hết tất cả các ký tự của các ngôn ngữ trên thế giới. Nó tương thích với mã ASCII 7 bits ở 127 ký tự đầu tiên. Hiện nay mã Unicode bắt đầu được sử dụng rộng rãi.

Số hóa hình ảnh tĩnh

Ảnh số thật sự là một ảnh được vẽ nên từ các đường thẳng và mỗi đường thẳng được xây dựng bằng các điểm. Một ảnh theo chuẩn VGA với độ phân giải 640x480 có nghĩa là một ma trận gồm 480 đường ngang và mỗi đường gồm 640 điểm ảnh (pixel).



Mã hóa hình ảnh tĩnh

Một điểm ảnh được mã hóa tùy thuộc vào chất lượng của ảnh:

Ảnh đen trắng : sử dụng một bit để mã hóa một điểm : giá trị 0 cho điểm ảnh màu đen và 1 cho điểm ảnh màu trắng.

Ảnh gồm 256 mức xám: mỗi điểm được thể hiện bằng một byte (8 bits) ;

Ảnh màu: người ta chứng minh rằng một màu là sự phối hợp của ba màu cơ bản là đỏ (Red), xanh lá (Green) và xanh dương (Blue). Vì thế một màu bất kỳ có thể được biểu biễn bởi biểu thức:

$x = aR + bG + cB$

Trong đó a, b, c là các lượng của các màu cơ bản. Thông thường một ảnh đẹp sẽ có lượng màu với giá trị từ 0 đến 255. Và như thế, một ảnh màu thuộc loại này được thể hiện bằng 3 ma trận tương ứng cho 3 loại màu cơ bản. Mỗi phần tử của mảng có giá trị của 8 bits. Chính vì thế cần có 24 bit để mã hóa cho một điểm ảnh màu.

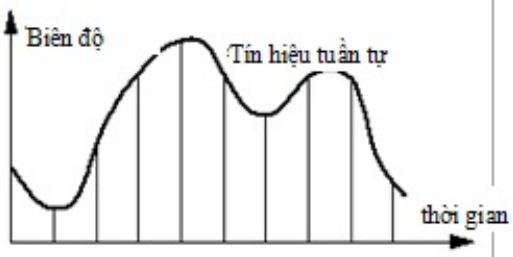
Kích thước của các ảnh màu là đáng kể, vì thế người ta cần có phương pháp mã hóa để giảm kích thước của các ảnh.

Số hóa âm thanh và phim ảnh

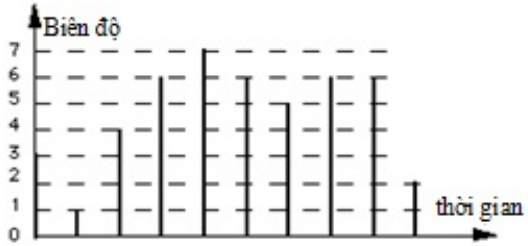
Dữ liệu kiểu âm thanh và phim ảnh thuộc kiểu tín hiệu tuần tự. Các tín hiệu tuần tự được số hóa theo cách thức sau đây:

1 - Lấy mẫu

Tín hiệu được lấy mẫu: với tần số f, ta đo biên độ của tín hiệu, như thế ta được một loạt các số đo.



H3.6 Lấy mẫu



H3.7 Lượng hóa

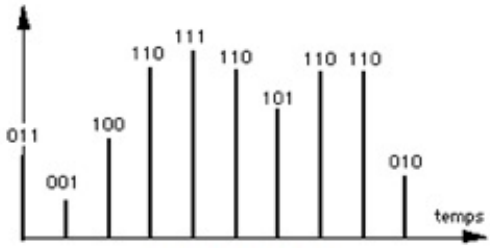
2 - Lượng hóa

Ta xác định một thang đo với các giá trị là lũy thừa của 2 (2^p) và thực hiện việc lấy tương ứng các số đo vào giá trị thanh đo.

3- Mã hóa

Mỗi một giá trị sau đó được mã hóa thành các giá trị nhị phân và đặt vào trong các tập tin.

011001100110111110101110110010.....



H3.8 Mã hóa

Dung lượng tập tin nhận được phụ thuộc hoàn toàn vào tần số lấy mẫu f và số lượng bit dùng để mã hóa giá trị thang đo p (chiều dài mã cho mỗi giá trị).