

Appunti di informatica: la codifica dell'informazione

Per codifica dell'informazione si intende la rappresentazione di un'informazione numerica, grafica, testuale o vocale.

I segnali possono essere di due tipi : analogico e digitale.

I segnali analogici sono paragonabili ad un'onda e trasmettono un ampio spettro di informazioni, segnali continui comprendenti tutte le variazioni tra il massimo ed il minimo.

Questi segnali (informazioni) spesso sono deformati rispetto agli originali in quanto le onde analogiche sono molto sensibili alle interferenze.

I segnali digitali possono assumere due soli stati significativi, ognuno dei quali denota un intervallo di valori di una grandezza osservata come una differente tensione elettrica (alta o bassa) oppure una diversa polarizzazione magnetica (positiva o negativa).

Il vantaggio che si ha con i segnali digitali consiste nel fatto che non sono soggetti ad interferenze, possono essere interpretati univocamente da un dispositivo come un computer, costituito da numerosi transistor ciascuno dei quali può assumere solo due stati: acceso o spento.

In un calcolatore l'informazione viene rappresentata mediante l'uso della tecnologia digitale, ossia dai 2 possibili valori della tensione elettrica $\{V_{high}, V_{low}\}$.

In generale, a seconda del dispositivo considerato i valori 0 e 1 possono essere rappresentati da una tensione elettrica (alta, bassa), da un differente stato di polarizzazione magnetica (positiva, negativa), da luce e buio, etc..

Ogni informazione viene trasformata nel calcolatore in una sequenza di bit, cioè in una sequenza di 0 e 1.

Codifica dei numeri

Un sistema di numerazione è caratterizzato dalla base di numerazione, cioè dal numero dei simboli differenti o cifre da esso utilizzati, e dalle regole che stabiliscono una corrispondenza biunivoca tra un numero e la sua rappresentazione espressa come combinazione delle cifre.

Il nostro sistema, introdotto in Europa dagli arabi nel Medio Evo, è decimale e posizionale:

- decimale (o in base 10): ogni numero anche molto grande può essere scritto usando dieci simboli cioè le cifre che vanno da 0 a 9;

- posizionale: il valore di ogni cifra dipende dalla posizione che occupa il numero;
- dieci unità di un ordine formano una unità dell'ordine immediatamente superiore.

I sistemi di numerazione posizionale sono caratterizzati da una base b e un alfabeto α .

La base (b) è il numero degli elementi che compongono l'alfabeto, ad esempio, nel caso decimale, $b=10$.

L'Alfabeto (α) è l'insieme delle cifre disponibili per esprimere i numeri.

A ogni cifra corrisponde un valore compreso tra 0 e ($b-1$), ad esempio, nella numerazione decimale l'alfabeto è $\alpha=\{0,1,2,3,4,5,6,7,8,9\}$.

Esempi:

Numerazione ottale: $b = 8 \rightarrow \alpha=\{0,1,2,3,4,5,6,7\}$

Numerazione esadecimale: $b = 16 \rightarrow \alpha=\{0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F\}$

$$35475 = 3 \cdot 10000 + 5 \cdot 1000 + 4 \cdot 100 + 7 \cdot 10 + 5 \cdot 1$$

$$35475 = 3 \cdot 10^4 + 5 \cdot 10^3 + 4 \cdot 10^2 + 7 \cdot 10^1 + 5 \cdot 10^0$$

Nei sistemi additivi ogni unità è rappresentata da un unico simbolo o non-posizionale come quello romano, e ogni simbolo mantiene sempre lo stesso valore indipendentemente dalla posizione, ad esempio $16 = XVI$.

Un codice è un sistema di simboli per rappresentare un'informazione di qualsiasi genere (parole, numeri, caratteri, etc.).

Il codice binario, è un codice che utilizza come simboli le cifre binarie 0 ed 1.

Con n cifre binarie è possibile codificare 2^n oggetti distinti, esistono infatti 2^n configurazioni distinte che possono essere poste in corrispondenza biunivoca con i naturali da 0 a $2^n - 1$.

Definire un codice significa attribuire un significato unico e convenzionale alle configurazioni binarie.

Ad esempio, la codifica di quattro colori (rosso, giallo, blu, bianco):

Oggetto	Codifica
bianco	00
rosso	01
giallo	10
blu	11

Nel caso dei sistemi di calcolo è stata introdotta la rappresentazione digitale, il bit (binary digit – cifra binaria): 0 o 1.

Per poter rappresentare un numero maggiore di informazioni si usano sequenze di bit: 00 10 01 11.

Nella rappresentazione digitale, il processo secondo cui si fa corrispondere ad una informazione una configurazione di bit prende il nome di codifica dell'informazione.

Codifica dei caratteri

I caratteri di un testo vengono codificati tramite sequenze di bit, utilizzando un codice di traduzione.

Quello più usato è il codice ASCII (si pronuncia Aschii) (American Standard Code for Information Interchange).

Utilizza 7 bit e rappresenta 128 caratteri. Mancano, ad esempio i caratteri accentati, greci etc.

Esempio:

0 è codificato come 048_{10} e 30_{16} (e $011\ 0000_2$)

9 è codificato come 057_{10} e 39_{16}

a è codificato come 097_{10} e 61_{16}

z è codificato come 122_{10} e $7A_{16}$

Vengono inseriti in un byte per cui 1 bit viene normalmente ignorato.

In trasmissione funziona come bit di parità per individuare errori.

Il valore (0 o 1) del bit di parità è scelto in modo che la sequenza di 1 sia pari.

Ad esempio: $0 \Rightarrow 30_{16} \Rightarrow 0\ 011\ 0000_2$

Il codice ASCII esteso utilizza invece 8 bit (256 caratteri).

Per codificare ogni simbolo dell'alfabeto anglosassone sono sufficienti 7 bit usando l'ASCII standard, 8 bit con l'ASCII esteso, 16 bit con il codice UNICODE. Microsoft Windows usa un codice proprietario a 16 bit simile ad UNICODE.

Nella codifica dell'informazione con 2 bit si codificano 4 informazioni (2^2), con 3 bit si codificano 8 informazioni (2^3), con N bit si codificano (2^N) informazioni.

L'insieme dei bit utilizzati per rappresentare un carattere è detto byte, ed è la minima unità di rappresentazione dell'informazione.

Il numero dei bit che costituiscono un byte può variare in dipendenza del codice, ma in genere è formato da 8 bit ($2^8 = 256$ caratteri)

Le parole sono sequenze di caratteri, quindi volendo rappresentare, per esempio, la parola informatica secondo la codifica binaria:

i	n	f	o	r	m	a	t	i	c	a
01101001	01101110	01100110	01101111	01110010	01101101	01100001	01110100	01101001	01100011	01100001

Nella seguente tabella le codifiche binario, ottale e esadecimale dei numeri da 0 a 15

Decimale	Binario	Ottale	Esadecimale
0	0 0 0 0	00	0
1	0 0 0 1	01	1
2	0 0 1 0	02	2
3	0 0 1 1	03	3
4	0 1 0 0	04	4
5	0 1 0 1	05	5
6	0 1 1 0	06	6
7	0 1 1 1	07	7
8	1 0 0 0	10	8
9	1 0 0 1	11	9
10	1 0 1 0	12	A
11	1 0 1 1	13	B
12	1 1 0 0	14	C
13	1 1 0 1	15	D
14	1 1 1 0	16	E
15	1 1 1 1	17	F

Consideriamo l'insieme dei numeri relativi (interi) $Z = \{ -\infty, \dots, -1, 0, +1, \dots, +\infty \}$

Se vogliamo rappresentare gli elementi di questo insieme in forma binaria, avendo sempre un numero limitato di bit, possiamo utilizzarne uno per la rappresentazione del segno (+ o -).

Il primo bit di un numero intero viene utilizzato come bit di segno, 0 positivo 1 negativo.
Gli altri bit indicano il modulo (valore assoluto) del numero e vengono convertiti secondo la notazione posizionale:

1100 -> $\boxed{1}$ 100 -> $\boxed{-}$ 4

Ad esempio con parole di 5 bit:

+5 = 00101

-10 = 11010

con parole a 16 bit:

+13 = 0000000000001101

-13 = 1000000000001101

Codifica binaria dei caratteri

Codificazione delle parole dall'alfabeto esterno all'alfabeto interno dell'elaboratore.

Sappiamo che l'alfabeto esterno è formato da lettere dell'alfabeto, 10 cifre decimali, 28 caratteri speciali, e che un messaggio digitale necessita di un codice per essere interpretato.

Un codice è un insieme di caratteri alfanumerici.

In informatica si usano vari sistemi di codifica dei caratteri, e in tutti i codici ad ogni lettera è associata una configurazione di bit:

- Il codice ASCII (American Standard Code for Information Interchange), sviluppato dall'ANSI (American National Standard Institute) è il più diffuso. Esso è basato su successioni di 7 bit $2^7=128$ caratteri, numeri, codici di controllo, segni di interpunzione, etc..
- Il codice ASCII esteso utilizza tutti gli otto bit disponibili in un byte ($2^8=256$).
- Il Codice UNICODE è basato su 16 bit e quindi in grado di rappresentare $2^{16}=65.536$ caratteri.

Ad esempio, nel codice ASCII esteso le lettere A, B, C sono rappresentate dalle seguenti configurazioni:

A -> 0100 0001 corrispondente al decimale 65

B -> 0100 0010 corrispondente al decimale 66

C ->0100 0011 corrispondente al decimale 67

La rappresentazione delle lettere rispetta l'ordinamento alfabetico facendo corrispondere a lettere in sequenza alfabetica crescente una successione ordinata di numeri binari crescenti.

La codifica delle immagini

Le immagini vengono codificate in due formati: bitmap e vettoriale.

Nel formato bitmap (mappa di pixel) vengono memorizzate come insieme di pixel, ognuno con le sue caratteristiche, relativamente a colore, luminosità, trasparenza, raggruppate in una matrice di pixel.

Le immagini vengono divise in punti, detti pixel, ciascun punto viene codificato con un numero che corrisponde a un particolare colore.

Esse vengono codificate come sequenze molto lunghe di bit; per interpretare le quali è necessario conoscere le dimensioni dell'immagine, ovvero il numero di pixel della base e il numero di pixel dell'altezza, la risoluzione (che si misura in dpi, dot per inch, cioè punti per pollice quadrato) e il numero di colori.

Nel formato vettoriale vengono archiviate le informazioni relative alla descrizione geometrica e a una serie di caratteristiche di apparenza quali dimensione della penna, colore, tipo di linea, etc..

L'immagine vettoriale ha il pregio di poter essere raffigurata a qualsiasi risoluzione. La descrizione vettoriale si adatta però a figure geometriche o immagini ottenute per mezzo di forme geometriche elementari.

Se volessimo calcolare lo spazio occupato su disco da una fotografia a colori salvata in modalità RGB, sapendo che questa modalità richiede l'uso di tre byte per memorizzare le informazioni di colore e luminosità relative a ciascun pixel dell'immagine, ci basta moltiplicare per tre il numero complessivo di punti da cui è costituita l'immagine, il quale

numero si ottiene moltiplicando il numero delle colonne per il numero delle righe che compongono il rettangolo occupato dall'immagine.

Ammettendo che sia una foto di 1024 pixel orizzontali per 768 verticali, lo spazio che essa occupa sarà di $1024 \times 768 \times 3 = 2.359.296$ byte, ossia 2.304 Kilobyte o, ancora, 2,25 Megabyte. Il risultato ottenuto rappresenta lo spazio occupato dal file grafico in forma non compressa.

Come per i caratteri si è cercato di stabilire alcuni standard per la codifica delle immagini:

- GIF è un formato proprietario permette di archiviare in formato compresso immagini con colori a 8 bit per pixel, ossia 256 colori, permette la codifica di immagini animate, usa un algoritmo di compressione lossless ossia senza perdita di informazioni
- BMP formato per sistemi operativi Microsoft e OS/2 utilizzato ormai in tutti i sistemi operativi permette di archiviare con diversa profondità di colore: 1, 4, 8, 24 bit;
- PNG dovrebbe sostituire il formato GIF ma non ammette l'animazione;
- TIFF nato per semplificare lo scambio di immagini tra scanner e programmi applicativi è applicato in diversi ambiti professionali è senza perdita di informazioni.
- JPG formato di archiviazione con perdita di informazione, può comprimere con un rapporto elevatissimo, viene visualizzata in RGB a 24 bit per pixel.

I file audio

Il suono è un segnale continuo, per essere memorizzato deve essere campionato ottenendo così un segnale digitale.

I parametri che caratterizzano il campionamento sono:

1. Il numero di canali:

la modalità mono ha uno solo canale mentre quella Stereo ne ha due separati, sinistro e destro. Ovviamente un segnale stereo occupa il doppio di uno mono.

Nelle applicazioni più recenti il numero di canali è notevolmente aumentato (surround).

2. La risoluzione:

rappresenta il numero di bit utilizzati per rappresentare i campioni. Solitamente si utilizzano 8 o 16 bit per campione, nel primo caso si hanno 256 valori possibili, relativamente pochi perché offrono una qualità del suono inferiore a quella di un nastro, nel secondo si hanno circa 65.000 valori.

3. La frequenza di campionamento:

è il numero di campioni al secondo, può variare da 11 khz adatta alla registrazione della voce, a 22 khz adatta alla registrazione di un nastro fino a 44 khz per una registrazione a qualità cd. Questo parametro deve essere scelto in modo da poter ricostruire il segnale originario da quello campionato.

I formati dei file audio sono:

- WAVE (formato non compresso)
- MP3 (formato compresso)
- MIDI (formato vettoriale)
- Streaming audio RAM, RM, ASF, ASX

Lo streaming è il trasferimento in rete dei dati audiovisivi senza tempi d'attesa dovuti al download completo del file sull'Hard Disk del computer.

Con lo streaming non viene scaricato l'intero file audio prima di consentirne l'ascolto, ma la riproduzione inizia per ogni blocco di due secondi d'ascolto; nel frattempo viene scaricato il successivo.

Si possono verificare momentanee interruzioni nella riproduzione, nel caso in cui il traffico nella rete risulti congestionato.

Le due principali tecnologie d'audio streaming utilizzate sono Real (attraverso il real player), e Windows Media (Windows Media player).

La tecnologia dello streaming audio ha permesso alle emittenti audio e video di trasmettere i loro programmi sui propri siti web consentendone la fruizione sul PC connesso a Internet.