Compression

16 avril 2018

Table des matières

1 COMPRESSION DE DONNEES

site slides : `www.lrde.epita.fr/~gtochon/CODO/`

Compression + decompression

Sans compression de donnees, un film 720p d'une heure ferait presque 400Gio.

La naissance naive de la compression de donnees est nee a peu pres en meme temps que le code morse. La naissance mathematique est avec Shannon.

Autre acteurs : - Abraham Lempel - David Huffmann - Terry Welch

—

Le but de la compression est que lors de la decompression, ce soit le moins visible possible par l'utilisateur.

—

1.1 On data compressebility

On va chercher a eliminer les redondances, et les gaspillages. L'outil que l'on va utiliser est l'entropie.

Du point de vue de Shannon, plus un message va etre probable, moins il va contenir d'informations.

—

Prenons un alphabet de N symboles

$$\sum = \{s_1, s_2 s_N\}$$

de probabilite $p(s_1)$ = proba (le symbole s_i apparait) N valeurs de proba

$$P_{1'}...P_N$$

avec $\sum P_i = 1$

F un fichier construit sur cet alphabet $\sum$/ distribution de proba (P_1, P_{N1}) $\rightarrow$ qte d'information totale du symboles$s_i = -log_2(P_1)$

F contient N_F symboles $\rightarrow$ $S_i$est statiquement present $N_F P_i$

$$Q_{Tot}(S_i) = -N_{FP_i}Sh$$

$$Q_Tot(F) = Q_{Tot}(S_1) + Q_{Tot}(S_2) + ... + Q_{Tot}(SN_2)$$

$$Q_{Tot}(S_1) = -N_F log_2(P_i)$$

$$Q_{Tot}(S_2) = -N_{FP_2} log_2(P_2)$$

$$Q_{Tot}(SN_2) = -N_F P_N log_2(P_N)$$

$$Q_{Tot}(F) = \sum -N_{FP_i} = -N_F \sum P_i log_2(P_i) Sh$$

Un gros fichier "probable" contient plus d'info qu un petit fichier "improbable" Q_{Tot} $\rightarrow$ Sh

$\rightarrow$ H $\rightarrow$ Sh/Symbole $H = -\sum P_i log_2(P_i)$ $\rightarrow$ ne depend pas du fichier considere, mais uniquement de la distribution de proba des symboles composant le fichier.

X variable aleatoire de valeur $\{x_1, x_2...x_n \rightarrow P_i = P(X = x) \sum P_i = 1$

$E[X] = \sum_{i=1}^{N} x_1 P(X = x_i) = \sum_{i=1}^{N} x_i P_i$ $E[\phi(X)] = \sum_{i=1}^{N} \phi(x_i) P_i$ $H = -\sum_{i=1}^{N} P_i log_2 P_i = \sum_{i=1}^{N}(-log_2(Pi))P_i = \sum_{i=1}^{N} q_{S_i * P_i} H = E[q(s_i)]$ $\sum = 0, 1 P(0) = P_0 = P P(1) = P_1 = 1 - P$ $H = -plog_2(p) - (1-p)log_2(1-p) = H(p)$

— Missing things —

$\sum$ avec$N$symboles $s_i, i = 1, ..N$ (2^n)

$$P(S_i) = \frac{1}{N_{\sum}} = \frac{1}{2^n}$$

$$H = -\sum_{i=1}^{N_{\sum}} log_2(P_i) = -\sum_{i=1}^{2^n} = -\sum_{i=1}^{2^n} \frac{1}{2^n} log_2(\frac{1}{2_n})) = \frac{1}{2_n} * (-n) * \sum_{i=1}^{2^n} * \sum_{i=1}^{2^n} 1 = n 2^n 2^{-n} = n$$

(Sh/Symbole)