

Compressione dei dati

Generalità

Spesso risulta utile ridurre le dimensioni dei dati su cui si sta lavorando in modo da renderne più agevole l'archiviazione e la trasmissione. Una condizione tipica in cui viene applicata la compressione sui dati è la ridondanza. In poche parole, dei dati che contengono le stesse informazioni in qualche modo ripetute (ridondanza) sono facilmente comprimibili. Ad esempio, quando ci si trova di fronte a un testo scritto si incorre in esempi di ridondanza lampanti. Si consideri, per esempio, l'uso che la lingua italiana fa della lettera `q' che è *sempre* seguita da una `u': nessuno avrebbe difficoltà nel capire il significato della parola ``q*adro", nonostante essa non sia scritta correttamente. Quindi, una prima tecnica di compressione della lingua italiana potrebbe essere basata sull'omissione del carattere "u" tutte le volte che questo è preceduto da un carattere "q".

La presenza di ridondanza in una lingua, in un testo scritto, o in una qualunque stringa di simboli non è però nefasta, anzi, essa svolge il compito fondamentale di rendere *robusto* il messaggio contenuto in quella stringa. Per robusto si intende facilmente comprensibile e non incline ad essere male interpretato anche nel caso in cui il messaggio venga trasmesso in modo solo parziale. Errori di trasmissione su codici ridondanti sono molto più facilmente correggibili.

D'altra parte, un testo che sia stato compresso al massimo è molto fragile, in quanto, per definizione di massima compressione, ciascun simbolo sarà portatore di informazione, e quindi una sua alterazione porterà ad una perdita irrimediabile di quella stessa informazione.

In definitiva, l'uso di codifiche ridondanti può avere due motivazioni: la *flessibilità* e l'*affidabilità*.

- La *flessibilità* indica la possibilità di utilizzare la stessa codifica in situazioni diverse. Es: il cosiddetto "baco del millennio" (millennium bug) che costa al mondo migliaia di miliardi è dovuto alla scarsa lungimiranza con cui i produttori di software hanno deciso di utilizzare due sole cifre decimali per rappresentare l'anno nelle date, dando per scontate (e quindi prive di contenuto informativo) le prime 2, fino ad ora sempre uguali a 19.
- L'*affidabilità* deriva dalla capacità di alcune codifiche ridondanti di rivelare o correggere errori. Un codice non ridondante associa univocamente una configurazione ad un significato. Un errore nella configurazione comporta un errore di interpretazione del significato. Una codifica ridondante a rivelazione d'errore associa significati utili solo ad un sottoinsieme delle configurazioni possibili. Se un errore trasforma una configurazione significativa in una non significativa, chi la interpreta riconosce la presenza dell'errore. Una codifica ridondante a correzione d'errore associa molte configurazioni allo stesso significato. Se un errore trasforma una configurazione in una a cui è associato lo stesso significato, l'interpretazione resta corretta.

Codifiche ridondanti si prestano a compressione, come dimostrano gli esempi precedenti, ma esistono anche tecniche di compressione che agiscono su dati irridondanti sfruttando proprietà tipiche dell'informazione da rappresentare.

Tecniche generiche di compressione

Per quanto riguarda le tecniche di compressione una prima distinzione che bisogna fare è tra compressione *lossy* e *lossless*. La prima tecnica effettua una compressione dei dati definita “con perdita dell’informazione” o anche distruttiva in quanto, una volta applicata questa tecnica, non è più possibile ricostruire in maniera esatta i dati di partenza attraverso il processo di decompressione. In definitiva, c’è stata, una perdita irrimediabile di informazione.

Tecniche di tipo *lossy* vengono applicate nella compressione delle immagini, dei filmati e dei suoni dove perdere dell’informazione significa diminuire la qualità dell’immagine o del suono. Questa tecnica è molto utilizzata nel campo della compressione dei formati multimediali soprattutto laddove è spesso inutile ricostruire, come nel caso delle fotografie, l’informazione iniziale con una risoluzione maggiore di quella che l’occhio umano può apprezzare.

Tecniche di compressione di tipo *lossless* permettono invece di recuperare interamente l’informazione contenuta nel testo prima della sua compressione, ma la loro efficienza di compressione è minore. Queste tecniche sono comunque le uniche utilizzabili in quei casi in cui non è possibile accettare neppure la minima perdita delle informazioni. Ad esempio la tecnica di compressione che sta alla base del compressore Winzip è assolutamente una tecnica di tipo *lossless* in quanto non sarebbe accettabile perdere dell’informazione comprimendo ad esempio un file di installazione di un programma.

Tecniche di tipo lossless

Compressione run-length

La compressione di tipo *run-length* fa parte delle tecniche di compressione di tipo *lossless* cioè senza perdita di dati. Essa si basa su un algoritmo denominato di *Run Length Encoding (RLE)* che è uno dei più semplici algoritmi di compressione mai progettati. Si basa sul fatto che nei dati da comprimere esistono sequenze, definite *run*, che si ripetono costantemente. Una volta individuate le sequenze ripetute, vengono sostituite da un unico simbolo e dal numero delle ripetizioni presenti. Ad esempio data una stringa di bit del tipo “01110000” l’algoritmo RLE la comprimerebbe codificandola in “03*14*0” che sta ad indicare “0 tre volte 1 quattro volte 0”.

Esistono numerose varianti di RLE le cui principali differenze consistono nella lunghezza minima da attribuire ad un run. Nell’esempio appena descritto abbiamo usato una lunghezza del run pari a uno cioè si prendeva in considerazione un carattere alla volta e si cercava se questo veniva ripetuto consecutivamente. Allo stesso modo si possono prendere in considerazione gruppi di simboli e cercare se l’intero gruppo viene ripetuto all’interno della stringa.

Compressione a codifica differenziale

In alcuni casi, le informazioni sono costituite da blocchi di dati, ognuno dei quali differisce leggermente dal precedente come, ad esempio, i fotogrammi successivi di un filmato. In questo caso sono utili le tecniche di compressione che utilizzano la *codifica differenziale*. L’approccio di queste tecniche è quello di memorizzare non il blocco stesso ma le sue differenze rispetto al precedente. È evidente che in questo caso se si dovesse corrompere un blocco compresso durante la trasmissione, risulterebbe compromessa la ricostruzione/decompressione di tutti i blocchi successivi.

Codifiche adattative basate su dizionari

Il termine dizionario si riferisce all’insieme di elementi di base sui quali viene ricostruito il messaggio compresso. In pratica viene utilizzato un insieme di simboli (dizionario) per codificare un messaggio. I simboli del dizionario rappresentano particolari sequenze di bit e durante la compressione, ad ogni sequenza riconosciuta viene sostituito il simbolo corrispondente. La particolarità di queste tecniche sta nel fatto che il dizionario viene creato dinamicamente durante il processo di compressione. Tale tecnica sta alla base dell’algoritmo di compressione Lempel-Ziv che troviamo nella maggior parte dei tools di compressione quali WinZip.

Compressione delle immagini

Formato GIF

La sigla GIF è acronimo di Graphic Interchange Format. Questo tipo di compressione rientra nelle tecniche di tipo lossless, cioè senza perdita di dati. La caratteristica del formato GIF è che esso può esportare solo immagini che contengono al massimo 256 colori. Se l'originale contiene un numero più elevato di colori quindi, è necessario effettuare una riduzione e la perdita di qualità sarà significativa. Il formato GIF usa colori a 8 bit ed è efficace per comprimere immagini vettoriali, geometriche o testo. Questo formato fu diffuso negli anni Ottanta come metodo efficiente di trasmissione delle immagini su reti di dati. All'inizio degli anni Novanta i progettisti originali del web lo adottarono per l'efficienza che offriva. Oggi la stragrande maggioranza delle immagini sul web è in questo formato ed è supportato da tutti i browser web.

Il formato GIF usa una forma di compressione LZW che mantiene inalterata la qualità dell'immagine, ovvero riduce le dimensioni del file senza pregiudicare la qualità grafica dell'immagine. Questo formato è basato sull'uso di una tavolozza di colori: anche se il singolo colore può essere uno fra milioni di sfumature, solo un certo numero di essi è disponibile (al massimo $256=8\text{bit}$). I colori sono memorizzati in una 'tavolozza', una tabella che associa un numero ad un certo valore di colore.

La limitazione a 256 colori appariva ragionevole all'epoca della creazione del formato GIF perché non erano ancora diffusi dispositivi in grado di visualizzarne un numero superiore. Per disegni al tratto, fumetti, fotografie in bianco e nero sono di regola sufficienti 256 colori. Minore sarà il numero di colori presenti nell'immagine e maggiori saranno le possibilità di compressione, ovvero minori saranno le dimensioni del file in quanto sarà ridotta la dimensione della tavolozza e quindi il numero di codici per descrivere i colori.

Il formato GIF consente anche di salvare le immagini in un formato interlacciato. Il formato a interlacciamento produce una visualizzazione graduale di un'immagine in una serie di passate sempre più definite a mano a mano che i dati arrivano al browser. Ogni nuovo passo crea un'immagine più nitida fino al completamento dell'intera immagine.

Il formato GIF consente anche di definire un colore come trasparente. Nelle aree di colore contrassegnato come trasparente, verrà visualizzato il colore di sfondo. Questa proprietà viene utilizzata per le animazioni e per la sovrapposizione di più immagini.

Formato JPEG: tecnica di tipo lossy

Un formato grafico utilizzato di frequente sul Web per ridurre le dimensioni dei file grafici è lo schema di compressione JPEG (*Joint Photographic Expert Group*). A differenza delle immagini GIF, le immagini JPEG sono policrome (24 bit, o 16,8 milioni di colori). Questo tipo di immagini ha generato un altissimo interesse tra fotografi, artisti, progettisti grafici, specialisti della composizione di immagini mediche, storici dell'arte e altri gruppi per i quali la qualità dell'immagine è d'importanza fondamentale e per i quali non è possibile accettare compromessi sulla fedeltà dei colori tramite retinatura di un'immagine a colori a 8 bit.

Una forma più recente di JPEG, chiamata JPEG *progressivo*, conferisce alle immagini JPEG la stessa gradualità di visualizzazione delle immagini GIF interlacciate; al pari di queste ultime, le immagini JPEG progressive impiegano spesso un tempo maggiore per lo scaricamento sulla pagina rispetto ai JPEG standard, ma offrono al lettore un'anteprima più rapida.

La compressione JPEG utilizza una sofisticata tecnica matematica, chiamata trasformazione discreta del coseno, per produrre una scala scorrevole di compressione delle immagini. Tale tecnica si basa su di una codifica dell'immagine "*percettiva*" in cui viene distinta la luminosità dei pixel dal loro colore. Il motivo di tale distinzione è che l'occhio umano è più sensibile alle variazioni di luminosità che non a quelle di colore. Lo standard base di JPEG trae vantaggio da questo fenomeno codificando ogni componente della luminosità, ma dividendo l'immagine in blocchi di quattro pixel e registrando solo il colore medio di ogni blocco. La rappresentazione finale preserva dunque i

cambiamenti di luminosità, attenuando i repentini mutamenti cromatici dell'immagine originale. Il vantaggio è che ogni blocco di quattro pixel è rappresentato soltanto da 6 valori (4 di luminosità e 2 di colore) anziché 12 valori che sarebbero necessari in un sistema a 3 valori per pixel (immagini a 24 bit RGB).

È possibile scegliere il grado di compressione che si desidera applicare a un'immagine in formato JPEG, ma in questo modo si determina anche la qualità dell'immagine. Più si comprime un'immagine con la compressione JPEG, più si riduce la qualità dell'immagine stessa.

Formato PNG: tecnica di tipo lossless

Il formato PNG (Portable Network Graphic) è stato sviluppato appositamente per il Web. Questo formato è stato disponibile fin dal 1995 ma ha stentato ad acquisire popolarità a causa della mancanza di un supporto generalizzato da parte dei browser. Si tratta di un formato che secondo le intenzioni degli autori doveva sostituire il formato GIF. Questo formato senza perdita di informazioni comprime le immagini a 8 bit producendo file di dimensioni inferiori rispetto a GIF ma supporta anche immagini a 16 e 24 bit.

Anche se il formato PNG supporta il colore a 24 bit, la sua routine di compressione senza perdita di informazioni non è in grado di raggiungere l'efficienza del formato JPEG. Il formato PNG supporta, inoltre, le funzionalità di trasparenza e interallacciamento ma non l'animazione.

Un'utile caratteristica del formato PNG è la capacità di incorporare del testo per offrire la possibilità di eseguire ricerche sulle immagini; è infatti possibile memorizzare nel file dell'immagine una stringa che identifica l'immagine stessa. Purtroppo il formato grafico PNG non è ampiamente supportato e l'implementazione corrente delle immagini PNG in Netscape Navigator e Microsoft Internet Explorer non supporta completamente tutte le sue funzioni.

Ref: J. Gleen Brookshear, "Informatica una panoramica generale"
S. Conway, "HTML 4.01 guida per il programmatore"