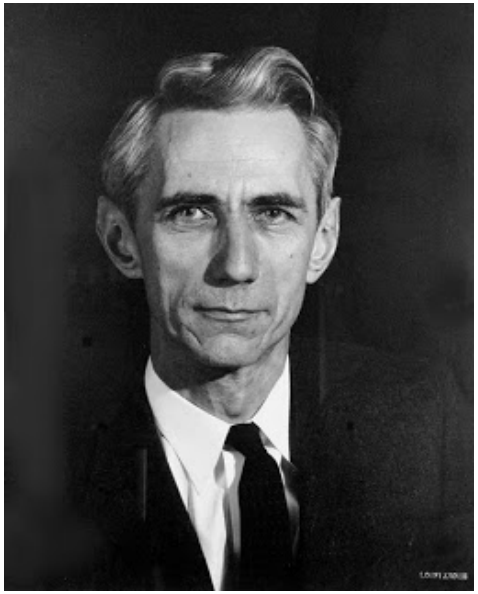
Information et quantité d'information

Vincent Vajnovszki

Sleon une édtue de l'Uvinertisé de Cmabrigde, l'odrre des ltteers dnas un mot n'a pas d'ipmrotncae, la suele coshe ipmrotnate est que la pmeirère et la drenèire lteetrs sinoet à la bnnoe pclae. Le rsete peut êrte dnas un dsérorde ttoal et vuos puoevz tujoruos lrie snas porblmèe. C'est prace que le creaveu hmauin ne lit pas chuaqe ltetre elle-mmêe, mias le mot cmome un tuot.

La théorie de l'information se préoccupe des systèmes de communication et de leur efficacité.



Ce domaine trouve son origine scientifique avec Claude Shannon qui en est un peu le père fondateur par son article *A Mathematical Theory of Communications* publié en 1948.

Parmi les branches importantes, on peut citer :

- le codage de l'information,
- codes correcteurs d'erreurs,
- la mesure quantitative de redondance d'un texte,
- la compression de données,
- la cryptographie.

L'information est-elle :

- un concept physique comme l'énergie ou la masse,

L'information est-elle :

- un concept physique comme l'énergie ou la masse, OU
- un concept purement mathématique qui n'a de sense que pour des êtres doués d'intelligence et qui n'intervient directement dans aucune loi proprement physique ?

Plusieurs scientifiques la considèrent même comme une grandeur physique fondamentale.

- Rolf Landauer (IBM) : *l'information est physique et elle doit avoir la même place, dans l'analyse du monde qui nous entoure, que la matière, l'énergie, l'espace et le temps.*
- Gilles Cohen-Tannoudji : *il faut introduire une quatrième constante fondamentale, une constante informationnelle ayant le même statut que la vitesse de la lumière c, la constante de gravitation G et la constante de Planck h pour rendre compte de l'importance de la notion d'information en physique théorique*.

La théorie de l'information a ainsi orienté l'évolution de certaines disciplines scientifiques.

- La biologie moléculaire qui a été dominée dans la seconde moitié du XXe siècle par des approches centrées sur l'information génétique,
- Les sciences humaines et sociales, des théories comme le structuralisme,
- Certaines conceptions néolibérales de l'économie s'appuient elles aussi sur une conception essentialiste de l'information.

Il faut moins de bits pour écrire *chien* que *mammifère*.
Pourtant l'indication Médor est un chien contient bien plus d'information que l'indication Médor est un mammifère : le contenu d'information sémantique d'un message dépend du contexte. En fait, c'est le couple message + contexte qui constitue le véritable porteur d'information.
Un évènement avec peu de probabilité représente beaucoup d'information
Exemple : Il neige en janvier contient beaucoup moins d'information que Il neige en août pour peu que l'on soit dans l'hémisphère nord.

Définition Une information désigne un ou plusieurs événements possibles parmi un ensemble fini d'événements. L'information permet de diminuer l'incertitude.

Exemple Considérons par exemple une source qui peut produire trois symboles *a*, *b* et *c*. Quand le destinataire attend un symbole, il est dans l'incertitude quant au symbole que la source va engendrer. Lorsque le symbole apparaît et qu'il arrive au destinataire, cette incertitude diminue.

Le but de la théorie de l'information est de mesurer cette incertitude avant réception.

Exemple On recherche une lettre dans une boîte. Si on précise que la lettre se trouve dans une enveloppe **bleue**, on fournit une information qui diminuera le temps de recherche du fait que le nombre de lettres dans des enveloppes bleues est plus restreint.
Si on ajoute l'information que la lettre dans une **grande** enveloppe, on pourra abréger d'autant plus le temps de la recherche.

Quantité d'information

Définition La quantité d'information est définie par

$$\log_2\left(\frac{N}{n}\right)$$

- N est le nombre d'événements possibles
- n est le cardinal du sous-ensemble dénoté par l'information

qui est exprimé en *logon*.

Observation La quantité d'information est une fonction croissante.

Exemple Dans une boîte il y a $N = 1050$ lettres dont

- $n_1 = 500$ lettres en enveloppes bleues
- $n_2 = 250$ en grandes enveloppes
- $n_3 = 40$ en grandes enveloppes bleues

L'information :

- *la lettres est dans une enveloppe dans la boîte* vaut
  $\log_2\left(\frac{N}{N}\right) = \log_2(1) = 0$
- *la lettres est dans une enveloppe bleue* vaut
  $\log_2\left(\frac{N}{n_1}\right) = \log_2(\frac{1050}{500}) = 1,07$
- *la lettres est dans une grande enveloppe* vaut
  $\log_2\left(\frac{N}{n_2}\right) = \log_2(\frac{1050}{250}) = 2,07$
- *la lettres est dans une grande enveloppe bleue* vaut
  $\log_2\left(\frac{N}{n_3}\right) = \log_2(\frac{1050}{40}) = 3,64$

0

1 3 5 7
9 11 13 15
17 19 21 23
25 27 29 31

1

2 3 6 7
10 11 14 15
18 19 22 23
26 27 30 31

2

4 5 6 7
12 13 14 15
20 21 22 23
28 29 30 31

3

8 9 10 11
12 13 14 15
24 25 26 27
28 29 30 31

4

16 17 18 19
20 21 22 23
24 25 26 27
28 29 30 31

Entropie

L'entropie nous permet de mesurer la quantité moyenne d'information contenue dans un ensemble de messages et de mesurer l'incertitude.
Soit X un ensemble partitionné en n sous-ensembles X_i, $1 \leq i \leq n$ (les messages), avec

$$X = \bigcup_{i=1}^{i=n} X_i$$

Par définition, la quantité d'information liée à chaque message de X_i est

$$I(X_i) = \log_2 \left(\frac{|X|}{|X_i|} \right) = \log_2 \left(\frac{N}{n_i} \right).$$

Définition L'entropie d'une *partition* est

$$H(partition) = \sum_{i=1}^{i=n} \frac{n_i}{N} \log_2 \left(\frac{N}{n_i} \right)$$

Observation Si

$$p_i = \frac{n_i}{N}$$

est la probabilité de l'apparition d'un message en X_i

$$H(partition) = -\sum_{i=1}^{i=n} p_i \log_2 (p_i)$$

Du fait que les X_i forment une partition de X,

$$\sum_{i=1}^{i=n} p_i = 1,$$

et l'entropie correspond à la distribution de probabilité de tous les messages possibles.

Exemple Soit une urne content $N = 100$ boules dont x blanches et $100 - x$ boules noires. On considère l'expérience qui consiste à tirer une boule.

- $I(b)$=quantité d'information liée à l'apparition d'une boule blanche
- $I(n)$=quantité d'information liée à l'apparition d'une boule noire
- H=quantite d'information moyenne par expérience = l'entropie

$$H = \frac{x}{100} \log_2 \left(\frac{100}{x} \right) + \frac{100 - x}{100} \log_2 \left(\frac{100}{100 - x} \right)$$

x	$100 - x$	$I(b)$	$I(n)$	H
50	50	1	1	1
40	60	1,32	0,73	0,97
1	99	6,64	0,014	0,080

• Quand un seul choix est possible (tous les p_i sont nuls, sauf un), l'incertitude est nulle; l'information apportée en spécifiant ce choix est donc elle aussi nulle.

• Dans le cas où les probabilités sont toutes égales (à p), on a

$$H = \log_2(1/p),$$

cela correspond à une entropie ou une incertitude maximale, et l'information apportée en spécifiant le choix effectué est maximale.

Claude Shannon

En 1948, l'Américain Claude Shannon publie l'article A Mathematical theory of communication.
Cette théorie mathématique de la communication a changé d'identité au fil du temps, puisqu'en français elle est connue sous le nom de théorie de l'information.

Shannon avait élaboré son modèle pour résoudre un problème de communication téléphonique (augmenter la vitesse de transmission des données et en diminuer les pertes). Dans ce modèle, la communication est définie en termes de transmission d'information, quelle qu'elle soit.

Thermodynamique

Rappelons que l'entropie est une grandeur fondamentale de la thermodynamique, science élaborée au XIXe siècle *le siècle de la Révolution industrielle*, qui permet de quantifier la part de l'énergie d'un système transformable en énergie mécanique.

Avec les travaux du physicien autrichien Ludwig Boltzmann, à la fin du XIXe siècle, l'entropie a pu être interprétée dans le cadre de la physique statistique comme une mesure du désordre d'un système.

Plus précisément, l'entropie (statistique) mesure le nombre d'états microscopiques auxquels peut avoir accès le système considéré, compte tenu des contraintes macroscopiques telles que l'énergie totale et le volume (un état microscopique est par exemple caractérisé par les vitesses et positions à un instant donné de chacun des atomes dont est constitué le système).

Codage pour un canal non bruité

Afin de transmettre un message sous la forme de signal, il faut le coder; c'est ce que l'on nomme également codage de source.

- écriture
- parole
- code Morse
- code ASCII
- unicode
- codes binaires i.e. sur l'alphabet $\{0, 1\}$

Soit Σ un alphabet et Σ^* l'ensemble des mots finis sur cet alphabet. Un code $C = \{c_1, ..., c_k\}$ est un sous-ensemble de Σ^*. Les éléments $c_i \in C$ sont appelés les mots du code.

Si

- tous les mots du code sont de même longueur, on dit que C est un code de longueur fixe ou code en bloc
- dans le cas contraire, C est un code à longueur variable
- si aucun mot du code n'est le préfixe d'un autre C est appelé préfixe

Exemple

- $C = \{0, 10, 11\}$ code préfixe de longueur variable
- $C = \{1, 01, 11\}$ code non-préfixe de longueur variable
- $C = \{000, 101\}$ code de longueur fixe

Codes optimaux

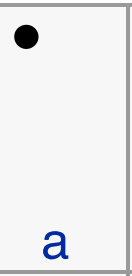
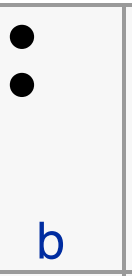
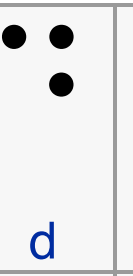
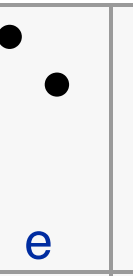
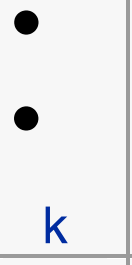
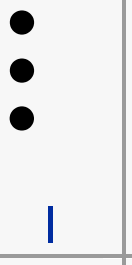
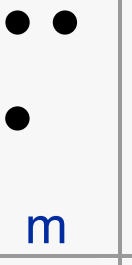
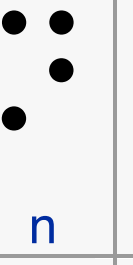
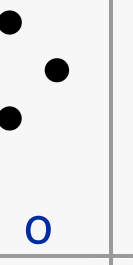
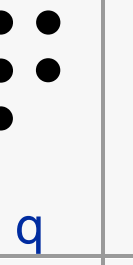
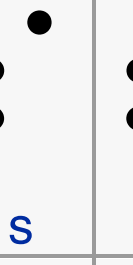
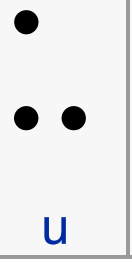
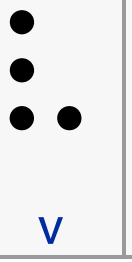
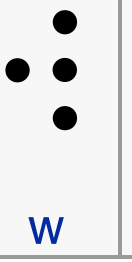
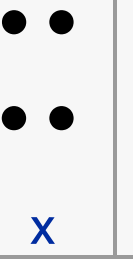
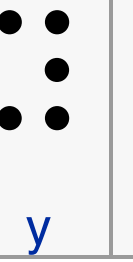
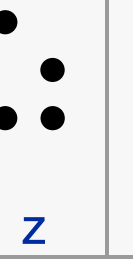
En associant des codes courts aux messages les plus fréquents et des codes longs aux messages les moins fréquents, on peut construire un code optimal : au sens où le nombre moyen de bits par symbole correspond précisément à l'entropie de l'ensemble des messages possibles.

Code de Morse

lettre	fréquence en français	fréquence en anglais	alphabet de Morse
a	6.16	8.05	·—
b	0.40	1.62	— ···
c	5.35	3.20	— · —·
d	3.86	3.65	— ··
e	18.61	12.31	·
f	2.24	2.28	·· —·
g	1.79	1.61	— — ·
h	1.48	5.14	····
i	6.35	7.18	··
j	0.04	0.10	· — — —
k	0.13	0.52	— · —

lettre	fréquence en français	fréquence en anglais	alphabet de Morse
l	5.26	4.03	· — ··
m	1.79	2.25	— —
n	6.02	7.19	—·
o	5.12	7.94	— — —
p	2.92	2.28	· — —·
q	0.62	2.29	— — ·—
r	5.35	6.03	· — ·
s	6.96	6.59	· · ·
t	7.41	9.59	—
u	5.03	3.10	· · —
v	1.03	0.93	· · ·—
w	0.20	0.25	· — —
y	1.39	1.88	— · — —
z	0.04	0.09	— — ··

Code de Braille

 a	 b	 c	 d	 e	 f	 g	 h	 i	 j
 k	 l	 m	 n	 o	 p	 q	 r	 s	 t
 u	 v	 w	 x	 y	 z				

Codage de Huffman

Huffman a proposé un algorithme qui construit un code préfixe optimal

Longueur moyenne, efficacité et redondance

Soit $A = \{a_1, a_2, \ldots, a_k\}$ un alphabet de source tel que

- p_i est la probabilité d'apparition du symbole a_i
- $a_i \rightarrow c_i$ est un code où c_i est un mot de longueur ℓ_i,

Définition La longueur moyenne du code est défini comme:

$$L = \sum_{i=1}^{k} p_i \ell_i$$

qui correspond à la somme pondérée des longueurs de tous les mots.

La longueur moyenne coïncide avec le rapport entre

- le nombre de symboles binaires du message codé, et
- le nombre de symboles de source

quand le message est suffisamment long pour que tous les symboles apparaissent avec une fréquence relative égale à leur probabilité.

Exemple

alphabet source	code	probabilité
a	0	$\frac{1}{2}$
b	10	$\frac{1}{4}$
c	11	$\frac{1}{4}$

- La longueur moyenne est

$$L = \sum_{i=1}^{3} p_i \ell_i = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = 1.5$$

- L'entropie est

$$H = \sum_{i=1}^{k} p_i \log_2(\frac{1}{p_i}) = \frac{1}{2} \cdot \log_2(2) + \frac{1}{4} \cdot \log_2(4) + \frac{1}{4} \cdot \log_2(4) = 1.5$$

Efficacité et redondance

Définition On définit l'*efficacité* comme le rapport

$$\eta = \frac{H}{L}$$

et la *redondance*

$$r = 1 - \eta$$

Exemple Soit l'alphabet de source $A = \{a, b\}$ tel que

alphabet source =symbole	probabilité	code
a	$p_a = 0,8$	0
b	$p_b = 0,2$	1

- L'entropie de la source est de

$$-0,8 \cdot \log_2(0,8) - 0,2 \cdot \log_2(0,2) = 0,72$$

- sa longueur moyenne est

$$0,8 \cdot 1 + 0,2 \cdot 1 = 1$$

- avec l'efficacité de $\eta = 0.72$ et la redondance de $r = 0.28$

Comment peut-on diminuer cette redondance et améliorer l'efficacité? Une idée est de coder des couples de symboles au lieu des symboles eux-mêmes.

symbole	probabilité	code
$aa = \alpha$	$p_{aa} = 0,64$	0
$ab = \beta$	$p_{ab} = 0,16$	11
$ba = \gamma$	$p_{ba} = 0,16$	100
$bb = \delta$	$p_{bb} = 0,04$	101

- L'entropie de la source est de

$$0,64 \cdot \log_2 \left(\frac{1}{0,64} \right) +$$

$$2 \cdot \left(0,16 \cdot \log_2 \left(\frac{1}{0,16} \right) \right) +$$

$$0,04 \cdot \log_2 \left(\frac{1}{0,04} \right) = 1,45$$

- sa longueur moyenne est

$$1 \cdot 0,64 + 2 \cdot 0,16 + 3 \cdot (0,16 + 0,04) = 1,56$$

- avec l'efficacité de $\eta = \frac{1,45}{1,56} = 0.93$ et la redondance de $r = 0.07$

En revanche, le coût à payer est une complexification des opérations de codage et de décodage (Arbre de Huffman).
On montre que, en faisant croître le nombre de symboles du'un code, l'efficacité du codage peut devenir aussi proche que possible de sa limite supérieure (égale à 1).
C'est précisément la signification du premier théorème de Shannon.
Théorèmes 1 de Shannon Théorème du codage de source: Sans perturbation, il est possible, à partir d'un alphabet quelconque, de coder les messages émis de telle sorte que le rendement soit aussi proche que souhaité de la valeur maximale, i.e. la capacité du canal.

Théorie algorithmique de l'information

ou théorie de la complexité de Chaitin-Kolmogorov-Solomonoff née dans la décennie 1960, à la frontière de la logique mathématique et de l'informatique théorique.

- Kolmogorov
- Chaitin
- Solomonoff

On veut transmettre en Australie deux types de renseignements :

1. les heures de lever et de coucher du soleil à Paris depuis dix ans;
2. les résultats du Loto à Paris tous les jours depuis dix ans.

Pour transmettre le premier résultat, il suffit d'une petite formule avec des sinus et des cosinus. Tandis que pour transmettre les résultats du Loto, depuis dix ans, il n'y a pas d'autre moyen que d'envoyer la liste. Personne ne trouve jamais de formule pour simplifier la tache.
La complexité d'une information, c'est la longueur du plus petit programme informatique qui la génère.
Une suite comme celle des heures de lever et de coucher du soleil est plus simple qu'une suite comme celle des résultats du Loto.

La complexité de Kolmogorov (parfois nommée complexité algorithmique), est une fonction permettant de quantifier la taille du plus petit algorithme nécessaire pour engendrer une suite de caractères. Cette quantité peut être vue comme une évaluation de la complexité de cette suite de caractères.

Profondeur logique

Les objets aléatoires ont le plus grand contenu en information : pour les mémoriser, rien n'est sensiblement meilleur que les stocker non modifiée. Puisque les objets aléatoires ont la plus grande complexité de Kolmogorov, cela signifie que cette notion de complexité ne mesure pas la richesse en structures d'un objet.

Profondeur logique

Les objets aléatoires ont le plus grand contenu en information : pour les mémoriser, rien n'est sensiblement meilleur que les stocker non modifiée. Puisque les objets aléatoires ont la plus grande complexité de Kolmogorov, cela signifie que cette notion de complexité ne mesure pas la richesse en structures d'un objet.
Charles Bennett a introduit (1988) le concept de profondeur logique.

Profondeur logique

Les objets aléatoires ont le plus grand contenu en information : pour les mémoriser, rien n'est sensiblement meilleur que les stocker non modifiée. Puisque les objets aléatoires ont la plus grande complexité de Kolmogorov, cela signifie que cette notion de complexité ne mesure pas la richesse en structures d'un objet.
Charles Bennett a introduit (1988) le concept de profondeur logique.
profondeur logique= le nombre de pas de calcul qu'il faut pour reconstituer l'objet *Ob* à partir de la représentation compressée optimale de *Ob*.

Plus un objet est structuré, plus il offre des moyens de le compresser, moyens qui, quand on les fait marcher à l'envers (la décompression), exigent de nombreux pas de calcul.

Plus un objet est structuré, plus il offre des moyens de le compresser, moyens qui, quand on les fait marcher à l'envers (la décompression), exigent de nombreux pas de calcul.
La richesse en structures ou complexité organisée est donc assez bien mesurée par le temps de calcul nécessaire pour passer de la forme compressée optimale de *Ob* à sa forme explicite.

La distance informationnelle entre deux suites A et B est définie comme la taille du plus court programme permettant de transformer A en B et B en A