Qu'est-ce qu'un moteur de recherche ? : sous le ronron, les machinistes et leur machination

Guillaume Sire

CHAPITRE 7. QU'EST-CE QU'UN MOTEUR DE RECHERCHE ?

Sous le ronron, les machinistes et leur machination

par Guillaume Sire

Plus de 90 % des requêtes en Europe, et plus de 80 % dans le monde, sont adressées au moteur de recherche Google. Dès qu'on cherche une réponse, dans le cadre de sa vie professionnelle ou pour des raisons personnelles, voire triviales – en fait pour toutes les questions, sans exception… –, on s'adresse à Google. Google, c'est l'oracle, le pont, le guide, le fil, la lumière dans le chaos informationnel, nettoyeur des écuries d'Augias. Il compare les prix, trace des itinéraires, agrège les informations. Bon nombre de nos actions et de nos idées dépendent, en partie, de ce qu'il a répondu, de ce qu'il nous répond. Comment ne pas l'aborder de façon critique ? Comment imaginer un monde où l'on ne connaîtrait rien de ce qui se trouve derrière pareil outil, les partis pris éditoriaux, les intérêts économiques et les valeurs défendues par ses concepteurs ? Qui sont-ils ? Que font-ils, pourquoi ?

Google assortit volontiers les résultats de son moteur de l'épithète « naturels » (*organic* en anglais) dès lors que ceux-ci ne dépendent pas, contrairement aux liens « sponsorisés » situés à droite de l'écran et en haut, dans un cadre légèrement rosé, d'une enchère, et ne génèrent aucun profit direct pour le moteur de recherche. Sous prétexte qu'ils ont été produits automatiquement, l'entreprise nous présente ces résultats comme s'ils avaient été produits par la nature elle-même, tombés d'un arbre et cueillis sous la canopée. C'est là une rouerie, évidemment, visant à nous faire oublier qu'il y a des ingénieurs dans l'arrière-boutique et que ceux-là ont fait des choix dont dépendent, en partie, certes, mais ils en dépendent, les résultats prétendument pertinents affichés à l'écran, leur hiérarchie, leur apparence. Ces ingénieurs ont des histoires, des valeurs, des projets. Ils ont des besoins et des envies. Autour d'eux il y a des chefs d'équipe, des grands patrons, des actionnaires, des objectifs, des partenaires, des concurrents. Il y a également des juges, des avocats, et des États, des députés et des sénateurs, et des commissions, des autorités comme la CNIL, l'Arcep et l'Autorité de la concurrence.

Il y a quelqu'un derrière le moteur : des machinistes, et des machinations. Les ingénieurs font des choix selon ce qui leur semble pertinent, et selon leurs intérêts. En *concevant* la machine, ils expriment leur *conception* de la société à laquelle ils la destinent, de sorte que le moteur est le reflet d'une sociologie implicite.

Ils conçoivent de petits logiciels, appelés *crawlers* ou *spiders* ou *bots*, et chargés de visiter le web en naviguant de lien en lien, et recueillant, quand ils arrivent sur une nouvelle page, un certain nombre d'informations. Ils peuvent mettre tout ce qu'ils trouvent en mémoire ou seulement une partie : les adresses URL, les titres, les sous-titres, les légendes des images. Le concepteur doit choisir. S'il veut tout mémoriser, il lui faudra des infrastructures conséquentes, et, donc, plus d'argent, que s'il ne mémorise qu'une partie. Il ne peut pas forcément faire tout ce qu'il veut. Il doit arbitrer. Google enregistre tout. Ses *crawlers*, les *Googlebots*, copient le web en permanence, toute l'information textuelle à leur disposition, sur leurs propres serveurs, où les contenus peuvent ensuite être analysés par des logiciels lexicométriques. Un index est généré, qui, pour chaque mot, dresse une liste des pages où il figure.

Une fois les pages connues et répertoriées, il s'agit de les classer selon un ordre de pertinence supposée, et, pour cela, de paramétrer un algorithme, c'est-à-dire « une marche à suivre ». D'abord, il faut choisir des critères pour mesurer la pertinence des documents les uns par rapport aux autres, pour une requête donnée. Ces critères, dans le cas de Google, sont principalement liés à la position du mot au sein d'un document, ainsi qu'à la position de ce document par rapport aux documents répertoriés, pour ce mot, dans l'index. Ce deuxième critère est appelé « centralité », il est mesuré grâce au célèbre « PageRank » [Brin et Page, 1998], qui consiste, pour dire vite, à considérer 1) que chaque lien hypertexte est un vote pour le document qu'il pointe ; 2) que plus un document est pointé par des liens hypertextes, plus il a lui-même de l'influence quand il pointe vers d'autres documents ; 3) que plus un document comporte de liens vers d'autres documents et moins chacun de ces liens est un facteur de pertinence aux yeux du système. D'autres critères existent : la vitesse de chargement, la fréquence de publication, les signaux sociaux (*like*, tweets, etc.), l'historique de navigation de l'internaute… Il y aurait pour Google entre 200 et 300 critères. Chacun d'eux procède d'un choix discutable, qui dépend des ingénieurs, de leur vision du monde et de leurs intérêts, et avec lequel l'utilisateur n'est peut-être, après tout, pas d'accord. La vitesse de chargement de pages, par exemple, fait de la performance du contenant un critère de pertinence du contenu, ce qui revient à faire courir un cent mètres à deux philosophes en décrétant que toutes choses étant égales par ailleurs, le gagnant sera le plus pertinent pour parler de Spinoza. On trouve ici une vision assez caractéristique des ingénieurs de Google, intéressés par l'efficacité du système et le confort des utilisateurs autant, voire davantage, que par la nature de la source et la fiabilité de l'information.

Une fois les critères déterminés, il faut les pondérer les uns par rapport aux autres. Là encore, les choix des ingénieurs dépendent de leur vision du monde et de leurs intérêts, et mériteraient d'être discutés. Quels sont les critères les plus importants ? Pourquoi ? Enfin, dans l'algorithme, certaines constantes peuvent intervenir. Cela veut dire que toutes les requêtes, quels que soient leurs motifs – rentabilité, divertissement, curiosité… –, sont traitées *de la même manière*. Dans le cas de Google, il est extrêmement intéressant de lire ce que Bernhard Rieder a écrit sur la constante *alpha* [Rieder, 2012].

ÊTRE OU NE PAS ÊTRE TRANSPARENT

Les algorithmes des moteurs de recherche doivent-ils, ou non, être transparents ? La question est épineuse. Comme on le sait, l'algorithme de Google n'est pas transparent. Autrement dit, la machine vous propose de répondre à toutes vos questions, sauf à celle-là : « comment fais-tu exactement, machine, pour répondre à toutes les questions ? » Le philosophe Paul Mathias parle d'inversement de l'adage platonicien : « C'est un peu comme si savoir exigeait qu'on ne sût pas pourquoi ni comment l'on sait ! » [Mathias, 2009, 41].

Le problème avec la transparence, ce sont les pratiques malveillantes. Si l'algorithme de Google était tout à fait transparent, les producteurs de contenus pourraient être tentés de modifier leurs pratiques de façon à duper le système, et apparaître dans des listes de résultats à l'intérieur desquelles leurs contenus n'ont normalement rien à faire. Cela ne se produirait pas s'il y avait atomicité de l'offre, c'est-à-dire un très grand nombre de moteurs de recherche différents ayant chacun une part de marché égale. Mais Google est en position de quasi-monopole et ne peut donc pas se payer le luxe de la transparence, car s'il le faisait, les comportements malveillants empêcheraient son moteur de fonctionner correctement et alors, sans doute, perdrait-il son monopole !

Les moteurs de recherche peuvent avoir un effet normatif, surtout quand le marché est concentré : les éditeurs ajustent leurs comportements dans le but de figurer en tête des classements. Les techniques qu'ils mettent en œuvre sont rassemblées sous le vocable *search engine optimization* (SEO). Des entreprises de conseil en ont fait leur cœur d'activité. Certaines entreprises comme les e-commerçants ou les éditeurs de presse internalisent cette fonction en embauchant un expert. Du point de vue de l'éducation aux médias, il est intéressant d'interroger la portée de ces effets normatifs sur l'organisation du travail de production médiatique et sur le contenu lui-même qui ne s'adresse pas seulement aux lecteurs, mais aussi, pour les atteindre, aux algorithmes des moteurs de recherche [Sire, 2016].

ÊTRE OU NE PAS ÊTRE RESPONSABLE

Quand un moteur de recherche génère un lien vers un contenu illégal, soit parce que le contenu a été publié à cet endroit sans l'accord de son ayant droit, soit parce que c'est un contenu qui contrevient à la loi, pédopornographique, néonazi, que faire ? Est-ce que le concepteur du moteur de recherche est responsable ? Dire qu'il est responsable au même titre que celui qui a piraté ce contenu dans le premier cas, ou qui l'a produit dans le second, est évidemment exagéré, cependant dire qu'il n'est absolument pas responsable en supposant que c'est un outil neutre ne peut pas convenir non plus.

Le juriste James Grimmelmann [2014] a écrit sur cette question un article très complet, dans lequel il nomme la première approche « théorie de l'éditeur », et la deuxième « théorie du tuyau ». Ces deux théories ont en commun l'absence de considération pour l'utilisateur. Dans l'une comme dans l'autre, on ne responsabilise pas en effet celui qui a formulé la requête ayant conduit vers un contenu violant le droit de la propriété, ou bien vers un contenu pédopornographique ou néonazi. Après avoir présenté comment et expliqué pourquoi les juges confrontés à cette question ont opté pour l'une ou l'autre de ces théories, James Grimmelmann plaide en faveur d'une troisième approche, « la théorie du conseiller », qui donnerait à l'utilisateur une forme de responsabilité, tout en responsabilisant les concepteurs des moteurs sans pour autant les assimiler aux auteurs du contenu illégal ou de la copie interdite.

Dans l'optique d'une éducation aux médias, il semble nécessaire de sensibiliser à la fois à la responsabilité du moteur – et de déconstruire, ce faisant, l'idée qu'un logiciel de traitement des contenus puisse être neutre sous prétexte qu'il fonctionne automatiquement – et à la responsabilité de l'utilisateur.

LES INCITATIONS ÉCONOMIQUES

Les moteurs de recherche appartiennent à des entreprises qui ont des modèles d'affaires, des actionnaires, et, donc, des intérêts économiques susceptibles de les conduire à préférer un contenu plutôt qu'un autre. Avec Bernhard Rieder, nous avons expliqué en détail comment, dans le cas de Google, il pouvait exister des incitations au biais dans le traitement informationnel. L'entreprise en effet est une régie publicitaire qui propose aux éditeurs, quels qu'ils soient, de devenir partenaires de son réseau d'annonceurs. Dans ce cas, ils n'auront pas besoin de chercher eux-mêmes des annonceurs pour monétiser leurs encarts ; il leur suffira d'indiquer à Google où sont ces encarts et de laisser l'entreprise y faire apparaître les publicités de ses clients annonceurs. Google a donc

objectivement intérêt à ce que les contenus des partenaires de sa régie publicitaire soient favorisés dans les résultats de son moteur de recherche [Rieder and Sire, 2014].

Étant donné l'opacité de l'algorithme, irrémédiable pour les raisons précédemment expliquées, le simple fait que Google ait objectivement intérêt à favoriser les éditeurs partenaires de sa régie dans ses résultats est problématique. D'autant que Larry Page et Sergey Brin ont eux-mêmes écrit en 1998, à l'époque où ils étaient encore de jeunes doctorants idéalistes, que le financement d'un moteur de recherche grâce à une activité de régie publicitaire n'était pas souhaitable dans la mesure où « un moteur de recherche pourrait ajouter un petit facteur à son algorithme pour avantager systématiquement les compagnies "amies" et soustraire le même facteur aux résultats de leurs concurrents. Ce type de biais serait très difficile à détecter mais pourrait avoir des effets significatifs sur le marché »[1] [Brin and Page, 1998, annexe A].

Là encore, il convient de situer les intérêts des concepteurs du moteur de recherche par rapport aux intérêts des éditeurs, non pas forcément pour identifier un biais, mais au moins pour avoir en tête qu'il existe des incitations à biaiser les résultats. C'est un cas finalement similaire aux entreprises de presse soupçonnées de louer les mérites des entreprises reliées à leurs actionnaires ou bien des entreprises auxquelles elles vendent leurs encarts publicitaires. Les arguments qui ont justifié que fût mise en place une éducation critique à la lecture de la presse justifient donc pareillement qu'une éducation critique à l'utilisation des moteurs de recherche y soit ajoutée.

LA PERSONNALISATION

Le dernier point essentiel à soulever concernant les moteurs de recherche a trait à la personnalisation. L'utilisateur doit se demander : à quel point les résultats qui apparaissent prennent en compte des données qui me sont propres, et sont, donc, taillés pour moi ? Le moteur est-il un outil d'information me permettant de m'informer à propos de ce que je ne connais pas, ou un moyen de confirmer ce que je pense déjà ? Comment puis-je le savoir ? Y a-t-il un moyen de contrôler le degré de personnalisation ?

Les moteurs peuvent avoir un intérêt à personnaliser les résultats qui n'est pas seulement lié à leur activité publicitaire. Notamment, cela rend impossible toute étude exhaustive du comportement d'un moteur. Lorsque les résultats sont personnalisés, on ne peut pas savoir, dans l'absolu, si le moteur préfère tel ou tel contenu, tel ou tel camp politique, ou tel ou tel produit, ce qui

1. Nous traduisons toutes les citations de ce chapitre.

ne veut pas dire qu'il ne préfère pas systématiquement un contenu, un camp ou un produit.

Pour autant, être opposé à toute forme de personnalisation serait absurde. Nous préférons avoir des résultats dans une langue que nous comprenons. Nous sommes heureux de pouvoir retrouver une page qu'on a déjà plusieurs fois visitée et que, pour cette raison, le moteur fait remonter. D'où l'importance d'une éducation critique aux médias, et d'une compréhension fine du débat concernant la personnalisation. Comme le dit Engin Bozdag,

> *nous devrions réfléchir à des solutions qui permettraient de minimiser les effets négatifs et de maximiser les effets positifs* [*des technologies de personnalisation*] *plutôt que d'essayer de nous en débarrasser totalement. La question n'est pas de savoir s'il doit y avoir personnalisation ou non, mais de comprendre comment mettre au point une technologie de personnalisation moralement acceptable.* [Bozdag, 2013, 221]

CONCLUSION

Une technique est toujours l'objet d'« une mise en tension entre la volonté d'élargir le champ des possibles et la tentation de réduire l'environnement à n'être qu'un milieu conditionné » [Bachimont, 2010, 175]. Les moteurs de recherche ne font pas exception. Ils sont nécessaires pour qui veut se déplacer dans le déluge des informations, et pourtant ils conditionnent le milieu dans lequel celui-là évolue. C'est pourquoi il est indispensable d'avoir une approche critique de l'outil, et d'apprendre aux élèves, aux étudiants et plus largement à tous les utilisateurs des moteurs à comprendre ce qu'il y a derrière, les questions juridiques, les intérêts économiques, et les prises de position techniques ou politiques dont le moteur est le résultat concret, et, ce, jusqu'à nouvel ordre.

Un moteur de recherche *fait faire* des choses aux internautes et aux éditeurs, et en *fait dire* aux contenus et à l'infrastructure. C'est pour ne pas être dupe de ce faire-faire et de ce faire-dire que les questions présentées ici méritent à mon avis d'être constamment posées, cela car aucun moteur, jamais, ne les aura résolues *une fois pour toutes*. Elles existent, quoi qu'il arrive, et elles existeront. Les concepteurs des moteurs de recherche et les législateurs, les juges, les autorités, quant à eux, nous proposent des réponses ; à nous de savoir si ces réponses nous satisfont, et à quel point. À nous de les critiquer.