La candela

La 38-ma puntata, che costituisce il mio record assoluto di lunghezza, nel formato TEX che uso di solito è lunga 40187 byte. Compressa con ZIP si riduce a 16179 byte: molto meno di metà. Com'è possibile questo miracolo? E non bisogna dimenticare che si tratta di una compressione *reversibile*: dal file "zippato" si può ricostruire fino all'ultimo bit quello originario...

L'osservazione che la compressione è reversibile andava fatta, perché oggi sono in uso corrente metodi di compressione assai più potenti, ma al prezzo di perdere informazione, ossia di non poter ricostruire in tutti i dettagli il messaggio originario. Esempi: il metodo JPEG per le immagini (che genera files con estensione .JPG); oppure la compressione MP3 per la musica, quella MPEG per film e televisione... Ma di questi non possiamo occuparci: ci porterebbero troppo lontano.

Torniamo dunque a ZIP. La risposta alla domanda che ho posto sta in una parola: *ridondanza*. Il concetto in termini qualitativi fa parte dell'esperienza comune: capita a volte di avere fra le mani uno scritto dove — o per la cattiva stampa, o a causa dell'invecchiamento, o per altri motivi — alcune lettere o addirittura intere parole riescono illeggibili. Eppure, se il danno non è troppo esteso, il testo può essere ricostruito con sicurezza. A titolo di esempio, provate a leggere quanto segue:

L'ultim v*lta ho ver*men** es*ger*to con la lung***za dei mi** spr*l*q**... Potr* c*rc*r* dell* att*n**nti, ma pref*ris** rim*tt**mi al** cl*m*nz* del** c*rt*, e m*strar* il mio ravved****to op**oso: in alt** p*rol*, ***sta vol** s*rò m*lt* p** brev*.*

Penso che non avrete avuto difficoltà a riconoscere la frase iniziale della scorsa puntata; eppure sul totale di 208 lettere, ne ho sostituite ben 62 con degli asterischi! E forse avrei potuto anche essere più drastico.

La stessa cosa succede anche quando si parla: è spesso possibile intendere ciò che dice un'altra persona anche se parti di parole o intere parole vanno perdute, per es. a causa di rumori. Tutto questo dipende dal fatto che molto di ciò che si dice o si scrive è *prevedibile dal contesto*: certe combinazioni di lettere sono pressoché uniche (come la "u" dopo una "q") o sono ricavabili facilmente dal resto (come ad es. la desinenza "-mente" di tanti avverbi). A volte interi gruppi di parole possono essere indovinati, perché fanno parte di modi di dire comuni (come la "clemenza della corte" nell'esempio che ho dato sopra).

Da un punto di vista meno vago, e più formalizzabile, si può far rientrare in primo luogo sotto la specie della ridondanza il fatto che le frequenze delle lettere non sono tutte uguali. Abbiamo già visto nella puntata precedente che questo

riduce l'informazione, e ho mostrato in un esempio particolare come si possa sfruttare la conoscenza delle diverse frequenze per condensare un messaggio.

Ma si può fare di più, e già Shannon ne parla nel lavoro che ho citato l'altra volta: si possono usare le *frequenze condizionate*. Con questa espressione s'intende il fatto che la frequenza con cui si presenta una lettera dipende dalla lettera precedente. Il caso estremo è quello della combinazione "qu": in italiano esiste (rarissima) anche la coppia "qq," ma dopo una "q" nessun'altra lettera è possibile. Ci sono poi moltissimi altri casi, meno evidenti ma pure importanti: per es. dopo una "b" solo poche consonanti sono ammesse; in generale dopo una consonante è relativamente raro che capiti uno spazio (segno di fine parola)...

Si può anche procedere oltre, ossia considerare non solo le frequenze condizionate dei "digrammi," ma quelle dei "trigrammi." Esempi: dopo "zi" è molto probabile una "o"; dopo "tr" ci sarà certamente una vocale, ecc.

Ma è inutile procedere a tentoni con singoli esempi: quello che si può e si deve fare è costruire una tabella delle frequenze condizionate (per digrammi, trigrammi, e magari anche oltre); cosa che chiaramente qui non è il caso di fare. Ci basta sapere che è stata fatta, per parecchie lingue. Dopo di che, si tratta di mostrare:

a) come la conoscenza di queste frequenze riduce la quantità d'informazione (la riduce di certo, stante il criterio che ogni prevedibilità va a scapito del contenuto d'informazione)

b) come si può sfruttare tale fatto per "comprimere" il messaggio, ossia per trasmetterlo impiegando meno caratteri.

Sempre nel lavoro di Shannon ci sono entrambe le cose; ma la seconda è di gran lunga più importante, e si esprime in un teorema, che vi enuncio un po' alla buona come segue. *Se un messaggio di lunghezza L ha un contenuto d'informazione di n bit per carattere, si può codificarlo senza perdita d'informazione con un numero L' di caratteri binari, purché $L' \geq nL$.*

Per essere più chiaro faccio un esempio specifico. Si stima che un messaggio in lingua inglese abbia un contenuto effettivo di 1.7 bit per lettera. Stando così le cose, invece di 27 simboli diversi, come avevamo fatto la volta scorsa, possiamo usarne soltanto 4 (che darebbero 2 bit, più del necessario). È ovvio che questo non si può fare con una codifica lettera per lettera, visto che le lettere sono comunque 27 (spazio incluso); bisogna invece procedere per *gruppi di lettere*, e il modo preciso di farlo è troppo complicato per spiegarlo qui. Ma il teorema di Shannon non solo ci assicura che si può fare, ma ci dice anche come: nel gergo dei matematici, è un teorema *costruttivo*.

Il fatto che si possa fare non significa però che sia una cosa semplice e pratica; di conseguenza sono stati cercati dei compromessi fra la compressione ottima teorica e la nessuna compressione del messaggio grezzo. Proprio questo fanno i programmi di compressione, di cui il più famoso è appunto ZIP: un compromesso

abbastanza efficiente quanto a bontà della compressione, ma al tempo stesso abbastanza semplice da lavorare rapidamente.

Per quei moltissimi che conoscono solo Windows: avete ZIP a portata di mouse nella versione denominata WINZIP, e potete provare a verificare quanto comprime un file e quanto tempo impiega. Dopo di che, fate un'altra prova: prendete un file generato dal vostro amico inseparabile, ossia WORD, e provate a comprimerlo: è piuttosto probabile che ne risulti una compressione enorme. Spiegazione: i file .DOC generalmente sono assai ridondanti, ossia sprecano un sacco di spazio. Si capisce che questo i venditori non ve lo dicono: hanno tutto l'interesse a farvi sprecare spazio, così comprate una macchina con più memoria e dischi fissi più capaci, che poi verranno a loro volta riempiti (a vostra insaputa) di spazzatura, e via consumando...

Tanto che sono nel discorso, spiego un altro grave difetto di WORD: non di rado in un file si trova (per chi sa guardarci ...) materiale che l'autore non intendeva mettere in circolazione. Per esempio, brani di lettere scritte ad altri, elenchi d'indirizzi, ecc. Perciò sarebbe buona pratica non mandare in giro files .DOC senza averli ripuliti... Come si fa? È molto semplice: basta un "copia-incolla" in un file nuovo.

Ma torniamo a usi più intelligenti della ridondanza. Da quanto ho già detto appare infatti che la ridondanza in un messaggio non è sempre uno spreco inutile: l'esempio di una conversazione in presenza di rumore dovrebbe darne un'idea. Se non ci fosse la ridondanza, basterebbe perdere una minima frazione per non poter ricostruire il messaggio. A condizione, si capisce, che il messaggio c'interessi davvero: a tutti noi è capitato di ascoltare ... con un orecchio solo, come si dice, qualcuno che parla; ossia prestandogli un'attenzione solo parziale. Salvo dirigere per intero l'attenzione sul parlatore se ci sembra che ciò che dice si stia facendo interessante...

* * *

Parlando sul serio (e sperando che la vostra attenzione sia ancora desta ...) ci sono situazioni pratiche in cui è proprio necessario garantire la fedele trasmissione di un messaggio *in presenza di rumore*. Se diamo al termine "rumore" un significato sufficientemente astratto, possiamo farci rientrare i casi più diversi:

- una linea telefonica disturbata
- idem per una trasmissione radio
- un CD graffiato
- una fotografia (all'antica, non digitale) che è stata un po' maltrattata in sede di sviluppo e stampa
- un DNA che subisce errori in fase di replicazione

e chi più ne ha più ne metta...

A prima vista sembra inevitabile che il rumore porti a un degrado dell'informazione trasmessa. Però bisogna stare attenti a non cadere in un tranello

linguistico, o psicologico, o filosofico ... non so bene: si potrebbe credere che il degrado stia in una qualche perdita di *qualità*. Non c'è dubbio che dei graffi su un CD potrebbero deteriorare la musica riprodotta; che delle macchie in una foto la rendono più "brutta"; che degli errori nel DNA causano la sintesi di proteine difettose, se non si trasmettono addirittura come mutazioni, molto probabilmente dannose... Ma invece non è questo il punto, almeno per quanto concerne la teoria dell'informazione: ricordiamo la frase di Shannon sull'irrilevanza degli aspetti semantici. Il vero problema è che il rumore *confonde i messaggi*: è per questo che riduce il contenuto d'informazione. Vediamo di capire perché.

Prima di accennare la dimostrazione, mi sembra utile precisare dei termini che ho già usato senza troppa attenzione: mi riferisco all'uso di "carattere" e "simbolo." Ho parlato e parlerò di *simbolo* per indicare le possibili alternative scelte in un "alfabeto": per es. le solite 27 lettere, o i due simboli di un alfabeto binario. Parlo invece di *carattere* per indicare uno degli elementi di un messaggio: il messaggio è una successione di L caratteri, ciascuno dei quali è scelto fra due o più simboli ammessi nell'alfabeto.

Venendo alla dimostrazione, darla in forma precisa e generale non è semplice, e del resto non credo che interessi a nessuno dei miei lettori; mi limito quindi a rendere plausibile il risultato con un esempio. Cominciamo a semplificare il problema considerando una trasmissione *binaria*: come sappiamo, ciò vuol dire che il messaggio consiste di una sequenza di caratteri presi da un alfabeto di due soli simboli, che posso designare con "0" e "1."

Di passaggio osservo che l'uso di "0" e "1" quando si parla di alfabeto binario è solo una convenzione: nella pratica ci sono innumerevoli modi in cui il messaggio può essere rappresentato e trasmesso. Possiamo avere due stati di magnetizzazione, due livelli di potenziale elettrico, la presenza o assenza di un foro in una striscia di carta o di una cavità sulla superficie di un CD, lo stato di conduzione o d'interdizione di un transistor... Ma si possono anche immaginare delle varianti biochimiche: mi viene in mente, perché debbo averla letta da qualche parte, la presenza o assenza di un gruppo metile (dove? in un aminoacido? in uno zucchero? la chimica organica è troppo complicata per me ...).

Assumiamo che i due simboli nel messaggio siano equiprobabili, che è il caso più semplice da studiare. Supponiamo poi che nella trasmissione possano intervenire degli errori (ecco il "rumore"!) consistenti nel fatto che in modo casuale un carattere trasmesso viene scambiato con l'altro: uno "0" arriva come "1," o viceversa; e che questo accada, per ogni carattere, con una probabilità p, indipendente da ciò che accade agli altri caratteri. Nel messaggio ricevuto, di lunghezza L, ci saranno dunque in media Lp errori: il solo problema è che *il ricevente non sa dove sono...*

Si vede dunque che tutto sarebbe risolto se il ricevente venisse informato circa la posizione dei caratteri errati; cosa che si può fare nel modo più semplice inviandogli, accanto al messaggio "principale," che è affetto da errori, un mes-

saggio "di correzione," anch'esso di lunghezza L (che dobbiamo supporre non corrotto) così fatto: esso porta uno "0" dove il messaggio principale non ha subito errore, mentre porta un "1" dove l'errore c'è stato. Dunque il messaggio di correzione è anch'esso binario, ma i suoi simboli non sono equiprobabili: infatti lo "1" ha probabilità p, e lo "0" ha probabilità $1 - p$.

In assenza di errori, il messaggio principale avrebbe avuto un contenuto d'informazione pari a L bit; ma ora vediamo che per riottenere la stessa informazione il ricevente abbisogna di un messaggio supplementare, che porta anch'esso una certa quantità d'informazione, anche se minore di L bit, visto che i suoi caratteri non sono equiprobabili. Se la chiamiamo L_1, è ragionevole assumere che il messaggio affetto da errore abbia appunto perduto altrettanta informazione, e che quindi trasmetta soltanto $L' = L - L_1$ bit.

Notate che nel caso estremo in cui $p = \frac{1}{2}$, si trova $L_1 = L$ e quindi $L' = 0$. Come mai? La risposta è facile: se $p = \frac{1}{2}$, vuol dire che c'è la stessa probabilità di ricevere il carattere giusto o quello sbagliato, ossia che i caratteri ricevuti sono *completamente casuali* e non hanno alcuna correlazione con quelli trasmessi. È dunque giusto che in questo caso il messaggio principale non trasmetta nessuna informazione, mentre la responsabilità di trasmetterla è caricata tutta, per così dire, sul messaggio di correzione.

* * *

Non dimentichiamo però che questo espediente del messaggio di correzione è servito solo per giustificare il teorema sull'informazione perduta in presenza di rumore: nei casi reali l'idea non funziona, per la solita ragione: il destinatario del messaggio non sa dove sono avvenuti gli errori; il mittente non è in grado di prevederlo. Quindi nessuno dei due può realmente costruire il messaggio di correzione.

Sembra dunque che non ci sia niente da fare: a causa del rumore una certa frazione dell'informazione contenuta nel messaggio va irrimediabilmente perduta. Eppure si sa che esistono vie d'uscita a questa spiacevole situazione; anche senza tanto armamentario teorico, chiunque capisce che se un messaggio rischia di arrivare disturbato, ripetendolo due volte si aumenta di molto la probabilità che arrivi giusto. Infatti la probabilità che un certo carattere risulti errato in entrambi gli invii è certo molto più piccola dell'errore in un invio singolo: mentre la seconda vale p, la prima vale p^2, e se $p \ll 1$ certamente $p^2 \ll p$.

Esempio: se c'è una probabilità dell'1% di un singolo errore ($p = 0.01$), quella di due errori nello stesso punto nei due invii è $p^2 = 0.0001$.

C'è a dire il vero ancora un problema: anche se siamo praticamente certi che ogni singolo carattere non può essere errato in entrambi gli invii, tutte le volte che troviamo una differenza non sappiamo quale sia quello giusto! Ma anche a questo si pone rimedio, se si è disposti ad aumentare il prezzo da pagare: si fanno tre invii, si spera che non capitino due errori su tre nello stesso punto, e si decide *a maggioranza*.

Il metodo della maggioranza viene realmente applicato in certi casi: per es. si montano sulle navicelle spaziali tre o più computer che eseguono lo stesso programma, e si prende per buono il risultato della maggioranza, magari qualificata... Ma si tratta di un metodo parecchio costoso, in quanto obbliga ad almeno triplicare gli apparati o la lunghezza dei messaggi; inoltre non funziona nei casi (non rari) in cui la probabilità del singolo errore non è molto piccola. In altre parole, nel caso di trasmissioni in presenza di forte rumore.

Si potrebbe credere che se il rumore è forte, per cui i caratteri sbagliati sono quasi quanti quelli giusti, ci sia poco da fare; e invece il solito Shannon ha dimostrato che è vero esattamente il contrario. Se il rumore riduce la quantità d'informazione trasmessa da L a L', allora è possibile codificare un messaggio con contenuto non superiore a L' bit in un messaggio ridondante di L bit, in modo tale da avere probabilità vicina a 1 quanto si vuole di poter ricostruire esattamente gli L' bit che interessano.

Si noti che questo teorema è un po' più debole, nel senso che non dà una certezza assoluta: garantisce soltanto che si può rendere piccola a piacere la probabilità di un errore residuo. E si noti anche che non è posto alcun limite al rapporto L'/L: il rumore potrebbe essere così intenso da ridurre poniamo a un millesimo il contenuto d'informazione di un messaggio. Poco male: ciò significa solo che se vogliamo trasmettere poniamo 100 bit, ne dovremo impiegare ben centomila (almeno).

A distanza di oltre mezzo secolo dalla prima formulazione teorica, questi risultati hanno ormai frequenti applicazioni pratiche. Per es. negli onnipresenti CD i graffi sono pressoché innocui (a differenza che per i vecchi dischi di vinile ...) in quanto la registrazione è codificata con opportuna ridondanza, e grazie anche ad altri accorgimenti che qui debbo tralasciare. Un'altra applicazione essenziale sono le trasmissioni spaziali: in quei casi il segnale ricevuto è estremamente debole, causa l'immensa distanza e le piccole potenze dei trasmettitori sulle sonde; di conseguenza i disturbi elettromagnetici presenti nello spazio, nell'atmosfera, nonché le sempre più abbondanti trasmissioni radio e TV di tutto il mondo producono un rumore assai alto. Eppure si riesce a ricevere e trasmettere dati vitali anche con sonde ormai ai confini del sistema solare.

* * *

Riepilogando, in queste due puntate vi ho presentato per cominciare il concetto base di quantità d'informazione, nella forma coniata da Shannon oltre 50 anni fa; vi ho poi illustrato alcuni dei risultati principali della teoria, e ho anche indicato alcune applicazioni pratiche. Ribadisco ancora una volta che tutta la teoria dell'informazione si basa sulla deliberata rinuncia a occuparsi del *significato* dei messaggi; e come avete potuto vedere anche dalla mia sommaria esposizione, i risultati mostrano a posteriori quanto feconda sia questa rinuncia, che a prima vista potrebbe apparire quasi autolesionista...

Arrivato a questo punto, dovrei affrontare un altro tema, sicuramente più interessante per chi mi legge: l'uso del concetto d'informazione in biologia. Anche coi miei evidenti limiti, non potrebbe essere un discorso breve, e perciò preferisco rimandarlo a una prossima puntata.