

# Compressione e validazione dei dati

## Lezione 7 di Fondamenti di informatica

Docente: Giuseppe Scollo

Università di Catania  
Facoltà di Scienze Matematiche, Fisiche e Naturali  
Corso di Laurea in Informatica, I livello, AA 2009-10

### Indice

1. Compressione e validazione dei dati
2. motivazioni per la compressione dei dati
3. tecniche generiche di compressione
4. compressione di immagini e suoni
5. codici per il controllo di errore
6. codici per la correzione di errore
7. temi per ulteriori approfondimenti

## **motivazioni per la compressione dei dati**

### **compressione dei dati:**

rappresentazione dell'informazione con sequenze più brevi (si spera ;)

risparmio di **spazio** nella memorizzazione dei dati

risparmio di **tempo** nella loro trasmissione

**costa** (spazio e tempo), tuttavia è spesso **vantaggiosa**

la compressione dei dati richiede algoritmi di **(de)codifica:**

$$d(c(T)) = T$$

**rapporto di compressione** di una codifica  $c$  per una data sequenza binaria  $T$ :

è il rapporto fra la lunghezza di  $T$  e quella di  $c(T)$

## **tecniche generiche di compressione**

**"generiche"**: tecniche di compressione applicabili a qualsiasi sequenza binaria  
tuttavia, con efficacia variabile

**codifica run-length**

codifica il **numero di occorrenze consecutive** del simbolo

**codifica differenziale, o relativa**

decompone la sequenza in **blocchi** e codifica le **differenze** del blocco  
relativamente al precedente

**codifica dipendente dalla frequenza** dei simboli

**esempi:** codice Morse, codici di Huffman

**codifica basata su dizionario, adattativa**

si codificano **riferimenti** (numerici) a un dizionario predefinito (ad es.,  
quello per il controllo ortografico in un elaboratore di testi) o, in quella  
adattativa, dinamicamente costruito sulla stessa sequenza da codificare  
**esempi:** codifiche di Lempel-Ziv (ad es. LZW: Lempel, Ziv & Welsh)

## compressione di immagini e suoni

sono largamente in uso tecniche di compressione **specifiche**  
cioè, specificamente progettate (e standardizzate) per dati multimediali

tecniche di compressione **audio**:

**MP3**: (v. appresso)

**OGG/VORBIS**: più efficace di MP3, a parità di qualità

**Speex**: molto efficace per la compressione del parlato

standard di compressione di **immagini**:

**GIF**: *Graphical Interchange Format*

**PNG**: *Portable Network Graphics*

**JPEG**: *Joint Photographic Experts Group*

compressione **A/V**:

**MPEG**: *Motion Picture Experts Group*

compressione video: ~JPEG + codifica relativa

compressione audio: MP3 = *MPEG-1 Layer 3*

## codici per il controllo di errore

**problema**:

nella comunicazione attraverso un mezzo trasmissivo, una sequenza binaria è suscettibile di **alterazioni** accidentali che possono corrompere l'informazione rappresentata

**soluzione**:

estendere la sequenza con **informazione di controllo** di errore  
efficace per la **rivelazione** di errori (in certi limiti),  
e magari anche per la loro **correzione** (in limiti più ristretti)

**esempi** (in ordine di crescente complessità algoritmica):

**bit di parità** aggiunto alla codifica ASCII dei caratteri

**byte di controllo**, per sequenze più lunghe di un byte

**somma di controllo** (ingl.: *checksum*)

**codici a ridondanza ciclica** (ingl.: *Cyclic Redundancy Code (CRC)*)

## codici per la correzione di errore

per configurazioni binarie di uguale lunghezza, **distanza di Hamming** :

di due configurazioni binarie: numero di bit in cui differiscono

di un insieme finito di configurazioni binarie:

minima distanza di Hamming fra due configurazioni nell'insieme

di un codice  $c : A \rightarrow 2^n$ : distanza di Hamming della sua immagine  $c(A)$

un codice con distanza di Hamming  $2n+1$  permette di rivelare fino a  $2n$  errori nella trasmissione della codifica di un simbolo, e di correggere fino a  $n$  errori

**esempio:** il codice appresso a sinistra ha distanza di Hamming 3

A	000000
B	001111
C	010011
D	011100
E	100110
F	101001
G	110101
H	111010

l'errore di un bit nella trasmissione della codifica del simbolo **D** provoca la ricezione della configurazione **010100**, estranea al codice

la correzione rimpiazza la configurazione ricevuta con quella valida a distanza di Hamming minima da essa

A	2
B	4
C	3
D	1
E	3
F	5
G	2
H	4

## temi per ulteriori approfondimenti

### 1. Tecniche generiche di compressione adattativa

L'algoritmo LZW di compressione generica, pubblicato nel 1984, è un raffinamento di algoritmi proposti in precedenza: è adottato in alcuni programmi di utilità, ad es. *compress* in sistemi Unix di vecchia data, ma nel corso del tempo questi hanno ceduto il posto ad altre tecniche generiche, quali ad es. *gzip*, *bzip2*, *rzip*, che vantano rapporti di compressione mediamente migliori. Meritano approfondimento gli algoritmi di compressione impiegati in tali tecniche e un'analisi comparativa dei relativi pregi e limiti.

### 2. Tecniche di compressione *lossy*

Si indicano con questo termine le tecniche caratterizzate dalla possibilità di perdita di informazione attraverso la codifica, ragion per cui la decodifica non è esattamente l'inversa della codifica, ma ne è solo un'approssimazione. Per la compressione di dati multimediali tale perdita, se contenuta entro certi limiti, è spesso accettabile, ad es. quando genera differenze impercettibili dai sensi umani interessati (visione, udito). Alcune tecniche *lossy* sono impiegate negli standard di cui si son dati cenni a lezione, altre sono tuttora oggetto di ricerca, ad es. le tecniche di compressione **frattale** o quelle basate su **ondelette** (ingl. *wavelet*). Si possono approfondire le idee alla base di tali tecniche, i relativi algoritmi e i campi d'impiego in cui risultano meglio applicabili.