Entropia e compressione dati

Published on February 24, 2018 on beBee

- **Read in English by Google Translate**

Introduzione

Uno dei primi algoritmi di compressione che scrissi, in realtà lo copiai da una rivista di programmazione intorno al 1991 e si trattava di una versione semplificata del PKZIP per Linguaggio C.

Non sapendo bene da dove cominciare per migliorarlo decisi di fare un analisi esadecimale dei dati in uscita dal compressore e mi stupì la sensazione di "*chaos*" ma non quello che potremmo trovare in un frattale o in una tempesta piuttosto nello schermo a "*neve*" di una TV non sintonizzata.

Un'altra domanda che mi affliggeva era se fosse possibile utilizzare l'algoritmo di compressione in modo iterattivo per ottenere una maggiore compressione.

Ovviamente, non quello nello specifico perché bastava applicarlo N volte per capire che dopo la prima volta la dimensione dei dati in uscita tendeva a crescere.

Ovviamente, un algoritmo iterattivo non avrebbe potuto funzionare all'infinito altrimenti tutto si sarebbe compresso a un solo bit e la cosa sarebbe stata assurda.

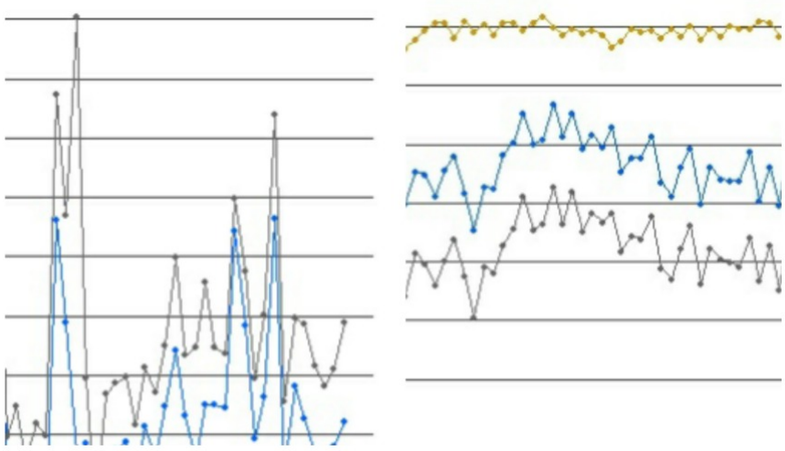
Ovviamente, fra prove pratiche con un'implementazione e una teoria più genrale le cose potevano essere molto diverse. Perciò, la curiosità permaneva.

Il prodotto del processo

Quando il prodotto (dati compressi) è intimamente legato al processo (algoritmo di compressione) e nell'eventualità che "rompere" il processo non è utile o non è fattibile allora vale la pena di concentrarsi sull'analisi del prodotto. Chiesto approccio si chiama blackbox analisys.

Così decisi di scrivere un programma che graficasse la frequenza relativa ai 256 valori del byte su 256 colonne.

I miei grafici erano realizzati in ASCII per semplicità di programmazione (erano tempi duri, bisognava scriversi tutto da soli) ma qui sotto sono riportati due grafici indicativi, giusto per dare un'idea del confronto fra ingresso e uscita.



Il risultato fu illuminante: il testo ASCII aveva caratteri ricorrenti e altri assenti mentre i dati compressi tendendevano a presentare valori uniformi su tutto lo spettro. Il massimo livello di compressione era asintotico ad una curva di spettro piatta (rumore bianco).

Dal cilindro nero, un coniglio bianco

Ora, se ci pensiamo un attimo – magari dall'alto della nostra cultura matematica – la cosa può apparire normale. Ma non è intuitiva.

Non è intuitivo pensare che un segnale di massima densità informativa assomigli in tutto e per tutto al segnale "*neve*" della TV analogica fuori sintonia.

Perché viceversa un rumore bianco potrebbe essere un segnale "*alieno*" da scompattare. Il rumore bianco assumeva così una valenza quasi mistica. Era il coniglio di Alice sbucato fuori da una scatola nera.

Informazione e caos

La differenza fra un file compresso e un blocco di dati generati con una variabile pseudo casuale non era tanto nella qualità del generatore di numeri pseudo casuale perché in termini di densità informativa già allora era equivalente ad un blocco dati conpresso.

La differenza sostanziale sta nel fatto che un file compresso ha un'intestazione (header) che specifica il tipo di algoritmo usato e la sua versione, poi segue una tavola dei simboli (table) che negli algoritmi di compressione primitivi coincideva con il blocco dati stesso e poi il blocco dati (data).

Quando la quantità di dati da comprimere è grande abbastanza, la tavola dei simboli si riempie e poi occorre ripartire con una tavola nuova (restart).

Perciò l'indice di densità informativa sale per assestarsi nell'intorno inferiore del massimo e poi crollava per risalire. Invece, nel rumore bianco l'indice rimaneva sempre nell'intorno inferiore del massimo.

Quindi, la possibilità di distinguere un segnale alieno dalla statica dipendeva dalla possibilità di identificare delle cadute di entropia all'incirca periodiche.

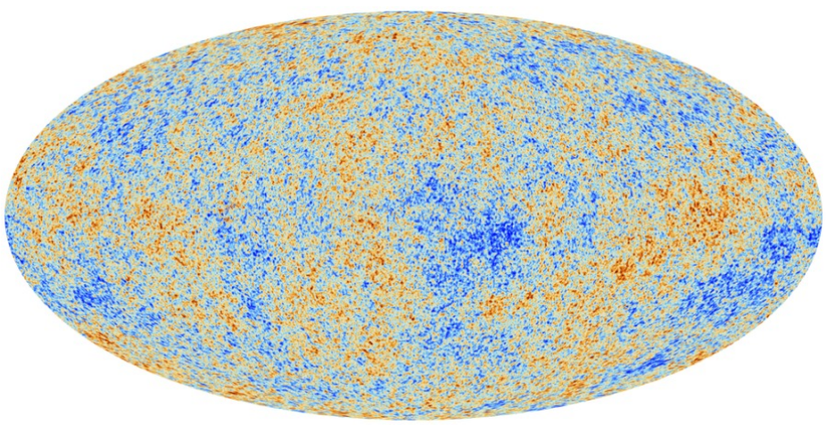
L'entropia come misura dell'informazione

Intanto, abbiamo introdotto un concetto quello di entropia che generalmente si incontra alle superiori o al più tardi all'università. Perciò già lo avevo.

Così, non solo la statica ma anche la statistica, la termodinamica e più in generale il concetto di varianza entrava di prepotenza nell'informtica – in realtà c'è sempre stato anche etimologicamente – ma guardiamo da un punto di vista di chi ha studiato a scuola fisica e chimica ma programmava prima di avere accesso a una teoria scientifica dell'informazione.

Non è una scoperta banale comprendere che il termosifone, la statica TV e l'algoritmo di compressione hanno qualcosa di molto intimo e profondo in comune.

Soprattutto perché nello spaccare un algoritmo si sarebbe potuto comprendere qualcosa anche di un termosifone mentre nello spaccare un termosifone si sarebbe compreso poco di utile.

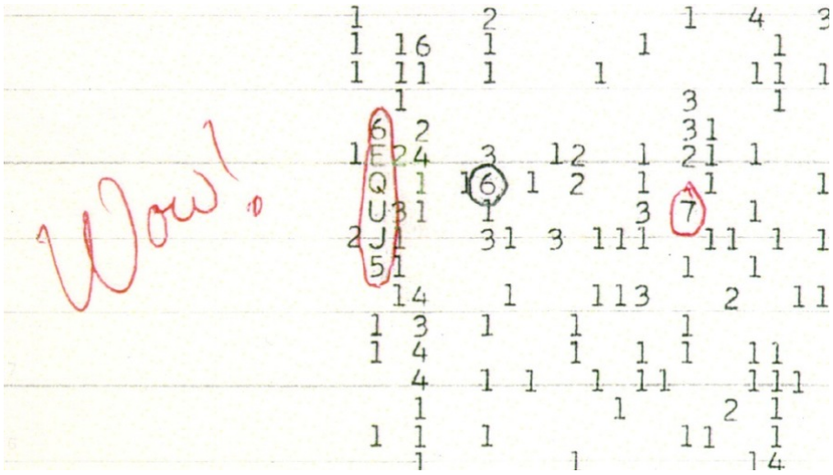


Qua sopra un'immagine della mappa celeste del**la radiazione di fondo residua del BigBang**. Certi caloriferi quando si spaccano, in effetti, producono qualcosa d'intrigante.

Seguire il bianconiglio nella sua tana

Qui, si apre un'altra finestra: se avessimo intestazione e tavola, potremmo interpretare quella che percepiamo come statica e finalmente leggere il messaggio che questi misteriosi alieni freneticamente ci inviano.

Qui sotto é rappresentata una "*caduta d'entropia*" [1] nel segnale d'ingresso di un radiotelescopio noto come "**Wow! Signal**".



Viceversa, se togliamo intestazione e tavola otteniamo un file cifrato. In realtà, come preannunciato sopra negli algoritmi primitivi di compressione non vi era distinzione fra tavola e blocco dati perciò il solo togliere l'intestazione era pleonastico come cambiare l'estensione di un file da .ZIP a .DAT.

Però concettualmente, si parte così. Perché se non si può togliere la tavola però si può utilizzare un algoritmo di scrambling. Poiché i blocchi di compressione non hanno tutti la medesima lunghezza perché la velocità di riempimento della tavola dei simboli dipende dalla variabilità del contenuto da comprimere, un'algoritmo a finestra fissa di scrambling tende a mascherare le variazioni periodiche di entropia rendendo il blocco cifrato ancora più simile una sequenza pseudo casuale.

Si può generalizzare, l'algoritmo di scrambling con una sequenza di byte che ne parametrizzino il funzionamento. Questi N byte costituiranno la chiave di decodifica che sarà forte quanto 2^(8N) quindi per rompere questa cifratura a chiave di 4 byte = 32 bit servono 2^32 = 4 miliardi di tentativi.

La crittografia è un'altra cosa

In realtà, basta decodificare con una chiave a caso solo della prima finestra di scrambling e poi passarli all'algoritmo di decompressione, se fallisce l'interpretazione prima di aver completato tutti i byte forniti allora la chiave non è quella giusta. Inoltre ci sono metodi anche più raffinati per rompere questo tipo di cifratura ma appunto abbiamo solo aperto il vaso di Pandora.

Ora, il vaso di Pandora è ben più profondo di così perché nel momento che si grafica la frequenza dei simboli in una finestra si è fatto una primordiale analisi di Fourier.

Se aggiungiamo che prima del PC c'era stato lo ZX Spectrum che caricava e salvava dati da cassetta magnetica in formato audio, la conversione analogico digitale e digitale analogica era già un fatto palese. Quindi abbiamo il teorema del campionamento di Shannon, etc.

Ma c'è di più di questo perché le funzioni statistiche sono anche funzioni di convoluzione tipiche nell'applicazione di filtri audio e quindi dei dispositivi elettronici di tipo RLC che per loro natura possono essere anche oscillatori. Insomma, prima ancora di arrivare alla meccanica quantistica, ormai la frittata era ormai fatta ed è stata tutta colpa di un algoritmo di compressione dati!

Conclusione

Non abbiamo ancora trovato vita intelligente al di fuori del nostro pianeta ma non ci arrenderemo facilmente!

Articoli correlati

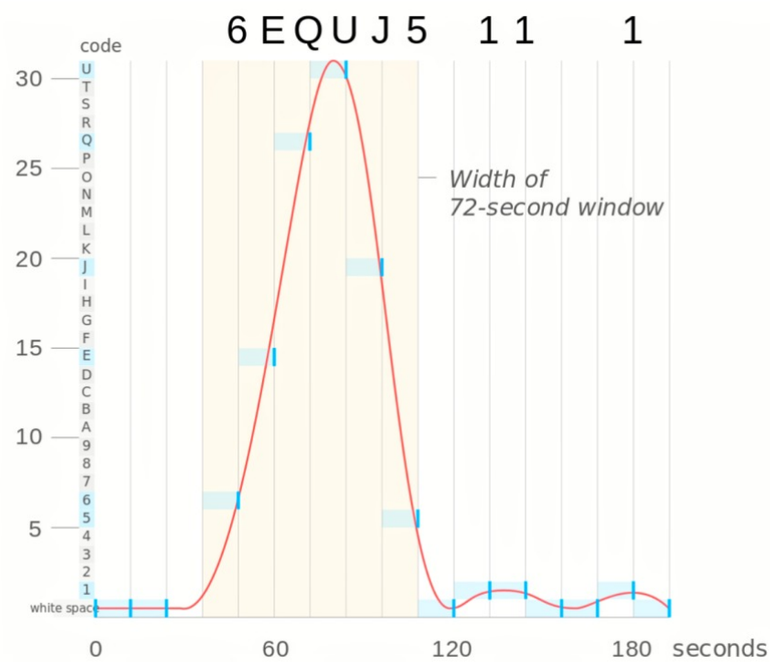
- Gli indicatori statistici di secondo livello (16 ottobre 2017, IT)
- Paradosso di Fermi e Singolarità A.I. (16 ottobre 2017, IT)
- ...e l'alba sorse su dune di spazzatura (17 ottobre 2017, IT)

Lettura esterna

- Stima entropica nei compressori dati

Note

[1] In realtà si tratta di una picco d'intensità su una banda di frequenza e in particolare quella dell'idrogeno atomico. Allora cosa c'entra l'entropia? Osservando il ritaglio della stampa si nota che tutto intorno al segnale Wow! i caratteri più frequenti sono spazio (0), uno, due e tre. In generale, interi tabulati avranno avuto una distribuzione di caratteri simile, poi d'improvviso sono apparsi caratteri (intensità) mai prima osservati. Non caratteri a caso ma con uno specifico pattern: "*a salire fino al massimo di picco e poi a scendere*" entro una finestra temporale che era quella di osservazione 72 secondi. In questo caso si osservano due fenomeni correlati: il primo é una sorpresa e l'altro un decremento di entropia ovvero una variazione improvvisa di uno schema da disordinato a ordinato. Questo ci fa comprendre che un conto é la densità di informazione, un altro conto é la significatività di un informazione, e un'altra cosa ancora é l'entropia. In un tabulato in cui i caratteri sono normalmente variabili in [0-3] la densità d'informazione é di 2bit per posizione.



Se andiamo a osservare il segnale Wow!, in quella finestra anche tutto l'alfabeto [0-9,A-Z] diventa accessibile, il segnale é prevedibile: SU per metà e GIÙ per metà. Circa 3÷4bit per posizione. La densità di informazione appare doppia ma su un alfabeto 8 volte più esteso. Si può obbiettare che anche prima l'intero set di caratteri fosse disponibile. Certo, disponibile nella telescrivente ma non nel messaggio. Perciò si é passati da un livello di entropia di circa 1:1 (rumore bianco) a 1:8 (6EQUJ5) e una densità di informazione doppia ma in termini significativi dipende da quanto quella specifica combinazione di segnale/ricevitore possa essere compatibile con un fenomeno naturale oppure no, se abbiamo osservato una trasmissione extraterrestre oppure no.