

Compression de texte

Thierry Lacroq

Université de Rouen
FRANCE

La compression

message \longrightarrow message codé

- gain d'espace de stockage
- gain de temps de transmission

La compression

2 types

- sans perte : le message décodé est égal au message d'origine
⇒ texte
- avec perte : le message décodé est différent du message d'origine
⇒ images et sons

Entropie

$A = \{a_0, \dots, a_{n-1}\}$ et $s = \text{card } A$

Soit une source d'informations $\mathcal{S} = (A, \mathcal{P})$

$\mathcal{P} = (p_0, \dots, p_{s-1})$

p_i : probabilité d'occurrence de a_i dans un mot sur A^+

\mathcal{S} est une source **sans mémoire** si les p_i sont indépendants et stables (source stationnaire)

\mathcal{S} est une source **markovienne** si les p_i dépendent des symboles précédemment émis

Définition

$$\begin{aligned}H(\mathcal{S}) &= H(p_0, \dots, p_{s-1}) \\&= - \sum_{i=0}^{s-1} p_i \log_2(p_i) \\&= \sum_{i=0}^{s-1} p_i \log_2\left(\frac{1}{p_i}\right)\end{aligned}$$

Entropie

Proposition

Soit $\mathcal{S} = (A, \mathcal{P})$ une source alors $0 \leq H(\mathcal{S}) \leq \log_2 s$

Longueur moyenne d'un code

C : code

f : fonction de codage

$\mathcal{S} = (A, \mathcal{P})$: source

$$|C| = \sum_{i=0}^{s-1} |f(a_i)| p_i$$

Exemple 1

$$\mathcal{S} = (\{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}, (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}))$$

C

$$f(\mathbf{a}) = 00$$

$$f(\mathbf{b}) = 01$$

$$f(\mathbf{c}) = 10$$

$$f(\mathbf{d}) = 11$$

$$|C| = 2$$

$$H(\mathcal{S}) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1,75$$

Exemple 2

$$\mathcal{S} = (\{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}, (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}))$$

C

$$f(\mathbf{a}) = 0$$

$$f(\mathbf{b}) = 10$$

$$f(\mathbf{c}) = 110$$

$$f(\mathbf{d}) = 1110$$

$$|C| = 1,875$$

$$H(\mathcal{S}) = 1,75$$

Théorème de Shannon

Théorème

Soit S une source sans mémoire d'entropie H . Tout code uniquement déchiffrable de S sur un alphabet A de cardinal s , de longueur moyenne ℓ vérifie :

$$\ell \geq \frac{H}{\log_2 s}.$$

De plus il existe un code uniquement déchiffrable de S sur un alphabet de cardinal s de longueur moyenne ℓ qui vérifie :

$$\ell < \frac{H}{\log_2 s} + 1.$$