Introduction à la théorie de l'information

Introduction à la théorie de l'information

Par Philippe Dubois, Zurich

Résumé

L'auteur expose les éléments de la théorie de l'information et traite quelques exemples simples d'application.

A l'origine de la théorie mathématique de l'information, créée par l'ingénieur américain Shannon [1] [1]) dans les années 1947/48, se trouvent des problèmes purement techniques de transmission de messages sous forme de signaux à l'aide de procédés radio-électriques. Etroitement apparentée au calcul des probabilités, la théorie de l'information a déjà été appliquée avec succès dans de nombreux domaines de la physique et de la technique (perfectionnement des machines à calculer électroniques, problèmes du codage dans les télécommunications, processus fondamental de l'observation scientifique, etc.). A l'heure actuelle, elle constitue le fondement d'une branche plus générale de la science connue sous le nom de *cybernétique* dont le champ d'application, grâce aux travaux fondamentaux du mathématicien américain Wiener, s'étend jusqu'à la psychologie (étude du comportement normal ou pathologique du système nerveux et, en particulier, analyse du mécanisme des réflexes) et même jusqu'à la biologie.

Le présent article se propose d'exposer d'une manière élémentaire les bases de la théorie de l'information et d'aborder quelques exemples simples d'application. Afin de faciliter la compréhension et la présentation des idées développées, la rigueur mathématique a été sacrifiée au profit de raisonnements empiriques et intuitifs. Le lecteur soucieux de s'orienter sur les essais entrepris récemment par de nombreux auteurs

[1]) Les chiffres entre crochets se rapportent à la liste bibliographique à la fin du présent article.

en vue de donner à cette théorie une base rigoureusement mathématique et même axiomatique, pourra se reporter de préférence à l'article [10] résumant les travaux publiés à ce sujet jusqu'à ce jour, et à l'ouvrage [5].

I. Eléments de la théorie de l'information

1. Remarques préliminaires [4]

La théorie mathématique de l'information part d'une définition précise et objective de la notion d'information et ne tient pas compte de la valeur subjective que peut revêtir une information. Plus exactement, l'information est considérée comme une grandeur absolue qui a la même valeur numérique pour tout observateur.

Ainsi, par exemple, l'information obtenue en choisissant une carte dans un jeu de cartes sera représentée par la même grandeur, que la carte tirée soit un as, un sept ou un roi. La valeur de la carte ainsi que la stratégie adoptée par les joueurs (éléments qui dépendent nécessairement des règles du jeu pratiqué) sont des notions qui relèvent de la théorie moderne des jeux et non de la théorie de l'information.

En fait, la définition de la notion d'information est fondée sur le critère de la rareté. Si une situation est rare, sa réalisation fournit de l'information, que cette information soit ou non dénuée de valeur.

L'interprétation restreinte donnée au mot «information» peut paraître sévère, mais elle répond, par exemple, aux préoccupations de l'ingénieur des télécommunications qui doit pouvoir transmettre toute information contenue dans un message sans se soucier de la valeur que présente cette information pour le destinataire. Seule l'élimination de tout élément subjectif a permis de donner une définition quantitative de l'information et de traiter cette dernière comme une grandeur mesurable. L'utilité d'une telle définition s'est affirmée d'une manière probante dans l'étude de nombreux problèmes scientifiques et techniques en permettant de parvenir à des conclusions d'ensemble de réelle valeur pratique et d'une grande généralité.

Ces remarques montrent clairement les limites de la théorie de l'information et elles doivent être présentes à l'esprit dans les applications.

2. Notion d'entropie d'une expérience [1] [4] [5] [7]

La définition de la notion d'information dérive de la notion d'entropie introduite par Shannon pour caractériser le degré d'indétermination d'une expérience. Considérons l'exemple simple suivant:

Soit une expérience α susceptible de prendre m valeurs A_i distinctes également probables *a priori*. Pour $m = 1$, l'expérience ne présente aucune indétermination, puisque le résultat de l'expérience est connu d'avance. Si m est grand, l'observateur ne disposant d'aucun élément d'information supplémentaire sur l'expérience α, ne sera en général pas en mesure de prédire l'issue de l'expérience envisagée. Plus grand sera m, plus grand sera le degré d'indétermination de l'expérience considérée. Le degré d'indétermination se présente donc comme une fonction $f(m)$ croissant avec m et s'annulant pour $m = 1$. Pour déterminer cette fonction complètement, il convient de lui imposer, en plus de la condition triviale de continuité, une propriété d'additivité.

Soient deux expériences indépendantes α et β, la première pouvant prendre comme précédemment m valeurs A_i et la seconde n valeurs B_j distinctes également probables *a priori*. Considérons l'expérience couplée $\alpha\beta$ consistant à réaliser simultanément les expériences α et β. Dans ces conditions, il est naturel d'admettre que l'indétermination de l'expérience couplée $\alpha\beta$ sera égale à la somme des indéterminations caractérisant les expériences α et β. Chaque valeur de la première expérience pouvant être couplée avec toute valeur de la seconde expérience, le nombre total de valeurs possibles également probables *a priori* de l'expérience $\alpha\beta$ sera donc égal à mn. Il en résulte la condition suivante pour la fonction $f(m)$:

$$f(mn) = f(m) + f(n).$$

Les conditions posées suffisent à déterminer complètement la structure de la fonction $f(m)$ qui est du type logarithmique. La mesure du degré d'indétermination d'une expérience susceptible de prendre m valeurs A_i distinctes également probables *a priori* s'exprime ainsi par la fonction

$$f(m) = \log m \qquad \text{avec } f(1) = 0.$$

Le choix de la base des logarithmes n'est pas essentiel, car le passage d'un système de logarithmes à un autre revient à multiplier la fonction $\log m$ par une constante égale au module du changement de base, c'est-à-dire à choisir une autre unité pour la mesure du degré d'indétermination d'une expérience.

Dans les applications techniques, il est d'usage de choisir les logarithmes de base 2. L'unité d'indétermination est appelée *unité binaire* ou *bit* et caractérise le degré d'indétermination d'une expérience présentant deux issues possibles également probables *a priori* (par exemple 0 et 1 dans le système de numération binaire). Le choix des logarithmes de base 10 conduit à l'*unité décimale* qui est environ $\frac{10}{3}$ plus grande que l'unité binaire (en effet: $\log_2 10 = 3{,}32$).

La fonction $\log m$ qui représente en définitive l'indétermination totale de l'expérience α peut aussi se mettre sous la forme

$$\log m = \sum \frac{1}{m} \log m = -\sum \frac{1}{m} \log \frac{1}{m},$$

la sommation portant sur m termes et $\frac{1}{m}$ désignant la probabilité que l'une quelconque des m issues possibles A_i également probables *a priori* soit sélectionnée. Dans la somme ci-dessus, chaque terme $-\frac{1}{m} \log \frac{1}{m}$ représente en quelque sorte l'indétermination engendrée par chacune des m issues possibles de l'expérience α.

La définition de la mesure d'indétermination relative à une expérience susceptible de prendre m valeurs distinctes également probables *a priori*, peut être immédiatement généralisée au cas d'une expérience α pouvant prendre m valeurs distinctes A_i avec une probabilité $p(A_i)$. L'expression obtenue dans le cas particulier permet de présumer que la mesure de l'indétermination de l'expérience α généralisée sera représentée par l'expression

$$H(\alpha) = -\sum_{i=1}^{m} p(A_i) \log p(A_i) \quad \text{avec} \quad \sum_{i=1}^{m} p(A_i) = 1.$$

Cette grandeur $H(\alpha)$ positive joue un rôle analogue à celui de la notion d'entropie en thermodynamique et a été, pour cette raison, appelée par Shannon l'*entropie* de l'expérience α. Elle peut aussi être interprétée comme la valeur probable d'une grandeur aléatoire qui prend la valeur $-\log p(A_i)$ avec la probabilité $p(A_i)$.

3. Propriété fondamentale de l'entropie $H(\alpha)$ [1] [4] [7]

Soit α_0 une expérience pouvant prendre m valeurs distinctes A_i également probables *a priori* et α une expérience susceptible de prendre aussi m valeurs distinctes A_i, mais avec une probabilité $p(A_i)$. A l'aide des propriétés élémentaires des fonctions convexes, il est facile de démontrer l'inégalité suivante

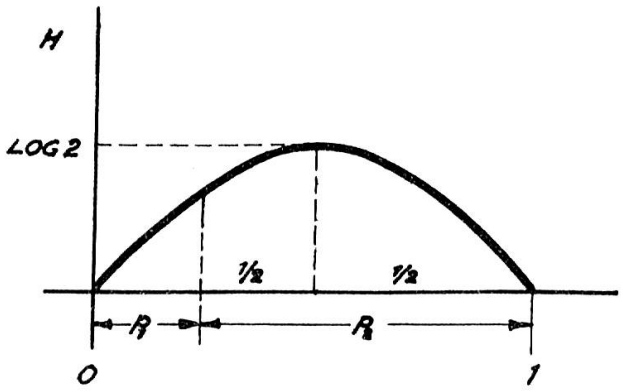
$$H(\alpha) \leqq H(\alpha_0) = \log m.$$

Le signe d'égalité ne peut avoir lieu que si $\alpha = \alpha_0$, c'est-à-dire si $p(A_i) = \frac{1}{m}$. Ce résultat est plausible, car l'issue de l'expérience α_0 est plus difficile à prévoir que celle de l'expérience α. Dans le cas où l'expérience α ne présente que deux issues possibles, la démonstration de l'inégalité ci-dessus est immédiate. En effet:

L'entropie d'une telle expérience

$$\begin{aligned} H(\alpha) &= -p_1 \log p_1 - p_2 \log p_2 \\ &= -p_1 \log p_1 - (1-p_1) \log (1-p_1) \quad (\text{car } p_1 + p_2 = 1) \end{aligned}$$

est une fonction de p_1 qui s'annule pour $p_1 = 0$ et $p_1 = 1$.



Représentation graphique de l'entropie d'une expérience α présentant deux issues possibles

En raison de sa structure symétrique (voir figure), la fonction $H(\alpha)$ admet pour $p_1 = \frac{1}{2} = p_2$ une valeur maximum égale précisément à

$$H(\alpha) = -\log \tfrac{1}{2} = \log 2 .$$

4. Entropie couplée et entropie conditionnelle [1] [4] [7]

Soient deux expériences α et β caractérisées par les tableaux de probabilités suivants:

$$\alpha = \begin{Bmatrix} A_i \\ p(A_i) \\ i = 1, 2, \ldots, m \end{Bmatrix}, \qquad \beta = \begin{Bmatrix} B_j \\ p(B_j) \\ j = 1, 2, \ldots, n \end{Bmatrix}.$$

Considérons l'expérience couplée définie par le tableau de probabilités:

$$\alpha\beta = \begin{Bmatrix} A_i B_j \\ p(A_i B_j) \\ i = 1, 2, \ldots, m \\ j = 1, 2, \ldots, n \end{Bmatrix}.$$

L'*entropie couplée* de l'expérience composée $\alpha\beta$ qui consiste à réaliser simultanément les expériences α et β, est définie par l'expression

$$H(\alpha\beta) = -\sum_{i,j=1}^{m,n} p(A_i B_j) \log p(A_i B_j),$$

où $p(A_i B_j)$ désigne la probabilité des issues A_i et B_j.

Si les expériences α et β sont indépendantes, c'est-à-dire si

$$p(A_i B_j) = p(A_i)\, p(B_j),$$

il est facile de démontrer que

$$H(\alpha\beta) = H(\alpha) + H(\beta).$$

En effet:

$$\begin{aligned} H(\alpha\beta) &= -\sum_{i,j=1}^{m,n} p(A_i)\, p(B_j) \left[\log p(A_i) + \log p(B_j)\right] \\ &= H(\alpha) \sum_{j=1}^{n} p(B_j) + H(\beta) \sum_{i=1}^{m} p(A_i) \\ &= H(\alpha) + H(\beta), \end{aligned}$$

car

$$\sum_{j=1}^{n} p(B_j) = \sum_{i=1}^{m} p(A_i) = 1.$$

Si les expériences α et β ne sont pas indépendantes, c'est-à-dire si

$$p(A_i B_j) = p(A_i)\, p(B_j | A_i),$$

où $p(B_j|A_i)$ désigne la probabilité de l'issue B_j lorsque l'issue A_i s'est produite, l'égalité ci-dessus n'est plus valable. D'une manière générale, on démontre que

$$H(\alpha\beta) \leqq H(\alpha) + H(\beta),$$

le signe d'égalité n'étant valable que si les expériences α et β sont précisément indépendantes.

L'expérience couplée $\alpha\beta$ impose en quelque sorte une contrainte supplémentaire à la réalisation des expériences α et β envisagées séparément. Dans ces conditions, il est logique que l'indétermination de l'expérience couplée $\alpha\beta$ soit inférieure ou au plus égale à la somme des indéterminations caractérisant les expériences α et β considérées individuellement.

L'inégalité ci-dessus peut être remplacée par une égalité d'une portée plus générale. A cet effet, Shannon a introduit la notion d'*entropie conditionnelle* $H(\beta|\alpha)$ qui est définie par les relations

$$H(\beta|\alpha) = \sum_{i=1}^{m} p(A_i)\, H(\beta|A_i)$$

avec

$$H(\beta|A_i) = -\sum_{j=1}^{n} p(B_j|A_i) \log p(B_j|A_i),$$

où $p(B_j|A_i)$ désigne comme précédemment la probabilité liée de l'événement B_j quand on sait que l'événement A_i s'est déjà réalisé. La grandeur $H(\beta|\alpha)$ représente donc l'entropie de l'expérience β liée à la réalisation de l'expérience α.

L'expression de l'entropie conditionnelle $H(\beta|\alpha)$ peut aussi se mettre sous la forme:

$$H(\beta|\alpha) = -\sum_{i,j=1}^{m,n} p(A_i B_j) \log p(B_j|A_i).$$

A partir de ces relations, on démontre aisément que l'entropie couplée $H(\alpha\beta)$ de l'expérience composée $\alpha\beta$ satisfait à l'égalité:

$$H(\alpha\beta) = H(\alpha) + H(\beta|\alpha).$$

Il en résulte immédiatement que

$$0 \leqq H(\beta|\alpha) \leqq H(\beta),$$

la limite supérieure n'étant valable que lorsque les expériences α et β sont indépendantes et la limite inférieure lorsque l'issue de l'expérience α détermine entièrement l'issue de l'expérience β. Dans ce dernier cas, on obtient:

$$H(\alpha\beta) = H(\alpha).$$

Soit l'expérience composée $\alpha = \alpha_1 \alpha_2 \ldots \alpha_k$ qui consiste à réaliser k expériences α_i. L'entropie $H(\alpha)$ satisfait à l'inégalité suivante qui est une généralisation du résultat indiqué précédemment dans le cas de deux expériences:

$$H(\alpha) \leqq H(\alpha_1) + H(\alpha_2) + \ldots + H(\alpha_k),$$

inégalité que nous aurons l'occasion d'utiliser dans la seconde partie du présent article.

5. Définition de la notion d'information [1] [4] [7]

Par définition, l'entropie $H(\beta)$ caractérise le degré d'indétermination de l'expérience β. $H(\beta)$ égal à zéro signifie que le résultat de l'expérience β est connu d'avance. Une expérience auxiliaire quelconque α (susceptible de présenter plusieurs issues possibles) précédant l'expérience β peut limiter le nombre des éventualités de cette dernière et ainsi diminuer son degré d'indétermination. Cette circonstance s'accorde avec le fait que l'entropie conditionnelle $H(\beta|\alpha)$ est plus petite ou au plus égale à l'entropie inconditionnelle $H(\beta)$ de l'expérience β.

Si le résultat de l'expérience auxiliaire α n'influe pas sur celui de l'expérience β, on obtient:

$$H(\beta|\alpha) = H(\beta).$$

Si, par contre, l'expérience α détermine complètement l'issue de l'expérience β, la relation suivante est valable:

$$H(\beta|\alpha) = 0.$$

Par conséquent, la différence

$$I(\alpha\beta) = H(\beta) - H(\beta|\alpha)$$

peut être interprétée comme une mesure indiquant de combien diminue l'indétermination de l'expérience β par la réalisation de l'expérience auxiliaire α. Cette mesure est appelée la *quantité d'information* que fournit l'expérience α au sujet de l'expérience β. Si, dans l'expression $I(\alpha\beta)$, l'expérience α coïncide avec l'expérience β, on obtient:

$$I(\beta\beta) = I(\beta) = H(\beta).$$

En effet, la réalisation de l'expérience β déterminant entièrement son issue, on a nécessairement:

$$H(\beta|\beta) = 0.$$

Il découle de ces considérations que la grandeur $H(\beta)$, définie à l'origine comme le degré d'indétermination de l'expérience β, peut être également utilisée pour mesurer la quantité d'information que recèle l'expérience β. En se reportant à la définition de l'entropie $H(\beta)$ sous sa forme générale, il apparaît que cette grandeur représente plus précisément l'information moyenne contenue dans l'expérience β (le caractère de valeur moyenne étant conditionné par les différentes issues

possibles de l'expérience β). Dans cet ordre d'idées, les notions d'entropie et d'information sont parfaitement identiques. Plus l'indétermination d'une expérience est grande, plus l'information obtenue par sa réalisation est grande. Ainsi, par exemple, si la probabilité p d'un événement est petite (décès d'une personne jeune), l'information obtenue par la réalisation de cet événement inattendu est sans aucun doute plus grande que si p est grand (décès d'une personne centenaire).

Il est facile de vérifier que

$$I(\alpha\beta) = I(\beta\alpha),$$

d'où résulte l'inégalité:

$$I(\alpha\beta) \leqq H(\alpha) = I(\alpha).$$

Cette inégalité est plausible, car l'information contenue dans une expérience α au sujet d'une autre expérience β ne peut être supérieure à l'information contenue dans l'expérience α envisagée en elle-même.

II. Applications de la théorie de l'information

Cette seconde partie est consacrée à l'application des notions d'entropie et d'information à la résolution de quelques problèmes simples provenant de différents domaines. Dans le cadre d'une introduction, il ne saurait être question de traiter dans toute sa généralité le problème fondamental du codage qui nécessite un appareil mathématique très avancé (analyse de Fourier, théorie des sondages et des filtres).

1. Problème de logique [7]

Soit un nombre entier quelconque x positif, inférieur ou égal à N. Combien faut-il poser de questions pour le deviner, celui qui a imaginé le nombre répondant par oui ou par non aux questions qui lui sont posées?

L'expérience β dont il s'agit de déterminer l'issue peut prendre N valeurs possibles, le nombre imaginé x étant l'un des nombres quelconques compris entre 1 et N (limites comprises). *A priori*, ces N valeurs peuvent être considérées comme également probables. L'expérience β contient donc une quantité d'information égale à

$$I(\beta) = \log N.$$

Considérons l'expérience auxiliaire composée

$$\alpha = \alpha_1 \alpha_2 \dots \alpha_k$$

qui consiste à poser k questions. L'expérience α_i n'ayant que deux issues possibles (réponse affirmative ou négative) contient une information maximum égale à

$$I(\alpha_i) = \log 2 .$$

En vertu de l'inégalité indiquée à la fin de la première partie (chiffre I), on obtient:

$$I(\alpha) \leqq k \log 2 .$$

Par ailleurs, on sait que

$$I(\alpha\beta) \leqq I(\alpha) .$$

Pour que l'expérience auxiliaire α détermine complètement l'issue de l'expérience β, il faut que

$$I(\alpha\beta) = I(\beta) .$$

De cette façon, nous obtenons l'inégalité suivante qui va permettre de déterminer le nombre k:

$$\log N \leqq k \log 2 = \log 2^k,$$

d'où il résulte que

$$2^k \geqq N,$$

c'est-à-dire

$$k \geqq \log_2 N = \frac{\log_{10} N}{\log_{10} 2},$$

le signe d'égalité n'étant valable que lorsque N est une puissance de 2.

Considérons, à titre d'exemple, le cas d'un nombre à deviner compris entre 1 et 10, limites incluses. Pour $N = 10$, l'inégalité ci-dessus donne

$$k \geqq 3{,}32.$$

Or k ne peut être qu'un nombre entier de sorte que le nombre minimum de questions nécessaires dans ce cas particulier est égal à

$$k = 4.$$

Montrons rapidement comment il est possible de déterminer effectivement un nombre x compris entre 1 et 10 à l'aide de quatre questions:

Partageons l'ensemble des nombres entiers compris entre 1 et 10 en deux parties égales. La première question permettra de situer le nombre x dans l'une de ces deux parties. La partie sélectionnée est à nouveau partagée en deux parts de trois et deux chiffres. A l'aide de la seconde question, il sera possible de déterminer dans quelle partie de ce sous-ensemble se trouve le nombre x. En continuant de cette manière, il est visible que le nombre x pourra être effectivement défini à l'aide de quatre questions au plus. Notons en passant que ces quatre questions suffiraient même à déterminer un nombre x compris entre 1 et 16, limites incluses.

Si l'interrogateur dispose d'une information préliminaire sur le nombre x, les N valeurs possibles ne peuvent plus être considérées comme également probables *a priori*. Le nombre minimum de questions qui permet de déterminer dans tous les cas le nombre x, sera néanmoins donné par l'inégalité établie précédemment. Le procédé appliqué dans l'exemple particulier ci-dessus pourra être maintenu et permettra d'arriver au but quelle que soit la répartition des probabilités attachées aux différentes valeurs possibles. Suivant les circonstances, il est cependant possible qu'une stratégie plus avantageuse soit concevable en vue de déterminer le nombre x plus rapidement. En cas d'échec de cette stratégie, il faut s'attendre alors à ce que le nombre de questions nécessaires devienne supérieur à la valeur k résultant de l'inégalité établie antérieurement. A ce point de vue, cette dernière valeur peut être considérée comme une valeur moyenne du nombre de questions indispensables pour déterminer le nombre x. Illustrons cet aspect du problème à l'aide de l'exemple simple suivant:

Soit $N = 4$ et x_1, x_2, x_3 et x_4 les valeurs possibles du nombre cherché. En vertu de l'inégalité définissant k, deux questions seront nécessaires pour déterminer complètement le nombre x, si le procédé de partage en sous-ensembles est adopté. Supposons que la valeur x_1 soit plus probable que les trois autres valeurs x_2, x_3 et x_4. Si la première question posée consiste à demander si $x = x_1$ et que la réponse soit affirmative, le problème aura été résolu au moyen d'une seule question. Si, par contre, la réponse est négative, une nouvelle question ne suffira plus en général à déterminer le nombre x. Dans le cas le plus défavorable, trois questions au total seront nécessaires. En fait, la stratégie la plus avantageuse à adopter dans ce jeu dépendra des probabilités attachées aux différentes valeurs x_i.

2. Problème élémentaire du codage [1] [4] [7]

Soit un message de N nombres, exprimé à l'aide des chiffres $(0, 1, 2, \ldots, 9)$ du système décimal. Dans le domaine des communications, le problème du codage le plus simple consiste à transcrire ce message donné dans le système de numération binaire, le chiffre 5, par exemple, étant représenté par le symbole 101. Quel est le nombre minimum de signes du système binaire nécessaires pour transmettre le message de N chiffres du système décimal?

Ce problème revient à rechercher le code le plus avantageux qui permette une transmission plus rapide du message et, par conséquent, une utilisation plus rationnelle de la ligne de transmission. Il est facile de se rendre compte que le nombre de signes binaires nécessaires pour représenter un nombre compris entre 0 et 9 (limites incluses) est exactement équivalent au nombre de questions qu'il faut poser pour deviner un chiffre compris entre 1 et 10 (voir développements sous 1.). Pour transmettre un nombre du système décimal, nous aurons donc besoin de 4 signaux binaires (le nombre 5 étant alors représenté par le symbole 0101). Un message de N nombres du système décimal nécessitera ainsi $4N$ signaux binaires. L'information contenue dans un nombre exprimé dans le système décimal est égale au plus à

$$\log 10 = 1 \text{ unité décimale,}$$

cette valeur maximum n'étant valable que si les dix valeurs possibles $(0, 1, \ldots, 9)$ sont indépendantes entre elles et également probables *a priori*. Un message de N chiffres en système décimal contiendra donc au maximum une information égale à

$$N \text{ unités décimales} = \tfrac{10}{3} N \text{ unités binaires,}$$

puisqu'une unité décimale vaut approximativement $\frac{10}{3}$ unités binaires.

Chaque élément du message codé peut prendre deux valeurs possibles (0 ou 1) et fournit donc au maximum une information égale à $\log 2$, soit une unité binaire. Un message composé de $4\ N$ signaux binaires fournira ainsi au maximum une information de

$$4N \text{ unités binaires,}$$

soit $\frac{2}{3} N$ unités binaires de plus que l'information maximum contenue dans un message de N nombres du système décimal. Il s'ensuit que le code envisagé (quatre signes binaires par chiffre du système décimal) n'est pas le plus avantageux du point de vue de la théorie

de l'information. Il est aisé de concevoir comment un code plus avantageux peut être construit. A cet effet, il suffit de partager le message de N chiffres du système décimal en tranches de deux chiffres et de coder chacune de ces tranches directement suivant le système de numération binaire. Le nombre de signaux binaires nécessaires pour transmettre une tranche de deux nombres du système décimal, comprise entre 00 et 99, est égal à 7 et correspond au nombre de questions qu'il faut poser pour deviner un nombre compris entre 1 et 100 (limites comprises). Un message de N nombres du système décimal nécessitera dans ces conditions (N étant supposé pair pour plus de simplicité)

$$7\,\frac{N}{2} = 3{,}5\,N \text{ signaux binaires,}$$

soit $\frac{N}{2}$ signaux binaires de moins que dans le procédé du codage envisagé initialement. Il serait théoriquement possible de réduire encore plus le nombre de signaux binaires en partageant le message de N nombres du système décimal en tranches de M chiffres et en transcrivant directement chacune de ces tranches suivant le système de numération binaire. Le nombre de signaux binaires nécessaires pour représenter une tranche de M chiffres du système décimal est égal à

$$k = \frac{\log_{10} 10^M}{\log_{10} 2},$$

où 10^M désigne le nombre de valeurs que peut prendre une tranche de M chiffres dans le système décimal. Dans ces conditions, le nombre de signaux binaires requis pour transcrire un message de N nombres du système décimal suivant le système de numération binaire à l'aide du procédé du codage par tranches, sera donné (N étant supposé divisible par M pour plus de simplicité) par l'expression:

$$\frac{N}{M}\,\frac{\log_{10} 10^M}{\log_{10} 2} = N\,\frac{\log_{10} 10}{\log_{10} 2} = 3{,}32\ldots N.$$

3. Problème du langage [1] [2] [4] [7]

La part la plus importante de l'information est transmise par le langage. Envisageons le problème du calcul de la quantité d'information que contient une phrase écrite. Ce problème d'une importance pratique très grande est compliqué. Le manque de données statistiques concernant le langage n'a pas encore permis de parvenir à une solution rigoureuse et complète de cette question.

Les résultats obtenus sous chiffre 2. peuvent être transposés au cas où le message ne se présente pas sous forme de chiffres, mais sous forme de lettres tirées d'un alphabet de n lettres par exemple. Si toutes les lettres de l'alphabet envisagé sont considérées comme également probables *a priori*, il sera possible de transmettre un message de N lettres, partagé en tranches de M lettres, à l'aide de

$$N \frac{\log_{10} n}{\log_{10} 2}$$

signaux binaires. Or il se trouve que l'hypothèse formulée n'est pas valable dans le domaine du langage. Ainsi, par exemple, dans un texte français quelconque, les lettres o ou e apparaissent plus souvent que les lettres y ou k. De telles anomalies se rencontrent dans toutes les langues.

Shannon et d'autres auteurs se sont penchés sur cet aspect du problème et ont étudié tout particulièrement la structure de la langue anglaise. Les résultats obtenus dans ce domaine sont exposés sommairement dans la suite.

L'alphabet de la langue anglaise se compose de 27 lettres (l'espace entre les mots étant assimilé à une lettre). Si toutes les lettres se présentaient dans un texte anglais avec la même probabilité, l'information contenue dans une lettre serait au plus égale à

$$I_0 = \log 27 = 1{,}431 \text{ unités décimales.}$$

Les statistiques établies à partir d'un texte anglais ayant un sens et suffisamment long ont permis de déterminer approximativement les probabilités de survenance des diverses lettres. L'extrait suivant donne une idée de l'ordre de grandeur de ces probabilités.

lettre i	probabilités de survenance p_i
espace entre les mots	0,200
E	0,105
T	0,072
I	0,055
D	0,035
P	0,018
K	0,003
Z	0,001

Si les lettres i de l'alphabet envisagé sont supposées indépendantes entre elles, l'information moyenne contenue dans une lettre transmise à partir d'un texte anglais sera alors donnée par

$$I_1 = H = -\sum_{i=1}^{27} p_i \log p_i = 1{,}213 \text{ unités décimales.}$$

Mais cette grandeur moyenne est bien supérieure à la valeur réellement transmise. En effet, dans un texte ayant un sens, la probabilité de transmettre une lettre donnée dépend essentiellement des lettres qui la précèdent. Ainsi, il est très probable que la lettre t soit suivie de la lettre h dans la langue anglaise. L'interdépendance des lettres dans toute langue est une forme particulière de la corrélation, notion fondamentale en statistique appliquée.

Soient deux expériences α_1 et α_2 qui consistent à étudier le comportement de deux lettres successives quelconques a_i et a_j, tirées d'un texte ayant un sens. L'information totale résultant de la réalisation de ces deux expériences dépendantes est donnée par la relation:

$$H(\alpha_1\alpha_2) = H(\alpha_1) + H(\alpha_2|\alpha_1) = H(\alpha_2|\alpha_1),$$

car on peut toujours admettre que la lettre a_i précédant la lettre a_j est connue à la réception du message (de sorte que $H(\alpha_1) = 0$). Dans ces conditions, l'information moyenne contenue dans une lettre transmise devra être calculée par la formule

$$I_2 = H(\alpha_2|\alpha_1) = -\sum_{i,j=1}^{n} p(a_i a_j) \log p(a_j|a_i),$$

où $p(a_i a_j)$ désigne la probabilité de survenance des lettres a_i et a_j, $p(a_j|a_i)$ la probabilité de survenance de la lettre a_j lorsqu'on sait qu'elle est précédée de la lettre a_i, et n le nombre de lettres de l'alphabet envisagé.

De même, si l'on considère un groupe de trois lettres successives quelconques a_i, a_j et a_k et si l'on tient compte de la corrélation liant la troisième lettre a_k aux deux premières a_i et a_j, l'information moyenne contenue dans une lettre transmise sera représentée par la formule

$$I_3 = H(\alpha_3|\alpha_1\alpha_2) = -\sum_{i,j,k=1}^{n} p(a_i a_j a_k) \log p(a_k|a_i a_j),$$

où les probabilités introduites dans cette expression ont des significations analogues à celles indiquées précédemment. La méthode de calcul

de l'information moyenne d'une lettre relative à un groupe de deux et trois lettres peut être immédiatement généralisée au cas d'un groupe de N lettres avec N quelconque.

Dans cet ordre d'idées, il convient de signaler que le problème de la corrélation posé par la transmission d'un texte ayant un sens est apparenté à l'étude de chaînes particulières de Markoff. Cette analogie a déjà été mise en évidence par Shannon dans son ouvrage fondamental [1] et tout récemment par Münzner [9].

En plus des probabilités de survenance des lettres simples (voir extrait ci-dessus), celles se rapportant aux groupes de deux et de trois lettres ont été calculées pour la langue anglaise. A partir de ces données, Shannon a déterminé les valeurs I_2 et I_3. Le tableau suivant résume les résultats obtenus:

$$I_0 = 1{,}431 \text{ unités décimales}$$
$$I_1 = 1{,}213 \text{ unités décimales}$$
$$I_2 = 0{,}999 \text{ unités décimales}$$
$$I_3 = 0{,}933 \text{ unités décimales.}$$

Si l'on envisage des groupes de lettres toujours plus importants et si l'on tient compte de la corrélation liant ces lettres, l'information moyenne contenue dans une lettre transmise décroîtra en conséquence. Les grandeurs I_m formeront ainsi une suite monotone décroissante et l'information moyenne réelle par lettre sera donnée par la limite théorique

$$I_\infty = \lim_{m \to \infty} I_m ,$$

lorsqu'un texte ayant un sens et suffisamment long est transmis. Cette valeur limite n'a pu être calculée exactement jusqu'à présent en raison du manque de données statistiques complètes pour les groupes de plus de trois lettres. En vue de son calcul approximatif, Brillouin, entre autres auteurs, a proposé une méthode fondée sur des statistiques relatives à la fréquence de rencontre d'un mot. De telles statistiques ont été établies pour la plupart des langues européennes dans l'intention de faciliter le codage de textes et l'enseignement des langues.

Pour caractériser l'interdépendance des lettres dans toute langue, Shannon a introduit la notion de *redondance* d'une langue qui est définie par l'expression

$$R = 1 - \frac{I_\infty}{I_0} = 1 - \frac{I_\infty}{\log_{10} n} ,$$

où n désigne comme par le passé le nombre de lettres de l'alphabet envisagé. Les recherches entreprises par Shannon et d'autres auteurs ont montré que la redondance R est de l'ordre de grandeur de 0,5 pour la langue anglaise. En d'autres termes, dans un texte anglais ayant un sens le 50% des lettres est déterminé par la structure même de la langue.

De l'application de ces considérations au problème du codage d'un texte, posé au début de la présente section, il résulte qu'un texte de N lettres (N suffisamment grand) ayant un sens peut être transmis en principe au moyen de

$$N \frac{I_\infty}{\log_{10} 2}$$

signaux binaires, nombre qui est notablement inférieur au nombre

$$N \frac{\log_{10} n}{\log_{10} 2},$$

valable lorsque les lettres de l'alphabet envisagé sont considérées comme également probables *a priori*. Un procédé de codage utilisant moins de $N \frac{I_\infty}{\log_{10} 2}$ signaux binaires serait théoriquement concevable, mais il provoquerait une perte d'information qui ne permettrait plus en général de rétablir la teneur du message transmis.

Les développements sommaires qui précèdent ainsi que les résultats présentés sous chiffre 2. sont à la base d'un théorème fondamental relatif à la vitesse de transmission des messages, dont la démonstration repose sur un procédé de codage approprié connu sous le nom de *code de Shannon-Fano*. Nous ne donnerons cependant pas l'énoncé de ce théorème, car il faudrait, au préalable, introduire la notion de *capacité* d'une ligne de transmission. Bien que cette notion soit étroitement liée à celle de l'information, elle sort du cadre du présent article. Le lecteur s'intéressant à ces questions fondamentales de la théorie des transmissions pourra se reporter aux ouvrages spécialisés indiqués dans la liste bibliographique ci-après.

Bibliographie

Ouvrages

[1] *Shannon and Weaver,* The mathematical theory of communication. Illinois Press, 1949.

[2] *Shannon,* Prediction and entropy of printed English. Bell System Tech. J. 30, 1951.

[3] *Mandelbrot,* Contribution à la théorie mathématique des jeux de communication. Institut de Statistique de l'Université de Paris, 1954.

[4] *Brillouin,* Science and information theory. Academic Press, 1956.

[5] *Feinstein,* Foundations of information theory. McGraw-Hill, 1958.

[6] *Booth, Brandwood, and Cleave,* Mechanical resolution of linguistic problems. Academic Press, 1959.

[7] *A. et J. Yaglom,* Probabilité et information. Dunod, 1959.

[8] *Brillouin,* Vie, matière et observation. Albin Michel, 1960.

Articles

[9] *Münzner,* Über einige Grundbegriffe der Informationstheorie. Blätter der deutschen Gesellschaft für Versicherungsmathematik, Band IV, Heft 3, 1960.

[10] *Schmetterer,* Literaturbericht zur Informationstheorie. Blätter der deutschen Gesellschaft für Versicherungsmathematik, Band IV, Heft 3, 1960.

Zusammenfassung

Der Verfasser legt die Grundbegriffe der Informationstheorie dar und behandelt einige einfache Anwendungsbeispiele.

Riassunto

L'autore expone gli elementi fondamentali della teoria dell'informazione e tratta alcuni esempi semplici di applicazione.

Summary

The paper deals with the elements of the theory of information. Its application is illustrated by some simple examples.