La théorie de l'information

Olivier RIOUL, professeur à Télécom ParisTech et à l'École Polytechnique

La révolution numérique que nous connaissons aujourd'hui doit énormément à la théorie de l'information de SHANNON. La question à la base de la théorie est toute naturelle : peut-on mesurer l'information, contenue dans un message ou transmise dans un canal de communication ? En se basant sur les probabilités et la notion d'entropie, la théorie décrit rigoureusement cette notion d'information pour résoudre des problèmes de compression et de transmission de données numériques et en trouver les limites fondamentales de performances.

Claude SHANNON (1916–2001) est un mathématicien et ingénieur américain haut en couleurs. Adepte du monocycle et du jonglage, il s'est amusé à construire des machines plus ou moins loufoques : une souris qui apprend et retrouve son chemin dans un labyrinthe, une machine à jouer aux échecs, à résoudre le Rubik's cube, une calculatrice en chiffres romains, un robot qui jongle avec trois balles, et même une « machine inutile » qui, dès qu'on l'allume, actionne une main pour s'éteindre elle-même… Dans le même temps, il a fait des avancées théoriques décisives dans le domaine des circuits logiques, de la cryptographie, de l'intelligence artificielle…

Mais surtout, SHANNON créé la théorie de l'information en 1948, dans un seul article – *A Mathematical Theory of Communication* – qui rassemble tellement d'avancées fondamentales et de coups de génie que SHANNON est aujourd'hui le héros de milliers de chercheurs. C'est le mathématicien dont les théorèmes ont rendu possible le monde du numérique que nous connaissons aujourd'hui.

Décrivons quelques unes de ses contributions les plus marquantes :

Le paradigme de communication. Dans le paradigme de la communication selon SHANNON, un message émis par une source d'information est transmise dans un canal bruité puis reçu par le destinataire (Fig. 1). Si cela peut paraître aujourd'hui naturel, ça ne l'était pas à cette époque : pour la première fois, on y distingue clairement les rôles de la source, du canal et du destinataire ; de l'émetteur et du récepteur ; et du signal et du bruit. Ce paradigme a eu un impact immense, jusqu'en psychologie, en linguistique ou en sciences sociales, si bien que dans les années 1950, SHANNON en vient lui-même à mettre en garde ses contemporains contre les dérives d'une telle popularité.

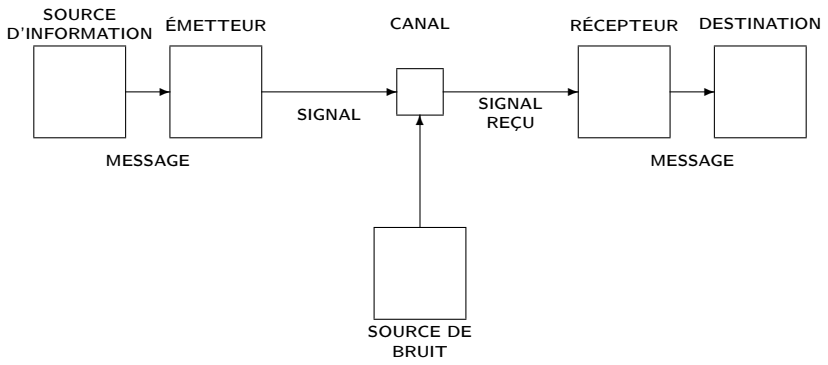

FIGURE 1 – Le paradigme de SHANNON.

L'aspect probabiliste. Laissant délibérément l'aspect sémantique de côté, SHANNON introduit pour la première fois un modèle probabiliste pour toutes les variables en jeu dans la communication et fonde ainsi sa théorie sur celle des probabilités, qui avait trouvé sa forme définitive quinze ans plus tôt avec les travaux d'Andreï KOLMOGOROV (1903–1987). KOLMOGOROV lui-même fut un ardent défenseur de la théorie de l'information, qui selon lui devait « précéder la théorie des probabilités », et non l'inverse.

L'unité d'information. SHANNON reprend l'idée exposée vingt ans auparavant par Ralph HARTLEY (1888–1970) d'une mesure logarithmique de l'information, en privilégiant l'unité binaire : puisqu'un chiffre binaire représente 2 symboles (0 et 1), deux chiffres 4 symboles, trois chiffres 8 symboles… représenter un message parmi N requiert donc $\log_2 N$ chiffres binaires, où $\log_2$ est le logarithme en base deux. SHANNON popularise à cette occasion le terme « bit » qu'il attribue à John TUKEY (1915–2000), comme contraction de *binary digit* (chiffre binaire). Mais le l'unité binaire d'information de SHANNON va plus loin que le simple chiffre binaire, car il prend en compte l'aspect *probabiliste* de l'information. Pour lui, un bit aléatoire peut très bien porter une information qui est en fait inférieure à un bit ! Aujourd'hui, l'unité officielle de mesure d'information s'appelle… le *Shannon* (sh).

Les limites de performances. SHANNON énonce et résout le problème *théorique* de la communication. Il ne propose quasiment aucune solution pratique, mais établit des *limites* de performances, ce qui est au moins aussi important. Avant SHANNON, des moyens de communication comme le télégraphe ont été développés pour ainsi dire dans le brouillard, sans le repère théorique permettant de savoir jusqu'où on peut aller. Depuis SHANNON, on sait que pour des ressources données, *quoique nous fassions*, le meilleur système de communication fiable ne pourra jamais dépasser une certaine limite sur le débit d'information. C'est comme si

SHANNON avait démontré la vitesse de la lumière sans dire comment construire la fusée qui pourrait s'en approcher. Cela a énormément stimulé la recherche de solutions pratiques permettant de s'approcher des limites de Shannon.

L'entropie. Un message se modélise comme une suite de variables aléatoires $X_1, X_2, X_3, \ldots$ correspondant chacune à un symbole qui peut prendre un nombre fini de valeurs. Si nous supposons que la source qui émet ce message est « sans mémoire », ces symboles sont choisis indépendamment les uns des autres et avec la même loi de probabilité. En notant $p(x)$ la probabilité qu'un symbole égale x, la probabilité d'un message donné $(x_1, x_2, \ldots, x_n)$ est le produit des probabilités individuelles :

$$p(x_1) \cdot p(x_2) \cdot \cdots\cdots \cdot p(x_n) = \prod_x p(x)^{n(x)}$$

où $n(x)$ est le nombre de symboles du message $(x_1, x_2, \ldots, x_n)$ qui sont égaux à x. Par la loi des grands nombres, quand la longueur n du message tend vers l'infini, le rapport $n(x)/n$ tend vers $p(x)$ si bien que la probabilité d'un message tiré au hasard vaut à peu près

$$\prod_x p(x)^{np(x)} = 2^{-nH}$$

où

$$H = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

est une quantité positive que SHANNON appelle l'*entropie* par analogie avec la notion étudiée en thermodynamique et en physique statistique. La petite histoire raconte que c'est John VON NEUMANN (1903–1957) qui recommande à SHANNON d'utiliser le terme « entropie » car lui dit-il, « personne ne sait vraiment ce qu'est l'entropie, de sorte qu'en cas de débat vous aurez toujours l'avantage. »

Le premier théorème de Shannon. La notion d'entropie permet à SHANNON de résoudre le problème théorique de la *compression* d'une source pour un canal sans bruit. Pour cela, il suffit de ne coder que les messages $(x_1, x_2, \ldots, x_n)$ typiques de probabilité $\approx 2^{-nH}$. En sommant les probabilités de ces N messages, on obtient une probabilité totale très proche de $1 \approx N \cdot 2^{-nH}$, d'où $N \approx 2^{nH}$, soit un débit de $(\log_2 N)/n \approx H$ bits par symbole. C'est le *premier théorème de Shannon* qui affirme que H bits par symbole suffisent pour comprimer fidèlement une source d'information. L'entropie H apparaît être une borne inférieure sur le débit nécessaire pour coder l'information de façon fiable.

Ce théorème est asymptotique (il n'est valable que lorsque n tend vers l'infini) et ne donne aucun moyen pratique pour comprimer un message. Mais SHANNON – et, indépendamment, Robert FANO (1917–2016) – ont l'idée de considérer un code à longueur variable où les symboles les plus probables sont codés par les codes les plus courts, de sorte que le débit moyen devient assez proche de H. Quatre ans plus tard David HUFFMAN (1925–1999) décrira l'algorithme optimal de compression dans ce contexte.

L'information mutuelle. SHANNON fonde également sa théorie sur la quantité

$$I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)\,p(y)}$$

que FANO nomme *information mutuelle* et qui mesure la quantité moyenne d'*information* entre deux variables aléatoires X et Y. C'est la première fois que le concept – jusque là flou – d'information transmise dans un système trouve une théorie rigoureuse.

Le deuxième théorème de Shannon. La notion d'information mutuelle permet à SHANNON de résoudre le problème théorique de la *transmission* dans un canal bruité d'entrée X et de sortie Y. Il s'agit cette fois de maximiser le débit d'information transmis tout en garantissant une communication arbitrairement fiable du message au destinataire. Pour cela SHANNON démontre que l'on peut choisir un code au hasard selon une distribution de probabilité qui rend $I(X;Y)$ maximal et nomme ce maximum

$$C = \max_{p(x)} I(X;Y),$$

la *capacité du canal*. C'est le *deuxième théorème de Shannon* qui affirme qu'on transmettre l'information de façon fiable tant que le débit ne dépasse pas la capacité C du canal. Ce théorème est une véritable révolution qui a changé le monde : pour la première fois, on comprend que le bruit présent dans le canal ne limite pas la qualité de la communication, il ne limite que le débit de transmission. À la condition de ne pas dépasser la capacité, la communication numérique peut être quasi-parfaite ! Ce théorème à lui seul justifie l'explosion du numérique aujourd'hui.

Lorsque le bruit présent dans un canal de transmission est modélisé par du bruit blanc gaussien qui s'ajoute au signal à la réception, SHANNON trouve l'expression exacte et étonnamment simple :

$$C = W \cdot \log_2\left(1 + \frac{P}{N}\right) \quad \text{bit par seconde}$$

où W est la largeur de bande et P/N le rapport signal à bruit présent dans la transmission. C'est certainement la formule la plus connue de SHANNON : elle fournit un aspect concret de la théorie de l'information qui a séduit de nombreux ingénieurs dès sa parution. Pas moins de 7 autres chercheurs publient une formule similaire la même année 1948 !

L'héritage de SHANNON en a dérouté plus d'un : ses théorèmes prévoient qu'il existe de bons systèmes de codage pratiques, mais ne disent pas comment les construire. Paradoxalement, ses démonstrations suggèrent que des codes choisis au hasard forment des solutions quasi-optimales (mais irréalisables en pratique). Il a fallu 50 ans pour que Claude BERROU (né en 1951) et Alain GLAVIEUX (1949–2004) proposent une solution pratique (les turbo-codes) qui « imite » le codage aléatoire et permet de s'approcher de la capacité.

Tout ceci n'est qu'un aperçu. La théorie de l'information n'a jamais été aussi vivante et trouve de nombreuses applications en réseaux sans fil, en sécurité de systèmes embarqués, en gestion de portefeuilles, en séquençage génomique, et même en interactions homme-machine.