INFORMAZIONE E COMPLESSITA'

Antonio Restivo
Università degli Studi di Palermo

Lezioni Lincee di Scienze Informatiche
Palermo 26 Febbraio 2015

Concetti fondamentali delle Scienze Informatiche

Le Scienze Informatiche si basano sullo sviluppo di alcuni concetti fondamentali , come algoritmo, informazione, complessità, etc.

Tale sviluppo è motivato da (e rende possibili) diverse applicazioni tecnologiche.

Svolge anche un ruolo trasversale, intervenendo in altre discipline scientifiche, come Fisica, Biologia, Linguistica, etc.

L ‘Informazione

Partiamo da alcune domande:

Che cosa è l’ Informazione?

Si può misurare?

A cosa serve misurare l’informazione ?

L'Informazione come "sorpresa"

Se la probabilità di un evento x è p , l'informazione I(x) associata al verificarsi dell'evento è:

$I(x) = \log_2(1/p)$

Perché il logaritmo?

Perché le probabilità di due eventi indipendenti si moltiplicano, mentre le loro informazioni si sommano.

La scelta della base 2 per il logaritmo corrisponde a individuare una unità di misura: il bit

Sorgente di Informazione

Una sorgente di informazione S è individuata da:

A	B	C	D
p_1	p_2	p_3	p_4
1/2	1/4	1/8	1/8

Entropia di S:

$$H(S) = p_1 \log (1/p_1) + p_2 \log(1/p_2) + p_3 \log(1/p_3) + p_4 \log(1/p_4) =$$

$$1/2 \log 2 + 1/4 \log 4 + 1/8 \log 8 + 1/8 \log 8 = 7/4 = 1.75$$

H(S) è l'informazione che in media produce la sorgente S

Informazione contenuta in un testo

Supponiamo che S produca una successione di simboli, cioè un "testo" T. Ad esempio:

T = ABACABADABABADABAC (lunghezza 16)

Poiché l'informazione associata ad ogni simbolo è 7/4, il contenuto d'informazione del testo T è

7/4 x 16 = 28 bits

Informazione e Compressione Dati

Claude Shannon

A Mathematical Theory of Communication

Bell System Technical Journal (1948)

Claude Shannon

Compressione Dati

Devo inviare il testo T = ABACABADABADABAC

utilizzando un canale binario (con simboli 0,1).

Usando la codifica "standard":

A $\rightarrow$ 01

B $\rightarrow$ 10

C $\rightarrow$ 00

D $\rightarrow$ 11

Il testo T viene codificato con la sequenza

T $\rightarrow$ 01100100011001110110011101100100 (32 bits)

Compressione Dati

Se invece si usa la codifica:

A $\rightarrow$ 0

B $\rightarrow$ 10

C $\rightarrow$ 110

D $\rightarrow$ 111

Il testo T viene codificato con la sequenza

T $\rightarrow$ 0100110010011101001110100110 (28 bits)

Si è ottenuta una compressione rispetto alla codifica "standard", che era di 32 bits.

Si può fare meglio ? No!

Teorema (Shannon)

Per codificare un testo T di lunghezza n prodotto da una sorgente S di entropia H(S) non si possono usare meno di $n \cdot H(S)$ bits.

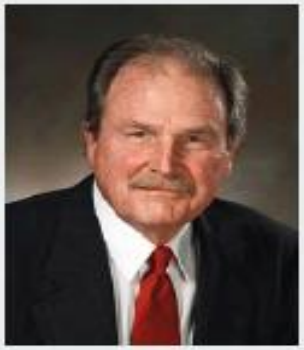
Un testo si può comprimere fino a una lunghezza che corrisponde al suo contenuto di informazione

Pertanto il codice dato nella slide precedente è ottimale.

Come costruire un codice ottimale ?

David Huffman (1952)

A Method for the Construction of Minimum-Redundancy Codes


David Huffman

L'algoritmo di Huffman produce un codice ottimale

L'algoritmo di Huffman è anche efficiente.

Limiti della Teoria di Shannon

Da un punto di vista concettuale:

L'informazione viene definita in base al fatto che un oggetto è stato selezionato in un insieme di possibili oggetti: non ha senso parlare di informazione di un singolo oggetto.

Dal punto di vista della compressione dati:

Utilizza solo proprietà statistiche di un testo.

L' Informazione come "Complessità"

Consideriamo i due testi

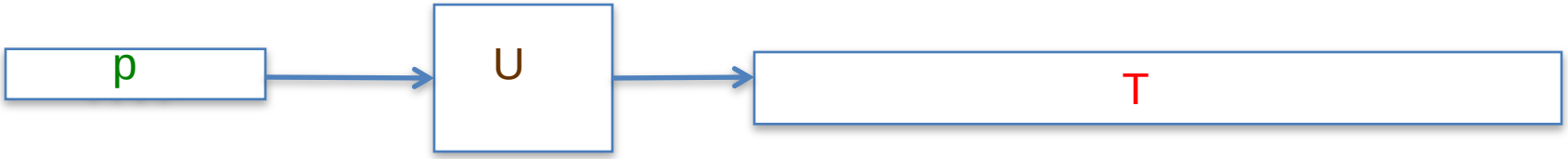
T_1 = 0001000100010001000100010001000100010001…

T_2 = 1110010111001101101001101110001111010010…

T_1 segue una "regola" "semplice"

T_2 segue una "regola" "complicata"

Alan Mathison Turing (1937)
Andrej Nikolaevic Kolmogorov (1965)



$U(p) = T$

Un programma p per T_1:

```
For n = 1 to 10000
        Print "0001"
```

Complessità di Kolmogorov:

$$K_U(T) = \min \{ |p| : U(p) = T \}$$

L’ Informazione come “complessità”

Dal “programma” p si può ricostruire il testo T.

Quindi p può essere interpretato come una versione compressa del testo T.

K(T) è la lunghezza minima del testo compresso, nel senso che qualunque compressore, applicato ad un testo T, produce un testo compresso di lunghezza $\geq$ K(T).

K(T) può quindi essere interpretata come una misura dell ‘informazione contenuta nel testo T.

Si tratta di una misura “assoluta”.

Informazione e Compressione

Compressore ideale:

Algoritmo che, per ogni testo T, produce un programma di lunghezza minima per T.

Teorema Il compressore ideale non esiste.

E' aperta la competizione per il miglior compressore!

In Compressione Dati non si può dire la parola definitiva

Informazione e Compressione

Osservazione:

Nell'approccio di Kolmogorov si parla di lunghezza minima del programma, ma non si parla della sua efficienza, cioè dei suoi tempi di esecuzione.

Compressione e Casualità

K(T) è la taglia di una *descrizione minima* di un testo T.

Definizione di testo casuale: T è *casuale* se K(T) = |T|.

Casuale = Incomprimibile

Teorema. Non si può *decidere* se un testo è casuale.

La teoria di Kolmogorov crea un legame tra la nozione di *Probabilità* e quella di *Algoritmo.*

Compressione e Induzione

La teoria di Kolmogorov mostra che esiste una stretta connessione tra la compressione e il "ragionamento induttivo"

Induzione: *processo di inferenza di una legge o principio generale da osservazioni particolari.*

Scienze naturali

Apprendimento di una lingua

Sequitur (Nevill-Manning, Witten, 1997)

Algoritmo che inferisce una struttura gerarchica (una *grammatica*) da una successione di simboli.

Compressione basata su Dizionari

La comunicazione avviene all'interno di un gruppo di persone che usa un vocabolario di 1000 parole.

Le parole vengono ordinate assegnando ad ogni parola un numero compreso tra 0 e 999.

precipitevolissimevolmente $\rightarrow$ 731

Osservazione: il mittente e il ricevente devono condividere lo stesso dizionario.

Compressione basata su Dizionari

Dizionari adattati al testo: il compressore crea allo stesso tempo il dizionario e la codifica del testo.

Abraham Lempel, Jacob Ziv

LZ 76 *On the Complexity of Finite Sequences*

LZ77 *A Universal Algorithm for Sequential data Compression*

LZ78 *Compression of Individual Sequences Via Variable-Rate Coding*

Molti programmi di compressione si basano su questi algoritmi: *gzip, zip, 7zip, pkzip, arj, rar, …..*

Di quale Complessità abbiamo parlato?

La complessità di Kolmogorov, detta anche complessità descrittiva, è una complessità statica.

Esiste, ed ha una grande importanza, anche una complessità dinamica, la complessità computazionale, che misura il tempo e la memoria utilizzate durante il calcolo.

Ma questa è un'altra storia …. o, meglio, un'altra lezione.