

INSUFFISANCE DE LA DISTANCE DE HAUSDORFF

Pourtant la distance de Hausdorff ne reconnaît pas certaines proximités entre formes que l'esprit humain identifie instinctivement. Ainsi la distance de Hausdorff entre un dessin et sa version en négatif (ou sa version où seul le contour des formes est gardé) est grande, alors que bien sûr nous, nous voyons très rapidement que le même objet est représenté. Pire, ajouter un unique point situé à l'extérieur d'un dessin donne un dessin très éloigné pour la distance de Hausdorff, alors que les deux objets nous apparaissent presque identiques. De même encore, la distance entre un objet et le même objet rapetissé d'un facteur constant est grande alors que nous percevons presque instantanément leur structure commune.

De même encore, la distance de Hausdorff ne permet pas de reconnaître la similitude entre des images d'un objet tridimensionnel (un visage, par exemple) vu sous divers angles ou des éclairages différents, alors que notre cerveau perçoit très vite l'analogie des formes.

Un défaut mineur, mais à signaler aussi de la distance de Hausdorff est qu'elle ne donne rien de bon avec des objets infinis en taille. Deux droites passant par un même point, ou bien sont confondues, et alors distantes de 0 (ce qui est satisfaisant), ou bien sont distantes de $+\infty$ même si leur angle est très faible (ce qui n'est pas acceptable).

Les déformations «caoutchouteuses» ne sont pas bien mesurées non plus : les montres molles de Dali ne sont pas proches des vraies montres pour la distance de Hausdorff, alors que, pour nous, elles le sont.

Remarquons qu'une combinaison de la notion de distance de Hausdorff avec la notion de similitude géométrique arrange certains de ces défauts : pour mesurer la distance entre A et B avec cette distance (appelée Hausdorff-bis), on envisage toutes les figures obtenues par application d'une similitude géométrique à B et l'on regarde celle qui, au sens de la distance de Hausdorff, est la plus proche de A ; c'est elle alors qui détermine la distance entre A et B pour Hausdorff-bis. Cette variante corrige les défauts concernant les déformations d'échelle et les déplacements, mais ne change rien aux insuffisances de la distance de Hausdorff concernant le passage au négatif, l'ajout d'un point à l'extérieur, les photos d'un même visage ou les déformations caoutchouteuses.

La dernière notion de ressemblance correspond à un progrès important sur le plan théorique, et en même temps

ouvre la porte à certaines applications pratiques. Cette notion, que nous appellerons distance informationnelle, est le résultat du travail commun de cinq chercheurs provenant de disciplines différentes : les physiciens Wojciech Zurek, du Santa Fe Institute, et Charles Bennett, du Centre de recherches IBM à New York ; le mathématicien Peter Gacs, de l'Université de Boston, aux États-Unis, et les informaticiens Ming Li, de l'Université de Waterloo, au Canada, et Paul Vitányi, de l'Université d'Amsterdam.

LA DISTANCE INFORMATIONNELLE

La distance informationnelle présuppose que tous les objets que l'on considère sont des ensembles finis de points pris dans un ensemble discret. Il faut par exemple imaginer qu'il s'agit des pixels (noirs ou blancs) d'une image. Pour la distance de Hausdorff, une telle hypothèse de finitude n'était pas nécessaire : pour affiner un concept, on est parfois obligé de renoncer à l'infini, ce qui n'est peut-être pas étonnant : selon certains philosophes des mathématiques, l'infini n'est qu'une illusion. Des généralisations de la distance informationnelle à plus de deux dimensions et prenant en compte les couleurs sont possibles, mais nous ne les examinons pas ici.

L'idée fondamentale est que, si deux objets sont semblables, on passe facilement de la description de l'un à la description de l'autre, et réciproquement. En revanche, plus le passage de l'une à l'autre est long à détailler, plus les objets doivent être considérés comme différents, c'est-à-dire éloignés.

La difficulté du passage d'une description de l'objet A à la description de l'objet B est mesurée par la longueur du plus court programme qui transforme la donnée des points de A (supposés fournis pas énumération) à la donnée des points de B . Et donc la distance informationnelle est la somme des longueurs du plus court programme permettant de transformer A en B , et du plus court programme permettant de transformer B en A . On montre que cette distance varie peu quand on change le langage de programmation utilisé pour mesurer la taille des programmes.

L'idée d'utiliser une notion de plus court programme provient de la théorie algorithmique de l'information (appelée aussi théorie de la complexité de Kolmogorov), où l'on mesure la complexité d'un objet A par la taille du plus court programme qui produit A . La théorie de la distance informationnelle utilise d'ailleurs

de nombreux résultats de la théorie algorithmique de l'information.

Pour illustrer la notion de distance informationnelle, considérons le dessin de la tour Eiffel et le même dessin en négatif (nous avons dit que la distance entre ces deux objets pour la distance de Hausdorff est grande). Ces deux objets sont-ils proches pour la distance informationnelle ?

Oui, car on écrit facilement un programme de «passage au négatif» et que ce programme est court : le plus court programme transformant le premier dessin en le second (qui est le même ici que celui transformant le second en le premier) sera encore plus court (notons qu'il n'est pas indispensable de connaître précisément le plus court programme), et donc les deux dessins seront proches au sens de la distance informationnelle.

Ajouter un point à l'objet A donne un objet B que notre cerveau considère proche. En accord avec cette perception (et contrairement encore à la distance de Hausdorff, qui se trompait dans un tel cas), la distance informationnelle trouve que A et B sont très proches, car le programme le plus court ajoutant un point, et le programme le plus court enlevant un point (pour passer de B à A) sont tous les deux petits.

Plus remarquable et plus intéressant est l'exemple suivant : si l'on considère une image de la tour Eiffel prise selon un autre angle de vue, alors là encore on pourra par l'utilisation d'un programme assez court (qui contiendra des informations sur la position tridimensionnelle de la tour Eiffel) passer d'une image à l'autre. La parenté entre deux images provenant d'un même objet pris sous différents angles, ou même sous différents éclairages, est correctement détectée par la distance informationnelle.

Si l'on considère deux textes, même très longs, le second étant le résultat d'une remise en page du premier (changement des polices de caractères, des sauts des lignes, etc.), ils seront proches pour la distance informationnelle, car le passage de l'un à l'autre est décrit – et donc programmé –, de manière concise.

De nombreux autres exemples prouvent que cette mesure de distance fondée sur la taille des programmes satisfait l'essentiel de nos attentes, et que la grande majorité des défauts de la distance de Hausdorff disparaissent avec la distance informationnelle. L'utilisation pratique de cette distance semble cependant poser de graves problèmes.

Peut-on mesurer facilement la distance informationnelle entre objets ? Malheureusement, non. On majore sans mal

la distance informationnelle entre objets : à chaque fois qu'on trouve des programmes qui transforment A en B et B en A , on en déduit une telle majoration. En revanche, trouver des programmes dont on puisse prouver qu'ils sont les plus courts possibles est exceptionnel, et donc mesurer la distance informationnelle exactement est très rare.

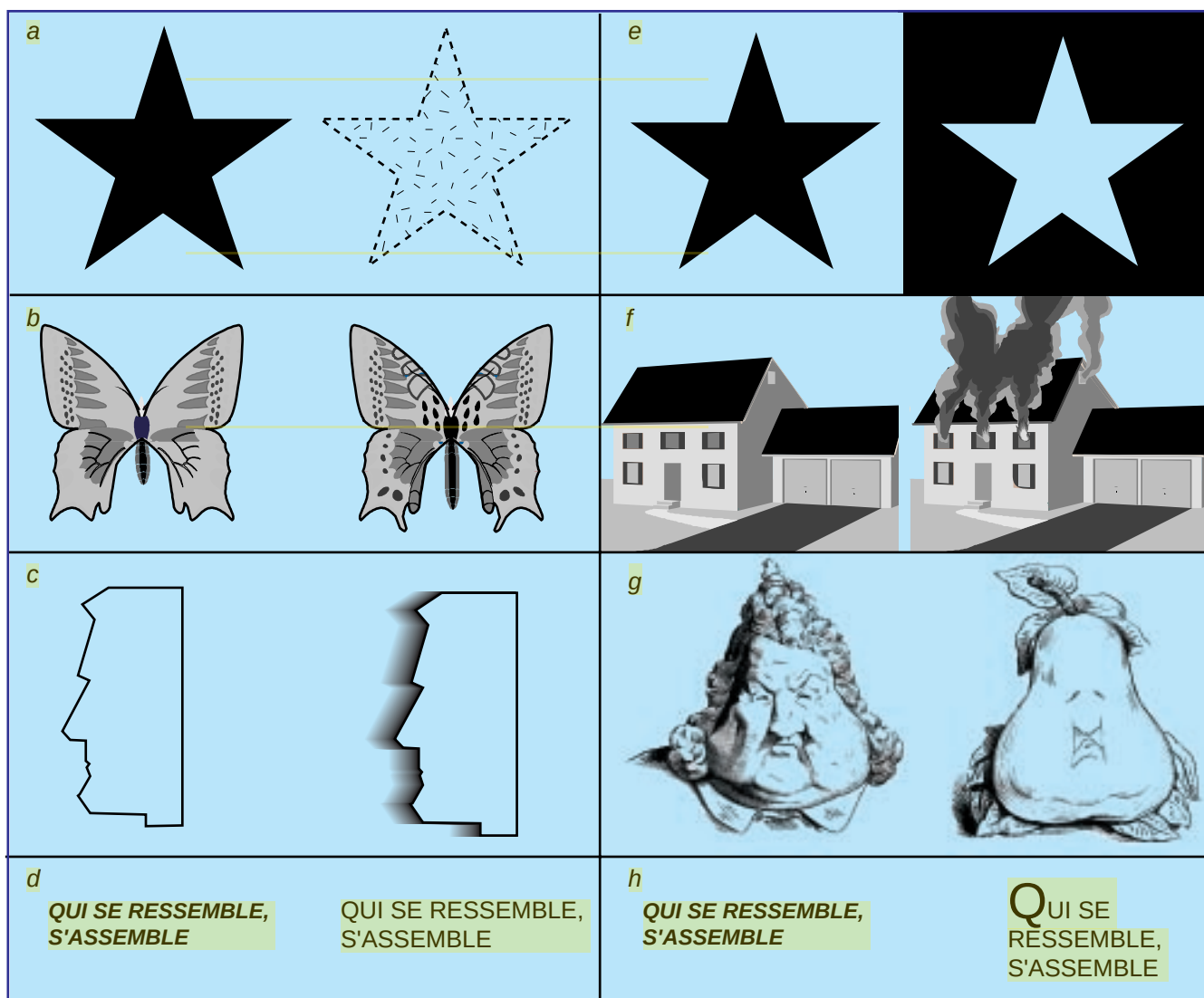
DISTANCE INFORMATIONNELLE ET INDÉCIDABILITÉ

Cette difficulté pratique est liée à l'indécidabilité logique dont le problème de l'arrêt des programmes proposé en 1936 par Alan Turing a été le premier

exemple. Le résultat de Turing nous dit que jamais on ne pourra faire un programme qui puisse, pour chaque programme qu'on lui soumettrait, calculer si oui ou non il s'arrête au bout d'un temps fini. Le résultat d'indécidabilité concernant la distance informationnelle nous dit que jamais nous ne pourrions concevoir un programme pour calculer, à chaque fois que nécessaire, la distance informationnelle de deux objets. Certains programmes pourront peut-être traiter quelques cas, mais aucun ne pourra les traiter tous, et donc aucun mécanisme général de raisonnement ou de calcul ne pourra servir d'instrument de mesure général pour la distance informationnelle.

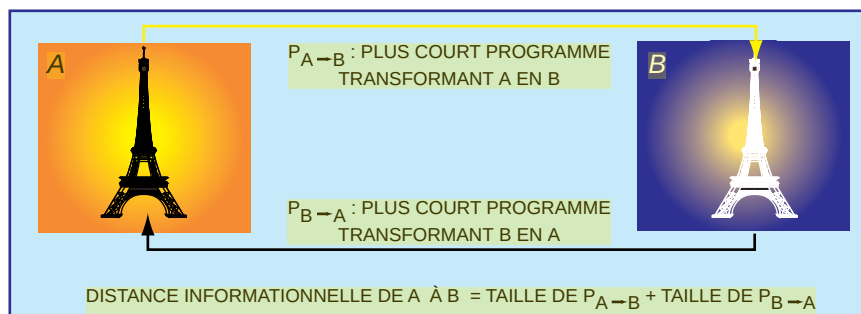
Cette impraticabilité reflète sans doute l'analyse intelligente nécessaire à la formulation des jugements de ressemblance : certains nous sont immédiats, d'autres ne sont perçus qu'après de longues analyses, d'autres encore ne le sont que par certains rares esprits. Est-il alors étonnant que la mesure du degré d'analogie entre des objets quelconques soit difficilement mécanisable et ne puisse pas, en définitive, l'être complètement.

L'intelligence est la cristallisation dans nos gènes, et ensuite dans notre cerveau, des très longs calculs réalisés par l'évolution et par le développement des cultures humaines. L'évolution biologique et le développement culturel sont



5. Pour nous, dans chaque paire les deux dessins se ressemblent clairement. La distance de Hausdorff saisit une partie de ces ressemblances (celles de la partie gauche (a), (b), (c), (d), mais passe à côté de celles de la partie droite (e), (f), (g), (h). La distance de Hausdorff est un moyen efficace mais pas infallible d'appréciation des ressemblances. En revanche, les cas où la distance de Hausdorff était impuissante à voir les ressemblances sont cette fois correctement traités par la distance informationnelle. Le programme pour transformer un dessin en son négatif (e) est très court, les

deux dessins sont donc proches pour la distance informationnelle. Le programme pour ajouter un peu de fumée et celui pour l'effacer (f) sont courts donc les deux dessins sont proches pour la distance informationnelle. Un programme permet de passer d'un dessin à l'autre (g), et ils sont relativement proches (et en tout cas bien plus proches l'un de l'autre que deux dessins quelconques). Quelques indications de remise en forme permettent de passer d'un texte imprimé à l'autre et réciproquement, qui sont donc proches l'un de l'autre pour la distance informationnelle (h).



6. La distance informationnelle entre deux objets A et B s'obtient en additionnant la taille des plus courts programmes permettant de passer de A à B et de B à A. Elle possède des propriétés de minimalité qui lui confère une importance particulière en théorie algorithmique de l'information. De plus les rapports qu'elle a avec la thermodynamique du calcul lui donnent un sens physique que les autres distances ne possèdent pas.

en effet de très longues séquences d'événements assimilables à des opérations élémentaires de calculs dont les produits sont le langage, les sciences et plus généralement nos facultés d'analyse. Ces très longs calculs, sans que nous sachions précisément comment, ont produit dans nos cerveaux une notion d'analogie très fine et très efficace, essentielle à notre survie. Cette cristallisation est incomplète (il n'y a pas de raisons sérieuses de croire que notre intelligence échappe aux limitations formulées par la logique), mais elle est bien meilleure que celle que nous réussissons à insérer dans nos programmes : aujourd'hui nous ne savons pas écrire des programmes qui égalent notre capacité à percevoir des analogies.

LA DISTANCE ENTRE SÉQUENCES GÉNÉTIQUES

L'indécidabilité rend-elle la distance informationnelle stérile ? Heureusement non, car les majorations que l'on peut faire la rendent utilisable dans des cas sans doute assez nombreux.

Un exemple d'application de la distance informationnelle est la distance définie dans l'équipe de bio-informatique de Lille, composée de Max Dauchet, Éric Rivals, Jean-Stéphane Varré et moi, pour évaluer la distance entre séquences génétiques. Depuis longtemps les généticiens, et tout particulièrement ceux qui souhaitent faire de la reconstitution d'arbres phylogénétiques (ce sont les arbres qui indiquent les parentés entre espèces animales et végétales), utilisent des mesures de ressemblance entre séquences génétiques. L'idée à la base des distances utilisées jusqu'à présent était le décompte des mutations, délétions ou insertions de nucléotides (les lettres de l'alphabet génétique) : plus le nombre des changements ponctuels nécessaires pour passer d'une séquence à l'autre

est grand, plus les séquences sont considérées comme éloignées. Récemment nous avons proposé une nouvelle distance qui mesure l'éloignement entre séquences en considérant d'autres événements possibles comme les déplacements de morceaux de séquences ou leur duplication. Cette distance, que nous avons conçue comme une approximation calculable de la distance informationnelle, est plus précise que les distances utilisées classiquement (qu'elle généralise). On peut donc espérer qu'elle fournira des arbres phylogénétiques plus exacts.

Notons encore que les algorithmes modernes de compression de données pour les images de films sont fondés sur l'idée que deux images successives A et B d'une séquence sont proches et que, pour gagner de l'espace, il suffit de ne coder que leur différence : ce qui change dans A pour obtenir B. Cette quantité d'informations permettant de passer de A à B est en fait la distance informationnelle entre A et B (ou presque, puisque l'on ne s'occupe pas du passage de B à A).

INTERPRÉTATION THERMODYNAMIQUE

L'interprétation thermodynamique de la distance informationnelle est une autre confirmation de son intérêt. Les cinq chercheurs physiciens, mathématiciens et informaticiens ont en effet démontré un résultat remarquable concernant la distance informationnelle : elle est liée aux coûts thermodynamiques minimaux de la transformation de A en B et de B en A.

On sait depuis quelques années, à la suite des travaux de Rolf Landauer et de Charles Bennett, qu'il est possible, en théorie, de réaliser des calculs sans dépense d'énergie et n'entraînant donc aucun accroissement de l'entropie physique. Lors d'un calcul, les seules dépenses d'énergie inévitables proviennent de l'utilisation d'opérations irréver-

sibles comme l'effacement d'une mémoire dont on peut se passer si l'on accepte de garder des informations inutiles en fin de calcul.

Il est alors intéressant de considérer le flux minimum d'informations (ajout d'informations à A au départ, effacement d'informations à la fin du calcul une fois B obtenu) pour transformer A en B de manière réversible. Cette distance déduite de considérations thermodynamiques est équivalente à la distance informationnelle : elle en diffère d'un facteur deux au plus. D'autres propriétés de minimalité confèrent, parmi toutes les distances envisageables, un statut privilégié à la distance informationnelle.

La distance informationnelle entre A et B est donc une quantité ayant un sens en thermodynamique, ce qu'aucune autre distance ou notion mathématique de similitude ne possédait jusqu'à présent. L'adéquation entre ce que donne la nouvelle distance et l'intuition, le lien avec le monde physique, ainsi que d'autres résultats sur la distance informationnelle trouvés par le groupe des cinq chercheurs autour de Charles Bennett montrent clairement qu'une notion profonde a été identifiée.

Jean-Paul DELAHAYE est directeur adjoint du Laboratoire d'informatique fondamentale de Lille du CNRS.

e-mail : delahaye@lil.fr

C. BENNETT, et al. *Thermodynamics of Computation and Information Distance*. Proc. 25th ACM Symp. Theory of Computation, pp. 21-30, 1993.

C. BERGE, *Espaces topologiques et fonctions multivoques*, Éditions Dunod, Paris, 1966 (voir le chapitre 6).

J.-P. DELAHAYE, *Information, complexité et hasard*, éditions Hermès, Paris, 1994.

M. LI et **P. M. B. VITANYI**, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer-Verlag, 1997 (voir le chapitre 8).

J. SERRA, *Image Analysis and Mathematical Morphology*, Academic Press, New York, 1982.

J.-S. VARRÉ, *Phylogénie et compression de données*, Publication du Laboratoire d'informatique fondamentale de Lille, URA CNRS 369, juillet 1996.

J.-S. VARRÉ, **E. RIVALS**, **M. DAUCHET** et **J.-P. DELAHAYE**, *Les distances transformationnelles et applications à la phylogénie*. Journées Analyse des séquences génomiques, École Polytechnique (Palaiseau), 20 et 21 juin 1996.

Articles de P. Vitanyi sur Internet :

<http://www.cwi.nl/~paulv/publications.html>