# Où est mon information ?

Jean-Louis DESSALLES (81)


***Le prix de certaines informations est devenu négatif : nous sommes prêts à payer pour ne pas recevoir la plupart des messages qui assaillent nos boîtes électroniques. À l'inverse, le coût de l'information pertinente, ou le temps nécessaire pour la trouver, risque d'augmenter indéfiniment. Pourrons-nous échapper à cette malédiction ?***

Quelle est la part qui nous intéresse dans le flot d'information qui nous atteint ? Prenons l'exemple de la messagerie. Pendant l'année académique 2007-2008, j'ai reçu 11 385 messages, soit quelque 31 messages par jour. Je dispose de plusieurs filtres qui n'ont laissé passer que 3201 pourriels[1], que j'ai dû classer comme tels à la main. Sur les 8184 messages restants, j'en ai jugé 6122 dignes d'être classés dans mes rubriques thématiques. On peut donc estimer que le taux de pertinence est de 54% pour cette année-là (*voir encadré 1*).

Lors de ma dernière recherche sur mon moteur préféré, sept des dix premiers résultats étaient sans aucun rapport avec ce que je recherchais. Les moteurs de recherche font ce qu'ils peuvent. Rien n'indique que leur taux de pertinence va s'améliorer avec le temps.

## L'opacité du monde de cristal

Nous payons le manque de pertinence des informations par le temps que nous perdons à les trier. Ne peut-on imaginer des systèmes qui le fassent pour nous ? Pas si simple. En matière d'information, nous sommes actuellement sur le flanc d'une vague grandissante et le rivage n'est pas en vue. L'émergence spontanée du Web au milieu des années 1990 a considérablement accru le réservoir des informations susceptibles de nous intéresser. Cela continue : nous sommes encore dans la phase de croissance exponentielle de services comme Wikipedia (*figure 1*). L'émergence, également exponentielle, de réseaux sociaux sur le Web (la croissance des dix réseaux les plus actifs est comprise entre 57 % et 343 % par an)[2] engendre une nouvelle classe d'informations : j'ai besoin de lire les nouvelles postées au sein de mon groupe, même les plus futiles, pour rester au courant. Il en va de mon existence sociale au sein de ce groupe. Et que nous réserve l'avenir ?

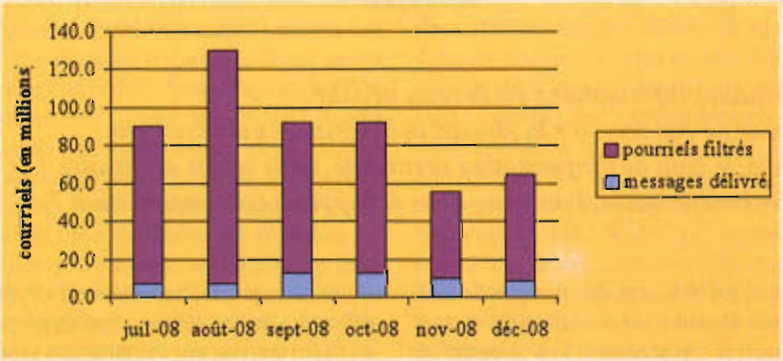
1 Les pourriels, ou spams, sont des courriels publicitaires envoyés en masse à des adresses collectées automatiquement sur le Web.
2 D'après blog.nielsen.com/nielsenwire/wp-content/uploads/2008/10/press_release23.pdf

**Encadré 1 :**

**En 2004, la CNIL (Commission Nationale de l'Informatique et des Libertés) dressait la typologie suivante pour les spams de langue française.**

| | |
|---|---|
| Messages à caractère pornographique / Rencontres | 55 % |
| Offres de biens et de services en ligne (achats biens de grande consommation et de services liés à internet) | 12.3 % |
| Tourisme | 6.5 % |
| Jeux/Casinos | 6.2 % |
| Crédits/Finances/Assurances | 5 % |
| "Escroquerie"/"Chaînes"(messages atypiques et/ou proposant des offres douteuses) | 3.3 % |
| Divertissement (messages incitant à consulter un site à caractère humoristique) | 2.3 % |
| Voyance | 2.3 % |
| Emploi | 1.1 % |
| Immobilier | 0.9 % |
| Santé | 0.5 % |
| Offres de biens et services autres | 4.6 % |

**La même année, le 24 janvier 2004, Bill Gates déclarait au forum économique mondial de Davos que les nuisances dues à ces pourriels seraient vaincues avant deux ans. Selon les mesures récentes effectuées en 2008 à l'Université de Yale (www.yale.edu/its/metrics/email), la proportion des pourriels reçus oscille autour de 90%, ce qui réfute la prédiction de Bill Gates.**



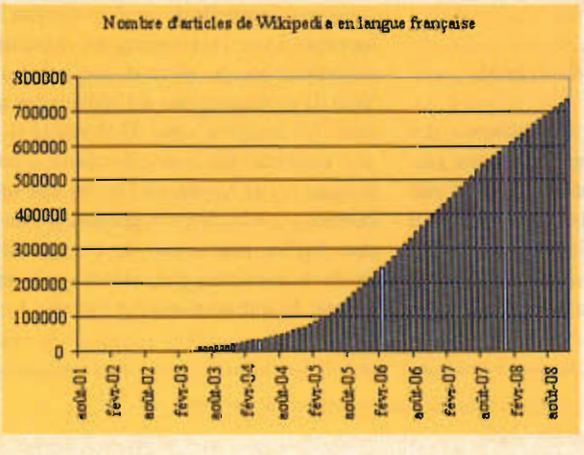




**Figure 1**

La sagesse, en matière de prévision, serait de ne pas en faire. Personne, pas même ses inventeurs, n'a imaginé l'émergence fulgurante du Web, et bien peu d'observateurs pariaient sur le succès de Wikipédia à ses débuts en 2001. Osons cependant une conjecture, celle du « monde de cristal », un monde où l'information ne peut être cachée. Chacun pourra, en principe, savoir tout sur tous les autres. Peut-être qu'en 2020, je saurai qu'une amie a acheté une bouteille de Sauternes dix minutes plus tôt dans une boutique de la rue du Bac. Son téléphone, sa carte de paiement, la bouteille et même ses habits (par les puces de radio-identification qu'ils contiendront) seront détectés en continu en milieu urbain. Il sera difficile d'empêcher les informations ainsi collectées d'atteindre Internet. Sommes-nous prêts à vivre dans un monde de cristal ? Notre seule protection sera, ironiquement, le fait que les informations nous concernant seront noyées dans un océan de non-pertinence. La transparence absolue pourrait ainsi rejoindre l'opacité.

Le constat est simple : la quantité d'information qu'une personne considère comme pertinente est bornée par le temps qu'elle souhaite passer à en prendre connaissance. Le réservoir des informations, lui, augmente de manière exponentielle. L'aiguille se fera de plus en plus petite dans une monstrueuse botte de foin. La croissance du Web et de ses rejetons serait-elle autodestructrice ?

## L'illusion du petit monde

La prédiction d'autodestruction semble démentie par les faits. Les moteurs de recherche parviennent à dénicher instantanément ma page personnelle parmi des dizaines de milliards de pages existantes. L'indexation peut faire des miracles que nous ne soupçonnions pas il y a quinze ans. Lorsque chaque brin de paille est répertorié, retrouver l'aiguille ne semble pas poser de problème.

Comment ne pas s'émerveiller ? Grâce aux moteurs de recherche, nous vivons réellement dans un village global. Chacun des internautes devient mon voisin, puisqu'il peut trouver ma page en quelques clics. En 1999, Albert-László Barabási et ses collègues ont mesuré pour la première fois le diamètre du Web, autrement dit le nombre maximal de clics nécessaires pour atteindre une page (Albert, Jeong & Ba-

rabási 1999). À leur grande surprise, ils ont constaté que le diamètre du Web n'excède pas vingt clics. Dans ce petit monde où tout est près de tout, ce qui m'intéresse n'est jamais loin. Si l'information est bien indexée, où est le problème ?

Le problème se situe précisément dans la structure du petit monde. L'incroyable réduction des distances est due à l'existence de hubs : des nœuds connectés à de très nombreux autres nœuds dans le graphe qui nous intéresse. Dans le Web, il existe des pages qui pointent vers des centaines ou des milliers d'autres pages. Inversement, certaines pages, particulièrement populaires, sont pointées par des centaines ou des milliers d'autres pages. L'émergence de hubs est due à une rétroaction positive, selon le principe rich get richer : plus une page est pointée, plus elle est connue, et plus elle a de chances d'être à nouveau pointée par de nouvelles pages (Barabási 2002).

Le moteur de recherche actuellement dominant, Google, a été créé en 1998 sur cette idée. La pertinence à la Google, où G-pertinence, d'une page est fonction du nombre de liens qui existent vers cette page. La G-pertinence semble fonctionner à merveille, et le succès initial de ce moteur de recherche lui est en bonne partie dû. Elle fonctionne selon un principe récursif : une information sera d'autant plus pertinente pour moi qu'elle est considérée par autrui comme pertinente. Est-ce vraiment là la pertinence dont on rêve ? Ou n'est-ce qu'une illusion ? Les moteurs de recherche prétendent fouiller pour moi l'océan des pages du Web ; en réalité, ils ne me permettent de surfer que sur la cime des vagues (*voir encadré 2*).

## Ma pertinence n'est pas la tienne

Il existe, perdues parmi les centaines de milliards de messages qui ne m'atteignent pas chaque année, parmi les dizaines de milliards de pages du Web, parmi les dizaines de millions de livres et les milliards d'articles scientifiques, quelques informations qui m'intéresseraient au plus haut point, qui pourraient peut-être modifier le cours de ma vie, et que pourtant je n'ai aucun moyen d'atteindre. Ces informations sont peut-être cruciales seulement pour moi et pour personne d'autre. Sur un plan plus quotidien, les informations les plus pertinentes qui pourraient s'afficher quand j'allume mon ordinateur le matin n'ont, actuellement, aucune chance de me parvenir.

J'ai visité récemment un village en Ouganda, où les habitants développaient un projet de vannerie artisanale. Tout événement inattendu ou émotionnel concernant ce projet, ce village, ses alentours, est susceptible de m'intéresser. Pour autant, je ne suis pas prêt à passer trente minutes sur un moteur de recherche pour retrouver la trace dans le Web de ce village et de son projet de vannerie, pour m'apercevoir ensuite qu'aucune nouvelle information digne d'intérêt n'est mentionnée à son propos.

À côté de la G-pertinence, qui est par nature imitative et indirecte, nous avons besoin de définir et de calculer une notion de pertinence personnalisée, ou P-pertinence. Les recherches que nous menons dans ce domaine à TELECOM ParisTech s'inspirent de l'observation des conversations quotidiennes. Lorsqu'ils conversent spontanément, les êtres humains exigent les uns des autres d'être pertinents et se détournent des individus qui ne le sont pas. La pertinence conversationnelle, dans la mesure où elle révèle ce qui intéresse les individus, nous a semblé offrir la meilleure définition de ce que pouvait être la P-pertinence.

## Rechercher l'inattendu

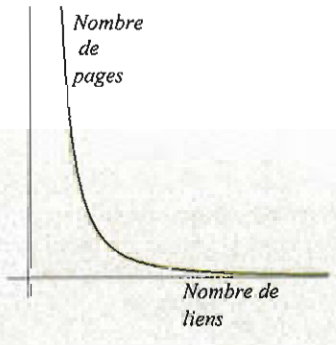
Il nous a été possible de modéliser la pertinence dans les conversations et d'en donner une définition mathématique (Dessalles 2008). La notion de P-pertinence que nous cherchons dépend de deux termes :

*P-pertinence = émotion typique + inattendu*

L'émotion typique permet de hiérarchiser les événements de manière relativement standard selon l'émotion qu'ils déclenchent en général. Par exemple, la mort d'un individu nous affecte généralement plus que le fait qu'il perde son emploi ou qu'il perde cent Euros. L'autre terme de la P-pertinence, l'inattendu, est plus intéressant car il relève d'un calcul égocentré.

La perception de l'inattendu semble systématiquement associée à un décalage de complexité. Une situation est inattendue lorsqu'elle est plus facile à décrire qu'à produire. Supposons qu'un monsieur Durand, que je ne connais pas, vienne à mourir lors d'un saut en parachute. L'événement ne m'intéresse pas vraiment. En revanche, si Durand se trouve habiter dans le même immeuble que moi, l'information me semble nettement plus intéressante. La proximité de Durand ne modifie en rien les conditions de production de l'événement, mais elle diminue la complexité de sa description. Ainsi, certains éléments, pourtant sans rapport avec l'événement, peuvent le rendre plus simple et donc plus intéressant à mes yeux. Par exemple, si l'on me dit que la victime du saut en parachute a occupé la même chambre d'étudiant que moi, son sort tragique m'intéressera un peu plus. Même chose si j'apprends qu'il s'agit du tout premier humain à avoir été conçu par fécondation in vitro. Ces propriétés rendent la désignation minimale de Durand plus concise que s'il s'agissait d'un quidam quelconque. L'inattendu résulte de cette simplicité.

**Encadré 2**




Les pages du Web qui reçoivent peu de liens sont très nombreuses (partie gauche du graphe). En 1999, les chercheurs ont eu la surprise de constater que la décroissance du nombre de pages en fonction du nombre de liens était beaucoup plus lente qu'attendu (décroissance en loi puissance). Ceci signifie qu'un nombre non négligeable de pages possèdent de très nombreux liens.

De par leur mode de fonctionnement, les moteurs de recherche atteignent préférentiellement les pages fortement connectées. En conséquence, l'immense majorité des pages ont une probabilité négligeable d'être atteintes. Certaines pages très pertinentes pour un utilisateur donné seront ainsi tout simplement ignorées.

**Encadré 3 :**

La P-pertinence, ou *pertinence personnalisée*, est la somme de l'émotion typique $E$ et de l'inattendu $U$ : $P = E + U$

L'émotion typique $E$ est relativement facile à mesurer. Elle correspond, pour les émotions négatives les plus fortes, à la prime d'assurance que l'on est prêt à payer pour se prémunir d'un risque. La mesure de $E$ fournit une hiérarchie *a priori* de la pertinence des événements, que l'inattendu peut bien entendu bouleverser. L'inattendu $U$ est défini comme un décalage de complexité : $U = C_w - C$

La complexité de production d'un événement, $C_w$, se mesure à l'ensemble des paramètres qu'il faut fixer pour que le « monde » produise cet événement. La complexité de production d'un crime donné dépend essentiellement de sa causalité (présence du criminel, choix de la cible, etc.) Un moyen commode pour approximer $C_w$ consiste parfois à voir le monde comme une loterie. Ainsi, s'il y a $N$ personnes dans une ville, la complexité du choix de la victime d'un crime donné peut être approchée par $C_w = \log_2 N$, car c'est le nombre de décisions binaires (penser à des lancers successifs d'une pièce de monnaie) dont la loterie a besoin pour désigner le malheureux élu.

La complexité de description $C$ correspond à la notion introduite par Gregory Chaitin et Andrei Kolmogorov dans les années 1960. Elle se mesure à la taille de la plus petite description qui permet de décrire la situation sans ambiguïté. Il s'agit d'une mesure égocentrée. Ainsi, la complexité d'une localisation augmente comme le logarithme de la distance à l'observateur, puisque celui-ci peut numéroter les positions selon leur distance. Cette loi vaut également pour les distances temporelles. Les personnes célèbres ou les endroits remarquables ont une complexité faible, que l'on peut mesurer par le logarithme de leur rang dans une liste (augmenté de la complexité de la liste elle-même). Ainsi, les monuments parisiens peuvent être rangés par complexité croissante en fonction du nombre de leurs visiteurs (1-Notre Dame ; 2-Sacré Cœur ; 3-Tour Eiffel ; etc.) Une manière prometteuse pour estimer la complexité par des moyens automatiques consiste à mesurer son impact sur le Web (Cilibrasi & Vitányi 2007). Un événement sera considéré comme d'autant plus complexe que les éléments qui le composent se trouvent rarement associés dans les textes du Web.

L'inattendu se convertit en probabilité subjective $p$, selon la formule : $p = 2^{-U}$

Cette formule inverse la relation classique de Claude Shannon, à condition de considérer que l'inattendu $U$ remplace l'information. La pertinence personnalisée $P$ combine l'émotion et l'information liée à la surprise pour fournir une nouvelle notion d'*information subjective*. C'est cette information subjective $P$ que les systèmes d'information personnalisée doivent maximiser.

Pour en savoir plus : www.unexpectedness.eu

Inversement, puisque l'inattendu repose sur un décalage de complexité, une situation anormalement complexe à produire m'apparaîtra également comme inattendue. Ainsi, si l'on me dit que Durand était âgé de 90 ans, je suis obligé d'imaginer des conditions complexes qui puissent amener un vieillard à sauter en parachute. La nouvelle revêt de ce fait un intérêt.

Cette définition de la P-pertinence comme décalage de complexité a une portée universelle. Nos recherches visent à permettre son calcul dans la plupart des cas d'intérêt pratique (*encadré 3*). Si nous y parvenons, nous disposerons d'une méthode radicalement nouvelle pour rechercher et pour trier des informations. Elle nous évitera de tomber dans la malédiction des petits mondes, puisque contrairement à la G-pertinence, son calcul est direct et ne repose pas sur l'imitation des préférences d'autrui. L'enjeu est le développement de nouveaux moteurs de recherche sur le Web et d'outils de veille informationnelle personnalisée. En comparant automatiquement la complexité de production et la complexité de description des situations, ces systèmes seront capables de découvrir, mêmes si elles se trouvent dans des endroits reculés du Web, les informations susceptibles de nous émouvoir et de nous surprendre.

La P-pertinence ne résoudra pas tous les problèmes. Celui qui connaît son mode de calcul peut en principe aisément la détourner à son profit. Au lieu de subir les assauts de pourriels génériques destinés au plus grand nombre, nous risquons ainsi d'être bombardés de fausses nouvelles spécialement conçues pour capter notre attention. Je risque bientôt de recevoir de faux messages dont l'entête me parlera de mon village ougandais préféré et dont le contenu sera une simple publicité. Que faire ? Il faudra sans doute concevoir des systèmes automatiques qui, à l'image de nos capacités de raisonnement, pourront tester par recoupements la cohérence des informations. Le progrès des techniques de traitement, de stockage et d'indexation nous réserve un avenir informationnel qui ne sera vraisemblablement pas un paradis, probablement pas un enfer, mais qui présente déjà des défis technologiques fascinants. ■

*Jean-Louis DESSALLES (81) est enseignant-chercheur à TELECOM ParisTech, où il s'attache à modéliser la pertinence dans la communication humaine spontanée. Il travaille également sur la simulation des processus évolutifs et sur la question fondamentale de l'origine du langage humain, considéré comme un jeu de communication entre agents égoïstes. Il est l'auteur de plusieurs livres, notamment « Aux origines du langage » et, tout récemment, « La pertinence et ses origines cognitives ».*

**Bibliographie**


- *Albert, R., Jeong, H. & Barabási, A.-L. (1999). Diameter of the World-Wide Web. Nature, 401, 130-131. http://www.es.emu.edu/~eairoldi/nets/public/albe_jeon_bara.1999.pdf*
- *Barabási, A.-L. (2002). Linked: The new science of networks. Cambridge, MA: Perseus.*
- *Cilibrasi, R. & Vitányi, P. (2007). The Google similarity distance. ACM Transactions on Knowledge and Data Engineering, . http://www.cwi.nl/~paulv/papers/amdug.pdf*
- *Dessalles, J-L. (2008). La pertinence et ses origines cognitives - Nouvelles théories. Paris: Hermes-Science Publications. http://pertinence.dessalles.fr*