# CODIFICA DI SORGENTE E IL CODICE DI HUFFMAN

Con l'avvento di nuove tecnologie nella codifica dei dati digitali, soprattutto quelli multimediali, quali immagini, audio e video, le tecniche di compressione acquistano sempre maggiore importanza. I file multimediali infatti comportano elevate quantità di dati che necessitano di spazio per la memorizzazione e di alte velocità trasmissive nel trasferimento ad altri sistemi digitali. Quindi la **compressione** dei dati porta ad un risparmio di memoria e soprattutto all'*agevolazione nella trasmissione*. Perciò sono state inventate e studiate **varie tecniche di compressione** che cercano appunto di ridurre il numero di bit necessari per immagazzinare un'informazione, organizzando in modo più efficiente le informazioni per ottenere una memorizzazione che richieda minor uso di risorse. Le tecniche di compressione vengono applicate da appositi programmi immediatamente prima della memorizzazione o trasmissione dei dati. Ovviamente in fase di lettura o ricezione analoghi programmi devono effettuare la decompressione.

**Codici a lunghezza variabile**

Se i simboli da trasmettere **non sono equiprobabili** vengono utilizzati codici a *lunghezza variabile.* Questo consente di usare codici brevi per rappresentare i simboli più frequenti e più lunghi per quelli meno frequenti. Questo consente di ottenere dei codici molto più efficienti rispetto a quelli a lunghezza fissa.
I codici a lunghezza variabile consentono di ottenere un minor numero medio di bit per trasmettere un certo messaggio.

**Codifica di Huffman**

La codifica di Huffman è basata sul metodo per la costruzione di codici con minima-ridondanza. L'algoritmo costruisce un **codice** prefisso ottimo **a lunghezza variabile**, detto codice di Huffman. Utilizzando un codice a lunghezza fissa per rappresentare ciascuno di N caratteri con una stringa binaria servono $\log_2 N$ bit.
Un codice a lunghezza variabile invece si comporta meglio di un codice a lunghezza fissa, proprio perché si può assegnare una parola del codice corta ai caratteri molto frequenti, e una parola lunga ai caratteri meno frequenti. Il codice di Huffman è un codice prefisso *univocamente decodificabile*. Un codice prefisso è contraddistinto dalla seguente proprietà: nessuna parola del codice è anche un prefisso di un'altra parola del codice.

**Algoritmo di Huffman (1951-52)**

Formalizziamo ora il metodo di costruzione di un codice a lunghezza variabile, in modo che soddisfi automaticamente la regola del prefisso. Ovviamente si devono conoscere le frequenze di trasmissione di ciascun simbolo.
Il metodo si basa sulla costruzione di un **albero binario**, i cui rami sono etichettati con 1 e 0 e le cui foglie sono tutti i simboli che può emettere la sorgente; si ottiene poi la codifica di ogni simbolo visitando l'albero dalla radice al simbolo stesso.
Il procedimento può essere descritto come segue:

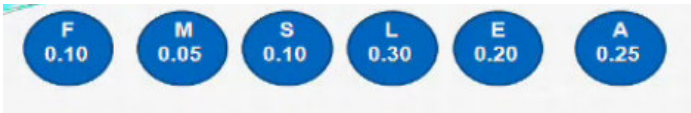
1. Ordina in modo decrescente (non crescente) i simboli in base alle rispettive probabilità creando una tabella;
2. *finché* c'è più di un elemento nella tabella *ripeti*:
   - associa a ciascuno dei due simboli con probabilità minore un nodo foglia dell'albero;
   - rimuovi dalla tabella i due simboli con probabilità più bassa;
   - crea un nuovo nodo interno all'albero con questi due nodi come figli, e con probabilità pari alla somma delle loro probabilità;
   - aggiungi il nuovo elemento alla tabella con la probabilità somma calcolata, rispettando l'ordinamento;

   *fine ripeti*;
3. il nodo rimanente è la radice con probabilità 1, e l'albero è completo.
4. A partire dalla radice si visita l'albero assegnando ogni volta 1 al ramo del sottoalbero di destra e 0 a quello di sinistra fino ad arrivare a tutte le foglie.
5. Si codifica ogni simbolo leggendo la sequenza di 0 e 1 che si incontrano scritti sui rami, attraversando l'albero dalla radice al simbolo stesso.

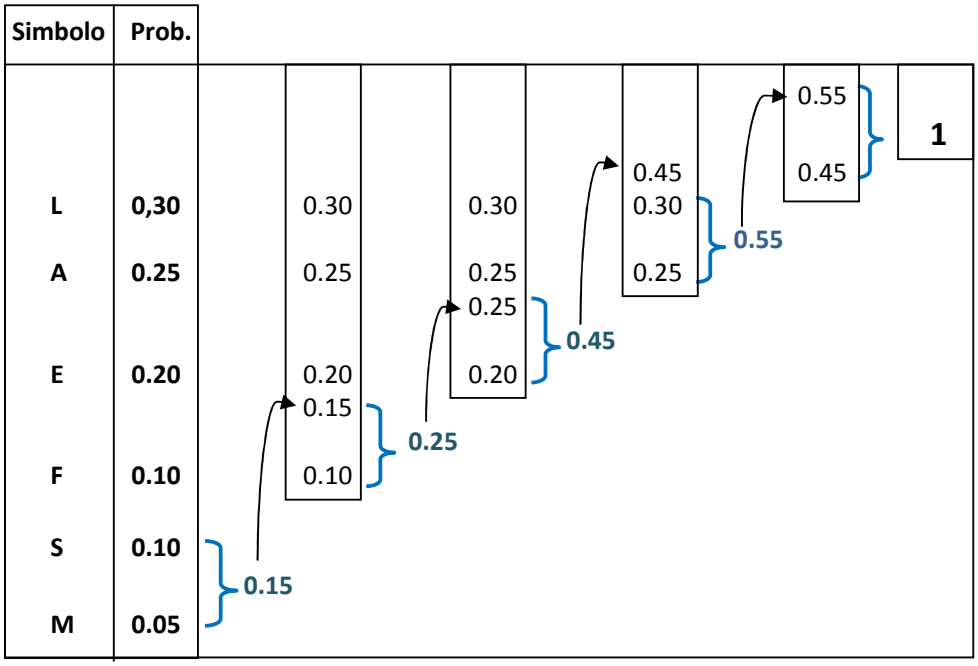
Si può dimostrare che il codice di Huffman generato in questo modo è il migliore possibile nel caso in cui la statistica dei simboli di sorgente sia nota a priori, nel senso che produce una codifica con il minor numero possibile di bit/simbolo medi. La codifica di Huffman è ampiamente utilizzata nel contesto di altri metodi di compressione (metodo DEFLATE di PKZIP) e di codec multimediali (JPEG e MP3), in virtù della sua semplicità, velocità ed assenza di brevetti. Ovviamente ci deve essere un accordo a priori tra sorgente e destinatario a riguardo delle corrispondenze tra parole di codice e simboli (o blocchi di simboli) della sorgente. Nel caso in cui ciò non sia vero, oppure nel caso in cui la statistica dei simboli della sorgente sia stimata a partire dal materiale da codificare, occorre inviare all'inizio della comunicazione anche la tabella di corrispondenza, eventualmente codificata a sua volta.

**Esempio di costruzione di un codice di Huffman**

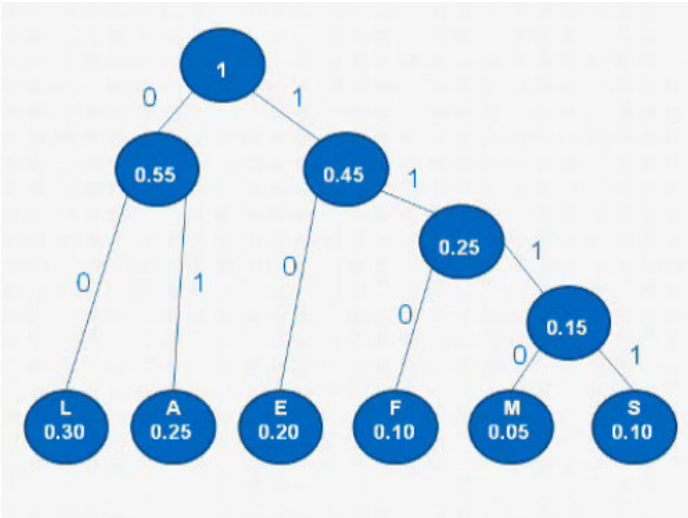
Sia **S** una sorgente che può trasmettere solo i seguenti caratteri dell'alfabeto con le probabilità indicate:




Si seguono i passi da 1. a 5.:




| Simbolo | Codice | Lungh.(bit) |
|---|---|---|
| L | 00 | 2 |
| A | 01 | 2 |
| E | 10 | 2 |
| F | 110 | 3 |
| S | 1111 | 4 |
| M | 1110 | 4 |




Calcolare ora:

La **lunghezza media** del codice

$$L = \sum_{i=1}^{N} p(s_i) \cdot l(s_i) \quad bit/simbolo$$

L' **entropia** della sorgente

$$H(S) = \sum_{i=1}^{N} p(s_i) \cdot \log_2 \frac{1}{p(s_i)} \quad bit/simbolo$$

L' **efficienza** del codice costruito

$$\eta = \frac{H(S)}{L}$$