L'ESPRIT DU TEMPS ET LA STRATÉGIE DE L'ANÉMONE COMMENT ET POURQUOI L'ENTREPRISE GOOGLE COMMUNIQUE-T-ELLE LES DONNÉES RELATIVES AUX REQUÊTES EFFECTUÉES SUR SON MOTEUR ?

Guillaume SIRE
Maître de conférences en Sciences de l'Information et de la Communication

■ Les données relatives aux requêtes des internautes sur le moteur de recherche Google constituent un trésor que le propriétaire du moteur est, de fait, seul à posséder. Or il se trouve que Google publie partiellement ces données, mais en protège les détails les plus valorisables sur le plan économique et informationnel dans un mouvement comparable à celui de l'anémone qui se déploie et se replie sur elle-même, de façon à ne montrer à son entourage immédiat, fait de proies et de prédateurs, que ce qu'elle veut bien leur montrer, pour mieux attirer les proies et écarter les prédateurs. Nous décrivons et analysons dans cet article le comment et le pourquoi d'une telle stratégie.

■ De gegevens over de zoekopdrachten van internetgebruikers via de zoekmachine Google zijn een ware schat, waarvan de eigenaar van de zoekmachine in feite de enige eigenaar is. Google publiceert een deel van deze gegevens, maar beschermt de meest waardevolle details op economisch en informatievlak als een anemoon die zichzelf openstelt en dan weer in zichzelf plooit om zijn onmiddellijke omgeving, die bestaat uit prooien en roofdieren, niet te laten zien wat ze eigenlijk wil laten zien om prooien aan te trekken en roofdieren weg te jagen. In dit artikel beschrijven en analyseren we het hoe en waarom van zo'n strategie.

Google propose aux internautes une multitude de services dont la plupart peuvent être utilisés gratuitement et dont l'efficacité est légendaire : moteur de recherche, messagerie électronique, cartographie, stockage, vidéos, réseau social. Bien entendu, la gratuité de ces services ne signifie pas que l'entreprise ne réalise aucun profit, mais qu'elle réussit à monnayer leur déploiement et leur fonctionnement grâce à d'autres activités qui, elles, sont lucratives, et même extrêmement lucratives. Les chiffres se passent de commentaire. La *holding* de Mountain View, qu'il convient d'appeler Alphabet depuis août 2015 (le nom Google renvoie désormais seulement aux activités liées au web au sein de cette *holding*) était valorisée à plus de 650 milliards de dollars au début de l'été 2017, ce qui en faisait la première capitalisation boursière du monde. Elle avait enregistré un chiffre d'affaires de 90 milliards de dollars en 2016, soit une croissance de 22 % par rapport à 2015. Et son résultat net avait atteint 19,5 milliards de dollars, en hausse de 23%.

Pour l'essentiel, ces revenus proviennent du marché de la publicité où Google valorise auprès des annonceurs le nombre d'utilisateurs sur ses services ainsi que sa capacité à afficher des publicités susceptibles d'intéresser chacun d'entre eux. Les revenus issus de ces activités représentent à peu près 90% de la totalité du chiffre d'affaires réalisé (exactement 87% sur le dernier trimestre de 2016).

Les parties en présence

Google a des relations avec plusieurs types d'acteurs, qu'il convient de lister, afin de bien comprendre sur qui, et comment, la firme de Mountain View est susceptible d'exercer son influence. Déjà il y a les éditeurs, qui créent et mettent en ligne des contenus, lesquels sont ensuite référencés par le moteur de recherche de Google. Ces éditeurs peuvent choisir d'ignorer Google, mais ils peuvent aussi essayer de créer et de mettre en ligne leurs contenus de telle sorte qu'ils soient le mieux référencés possible par le moteur de recherche où ils espèrent apparaître en tête des listes de résultats. On parle d'optimisation des contenus, et de stratégies de "*search engine optimization*" (SEO). En plus du rôle qui consiste à hiérarchiser les pages web par ordre de pertinence, Google peut donc exercer une influence sur les contenus présents sur le web et, donc également, sur les messages, sur ce qui est dit, en plus d'exercer son influence sur le degré d'attention accordé à tel ou tel éditeur[1].

Le deuxième type d'acteur avec qui Google interagit, ce sont les annonceurs. Les annonceurs sont ces acteurs qui payent pour que leurs annonces figurent sur les services de Google ou de ses partenaires. Google fixe des conditions contractuelles de l'affichage des publicités sur ses différents services, une manière de calculer les prix, et met à la disposition des annonceurs des outils leur permettant de cibler leurs campagnes et de mesurer leur efficacité. Google peut donc influencer la stratégie des annonceurs et, finalement, le succès de telle ou telle campagne et de tel ou tel produit.

Le troisième type d'acteurs avec qui Google entretient des relations, ce sont, bien entendu, les internautes. Google met à leur disposition des services très nombreux et gratuits pour la plupart : un moteur

de recherche, une messagerie électronique, un navigateur, etc. Ainsi, la firme peut influencer ce que font et ce que peuvent faire les internautes, ce à quoi ils ont accès et les modalités de cet accès. Les ingénieurs de Google paramètrent en effet les fonctionnalités de ces outils, et de leurs actions dépendent donc les capacités (ou les incapacités) d'agir des internautes.

Enfin, Google entretient des relations avec les institutions : les représentants, les tribunaux, les autorités de la concurrence de tous les pays où ses outils sont accessibles, c'est-à-dire que la firme entretient des relations et négocie avec l'ensemble des pays dont les citoyens ont accès à Internet.

Analyser la stratégie "data" de Google

Si nous souhaitons décrypter la stratégie menée par Google en matière de données, il est nécessaire de considérer les synergies entre, d'une part, les activités de production de services mis à disposition des internautes gratuitement, et, d'autre part, les activités publicitaires. Trois points fondamentaux ressortent d'une telle considération simultanée.

Premier point : les services gratuits sont autant de plateformes sur lesquelles la firme peut distribuer de la publicité, étant donné le principe des marchés à deux versants[2] : mettre à disposition un service gratuitement dans le but d'attirer le plus d'individus possible (premier versant) et monnayer ensuite leur présence auprès des publicitaires à qui on propose que leurs encarts soient juxtaposés au service en question (deuxième versant). C'est ce que font Google, Facebook, la presse gratuite et un nombre considérable de services en ligne dont le développement repose sur la gratuité pour l'internaute et le financement par la publicité.

Deuxième point : les concurrents de Google ne sont pas seulement et pas principalement les moteurs de recherche, les messageries électroniques, etc. La principale source de revenus de Google provient du marché de la publicité. Autrement dit, Google est en concurrence avec toutes les entreprises dont l'activité est financée (pour partie ou totalement) par la publicité. Cela concerne aussi bien la publicité en ligne que la publicité hors ligne. Car en effet, le budget des annonceurs ne se multiplie pas sous prétexte que les supports se multiplient. C'est une des caractéristiques de la nouvelle économie, dont un des effets relativement pervers est qu'une entreprise peut faire faillite alors que la qualité du bien ou du service grâce auquel elle a attiré les internautes est sans égal. En effet, si une telle entreprise a décidé de fournir ce service gratuitement et de se rémunérer par la publicité, elle se trouvera en concurrence avec des acteurs dont le cœur de métier n'aura peut-être rien à voir avec le sien.

Troisième point : Google n'attire pas les publicitaires avec seulement le nombre d'utilisateurs de ses services, mais aussi avec la connaissance qu'il a de chacun de ces utilisateurs et le ciblage qu'il est capable d'opérer pour le compte de l'annonceur. Ce dernier en effet ne cherche pas tant à maximiser sa visibilité qu'à optimiser ses chances d'augmenter les ventes de son produit. Pour cela, il lui faut montrer ses publicités à des individus susceptibles d'être intéressés. Et pour identifier lesquels sont susceptibles d'être intéressés par quoi, il faut avoir récolté des données à propos des internautes. C'est ce que Google fait grâce à ses services. L'enjeu est crucial, étant donné la concurrence à laquelle la firme de Mountain View fait face sur le marché de la publicité.

Les données relatives aux requêtes effectuées sur le moteur de recherche Google sont particulièrement précieuses, parce qu'elles donnent des indications très claires concernant les centres d'intérêt des internautes, dès lors que ce sont eux qui ont exprimé directement ces centres d'intérêt, et parce que le fait que le moteur de recherche Google jouisse d'une position hégémonique signifie qu'aucun autre acteur ne dispose de données comparables.

L'objectif ici sera de décrire et décrypter la stratégie de Google vis-à-vis des données récoltées par son moteur de recherche. La firme pourrait se contenter de conserver ces données et d'effectuer le ciblage pour ses annonceurs. Cependant, il se trouve qu'elle communique ces données, mais qu'elle ne les communique qu'en partie, de manière à garder la main sur leur valeur. Nous essaierons de comprendre pourquoi elle les communique, et pourquoi cette communication est partielle. Dans une première partie, nous expliquerons quelles sont les données et pourquoi elles ont de la valeur. Dans une seconde partie, nous expliquerons comment Google agrège ces données pour publier des comptesrendus adressés à tous, mais sans communiquer les données dans le détail. Dans une troisième partie, nous verrons comment Google communique à chaque éditeur les données le concernant, mais comment l'entreprise, traite, agrège et tronque ces données. Nous verrons ainsi comment la firme publie les données mais en protège les détails dans un mouvement comparable à celui de l'anémone qui se déploie et se replie sur elle-même, de façon à ne montrer à son entourage immédiat, fait de proies et de prédateurs, que ce qu'elle veut bien leur montrer, pour mieux attirer les proies et écarter les prédateurs.

Collecte des intentions

Les requêtes effectuées sur le moteur de recherche sont des besoins exprimés directement par les internautes : besoin d'information (un sujet m'intéresse et je veux en savoir plus) ou de consommation (je cherche à acheter une planche de surf à un bon prix). Pour chaque requête, Google peut enregistrer les réponses aux questions suivantes : *quoi, quand, où, qui.*

Quoi, car il connaît les mots employés. *Quand* et *où*, puisqu'il connaît le moment et le lieu précis où chaque requête a été formulée. *Qui*, car il connaît l'adresse IP du terminal utilisé. Pour passer d'une adresse IP à un profil individuel, Google a en revanche besoin que l'internaute soit également utilisateur d'un service authentifié : compte Gmail, Google+, Youtube. Cela permet de *reconnaître* un individu quand bien même celui-ci utiliserait plusieurs terminaux (ordinateur professionnel, ordinateur personnel, tablette, smartphone, ordinateur de quelqu'un d'autre, etc.), et de *dissocier* les différents individus qui utilisent un même terminal (par exemple plusieurs membres d'une famille qui utiliseraient à tour de rôle le même ordinateur). Cela augmente considérablement la qualité du ciblage, et, donc, l'attractivité des services vendus par Google aux annonceurs publicitaires.

Intérêt pour les professionnels du marketing

Il est aisé de comprendre pourquoi ces informations (*quoi, quand, où, qui*), compilées dans ce qu'on appelle les "*query log*s", peuvent s'avérer utiles pour un expert en marketing ou en communication, quel que soit son domaine de prédilection. Il pourra positionner sa marque et ses produits sur certains mots en particulier, certains secteurs géographiques et viser certaines périodes temporelles. Les individus authentifiés, même s'ils n'ont pas tout de suite acheté le produit pour lequel ils ont marqué un intérêt, pourront être soumis à nouveau à des publicités pour ce produit, plus tard dans leur navigation, grâce à la technique dite du *retargeting* : on suit à la trace les internautes et on attend le bon moment pour leur proposer un bien ou un service dont on sait qu'il les intéresse puisqu'ils ont fait une requête à son sujet une heure, un jour, une semaine auparavant.

En plus de cela, Google a les moyens techniques de savoir sur quel lien l'internaute a cliqué, et si oui ou non il a été satisfait du résultat. S'il est revenu tout de suite sur la page du moteur pour faire une nouvelle requête ou cliquer sur un autre lien, c'est qu'il n'a pas trouvé son bonheur. Si en revanche il est resté sur la page vers laquelle le moteur l'a dirigé, c'est que celle-ci, probablement, l'aura intéressé. Enfin, Google a les moyens de savoir, dans certains cas, si l'internaute a acheté le produit qui l'intéressait une fois redirigé par le moteur vers le site où se produit était en vente.

Intérêt pour Google

Pour résumer, Google dispose d'éléments lui permettant d'émettre des hypothèses crédibles à propos de ce que tel ou tel internaute veut savoir ou acheter. Ces informations constituent un trésor pour les producteurs de contenus et les annonceurs : "une banque de données des intentions", comme dit John Battelle[3], qui rend particulièrement intéressant les dispositifs de ciblage que Google propose à ses clients annonceurs, et à ses partenaires éditeurs (car Google pratique à la fois la vente directe sur le marché de la publicité, en affichant des encarts et des liens sur ses propres services, et le rôle d'intermédiaire, en proposant aux éditeurs de vendre à leur place aux annonceurs les encarts disponibles sur leurs pages en échange d'une partie des bénéfices réalisés par ces encarts).

Tout le monde ne voit pas d'un bon œil le fait que Google détienne un tel trésor. Par exemple, selon les économistes Cédric Argenton et Jens Prüfer[4], le fait que Google dispose de ces données, ajouté au fait que son moteur de recherche jouisse d'une position hégémonique qui fait qu'aucun acteur ne dispose de données comparables, pourrait créer une barrière à l'entrée du marché et empêcher toute tentative de concurrence du moteur de recherche. La firme jouirait par conséquent d'un avantage décisif sur tous les concurrents avérés et potentiels, ce qui conduit les économistes à plaider pour une publication des *query log*s. D'après Argenton et Prüfer, une telle mesure permettrait que les moteurs de recherche se concurrencent sur le seul plan de la qualité de leurs algorithmes, ce qui augmenterait à la fois le dynamisme de l'innovation dans le secteur, la qualité des moteurs, le surplus des consommateurs et la création de valeur ajoutée.

Dans les faits, il se trouve que Google communique à propos de *query log*s de deux façons différentes, l'une générale l'autre particulière. La première consiste à communiquer les "*tendances de recherche*" au grand public, et la seconde à communiquer à chaque éditeur intéressé les données qui le concerne lui, seulement lui. Aucune de ces façons de faire ne permet cependant de répondre au problème soulevé par Argenton et Prüfer. Nous verrons que dans les deux cas, en même temps qu'ouverture, il y a traitement, agrégation, normalisation. C'est ce double mouvement — je donne mais je choisis, je publie mais je traite, je communique à condition d'arranger — que nous appelons la stratégie de l'anémone.

L'esprit du temps

Google communique les données enregistrées à propos des requêtes des utilisateurs de son moteur de recherche. L'entreprise a lancé pour cela le site *Google Trends* en 2006 et une version plus perfectionnée en 2008, nommée *Google Insights for Search* (en français : *Tendances de recherche*). Le service renseigne quant aux volumes liés à des termes employés lors de requêtes, dans des pays et durant des périodes que l'on peut spécifier. Les informations sont actualisées en temps réel depuis juin 2015. Il est également possible d'effectuer des comparaisons entre différents termes afin de savoir quels sujets et quels mots, dans quels pays, quand, ont été les plus employés par les utilisateurs du moteur. L'outil donne également, pour certains termes, les prévisions de recherche pour les mois à venir. Il est par exemple possible de connaître quelles actualités font l'objet, à un moment donné, sur une zone géographique définie, de requêtes sur le moteur. Les informations sont présentées sous forme de graphique, avec en abscisse le temps et en ordonnée la fréquence d'apparition du mot-clé ou du groupe de mots-clés. L'outil permet également de comparer les termes entre eux : savoir par exemple si les internautes utilisent plus volontiers le terme "téléphone" ou "mobile".

Au moment du lancement de *Google Trends*, la vice-présidente de Google, Marissa Mayer, expliquait que grâce à cet outil il serait possible de savoir, et de voir, "*à quoi les gens étaient en train de penser*". Prenons un exemple extrêmement trivial pour illustrer ces propos : il est possible de voir qu'aux États-Unis l'intérêt pour la "plage" (*beach*) est plus grand que l'intérêt pour la "neige" (*snow*) et que les *tendances de recherche* les concernant sont inversées sur l'année : on s'intéresse plus à la plage en été qu'en hiver, et l'on s'intéresse plus à la neige en hiver qu'en été, sachant qu'on s'intéresse malgré tout plus à la plage qu'à la neige, même en hiver (Fig. 1).

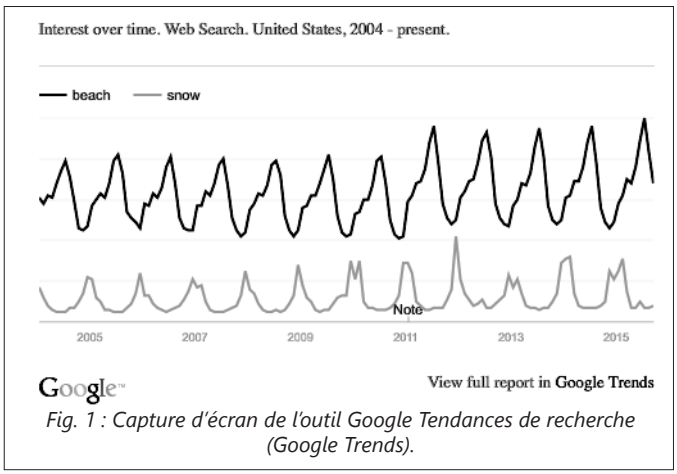

Fig. 1 : Capture d'écran de l'outil Google Tendances de recherche (Google Trends).

En plus des dirigeants de Google, certains observateurs ont fait les louanges de ce nouvel outil, qui selon eux permet de savoir en direct ce que font et ce que pensent les gens. Ce fut le cas par exemple de John Battelle, qui écrivit en 2005, à grand renfort de points d'exclamation, que Google était "*directement connecté au système nerveux de la culture*", et que le traitement des *query logs* constituait l'opportunité de savoir "*ce que la culture voulait*" (sic.). Ces propos étaient quelque peu excessifs, et plutôt naïfs lorsqu'ils présentaient la culture comme une entité homogène capable de volonté ; leur enthousiasme a toutefois le mérite d'indiquer assez clairement à quel point ces données récoltées par le moteur de recherche Google ont été perçues dès le milieu des années 2000 comme une nouvelle méthode pour comprendre les intérêts et les besoins des individus, fondamentalement différente des méthodes utilisées jusque-là dans les études de marché et les enquêtes d'opinion.

Les limites de la méthode

Le moteur de recherche agit en fait comme une sonde dont l'objectif serait de savoir ce que veulent les internautes et également de prévoir leur comportement avec, selon la firme, 12% de marge d'erreur en valeur absolue. Mais comme tous les sondages, la méthode de Google, aussi séduisante soit-elle, a ses défauts et ses zones d'ombre. Il est difficile dans bien des cas d'attribuer une intentionnalité aux requêtes effectuées sur le moteur. Si je fais une recherche sur "François Hollande", peut-être que je suis un sympathisant de sa politique, un opposant, ou peut-être encore que je n'ai aucun avis clair à ce sujet. Le fait que cela m'intéresse ne dit rien quant à *la nature* de mon intérêt.

Il est en outre important de noter que les données ne représentent pas tous les internautes, et encore moins "tous les gens" comme le suggère Marissa Mayer, mais les seuls internautes qui ont utilisé le moteur Google pour se renseigner sur un sujet en particulier. Ainsi, même si ces données sont intéressantes pour sonder ce qui intéresse les internautes, leur interprétation doit faire l'objet de grandes précautions, comme pour n'importe quel autre sondage. Rien ne permet d'affirmer en effet que les résultats publiés sur l'outil *Tendances de recherche* sont plus représentatifs que les résultats obtenus avec d'autres méthodes. Une étude a d'ailleurs été menée aux États-Unis, montrant que, dans le cas des élections du Congrès en 2008 et 2010, l'analyse des données fournies par *Google Trends* donnait des résultats moins pertinents que les méthodes "classiques" des enquêtes d'opinion[5].

Normalisation des données

Google ne donne pas accès aux données relatives à ses requêtes dans le détail. Nulle part sur l'outil *Tendances de recherche*, on ne peut télécharger les query logs. Ce à quoi on accède, c'est à un traitement des données, et plus exactement au résultat du traitement, car on ne connaît pas la méthode exacte employée par Google. C'est un peu comme si un statisticien communiquait le résultat de son enquête sans laisser avoir accès aux données brutes et sans rien dire, ou très peu, à propos de la méthode qu'il a employée.

De plus, l'information délivrée est normalisée et aucun chiffre en valeur absolue n'est jamais donné, comme cela est indiqué par Google sur son site : "Les nombres du graphique indiquent la quantité de recherches ayant été effectuées pour un terme donné, par rapport au nombre total de recherches effectuées sur Google au cours de la même période. Ils ne représentent pas le volume de recherche en valeur absolue, car les données sont normalisées et présentées sur une échelle allant de 0 à 100. Chaque point du graphique est divisé par le point le plus élevé ou par 100".

Tendances de recherche ressemble plus à un média qu'à un outil marketing, dès lors que les informations sont traitées avant d'être publiées. Nous voyons ici comment Google communique à propos des *query log*s, sans communiquer les *query log*s. Il y a ouverture mais il y a traitement, et il y a absence de transparence concernant ce traitement. Le chercheur qui veut utiliser les résultats publiés par *Tendances de recherche* n'a pas d'autre choix que de faire confiance aux employés de Google quand ils lui disent qu'ils n'ont pas manipulé les données.

Publications de rapports

Google publie des rapports à partir de l'analyse des *query log*s. Le plus connu de ces rapports est publié annuellement et se nomme Zeitgeist, qui signifie littéralement "esprit du temps". L'ambition est de rendre compte de "l'état d'esprit de l'époque", ainsi que l'indique la page d'accueil du service. Les sujets qui ont concentré le plus de requêtes chaque année sont présentés côte à côte, qu'ils soient liés à des sujets divertissants (cinéma, sport, musique), à des événements tragiques (guerres, terrorisme, catastrophes, épidémies), à des découvertes scientifiques majeures ou à des tendances de consommation. Des informations textuelles sont données à propos de chaque sujet pour expliquer le contexte, des photos sont publiées et un lien "Découvrir les tendances" renvoie vers l'outil *Tendances de recherche* où des requêtes ont été préprogrammées à propos du sujet concerné. Là encore, toutes les données sont normalisées. Impossible de savoir lequel, parmi les "sujets qui ont rendu cette année mémorable" (sic), a concentré le plus de requêtes. Impossible également de savoir s'il n'y a pas des sujets qui ont été évincés par Google, sous prétexte qu'ils concernaient par exemple des histoires jugées trop glauques (exemple : Luka Rocco Magnotta), ou des sujets jugés trop banals (ex. : la météo).

Google publie d'autres rapports, et crée des pages spéciales sur *Tendances de recherche* à propos de thème jugés (par les employés de Google) particulièrement pertinents. C'est ainsi que fut mis en place le dispositif Flu Trends dont l'objectif était de tracer les requêtes liées aux symptômes de la grippe, de manière à tenter de prévenir les épidémies. Idem avec Dengue Trends pour la grippe tropicale. Google avait également mis en place une page spéciale pour les élections présidentielles états-uniennes de 2016, où l'on apprenait quels étaient les sujets politiques qui intéressaient le plus les Américains, ce qui pouvait éventuellement aider les candidats à se positionner. (Fig. 2)

Google donne des informations pouvant intéresser directement les producteurs et les vendeurs de biens et de services. Ainsi, en avril 2015, un rapport à propos des tendances vestimentaires a été publié, dans lequel on apprenait que l'intérêt pour les jupes en tulle avait bondi de 34% entre janvier 2014 et janvier 2015, et que l'intérêt pour le jogging avait quant à lui augmenté de 165% en 2014.

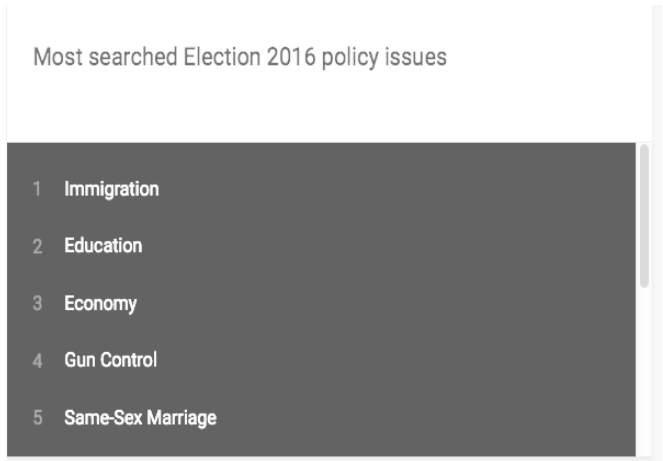

Fig. 2 : Les sujets politiques qui intéressaient le plus les utilisateurs de Google pendant la campagne américaine de 2016

Google publie également des données relatives aux requêtes effectuées durant des événements attirant beaucoup d'audience, et donc beaucoup de sponsors et d'annonceurs, comme par exemple le Super Bowl ou les Oscars. Enfin, Google a créé un site appelé "*Google Trends Data Store*" où il est

possible de télécharger les données (normalisées, comme les autres) concernant environ 80 sujets identifiés comme étant pertinents.

Toutes ces données sont susceptibles d'intéresser les éditeurs de contenus, les producteurs de biens et services, les vendeurs, les médecins, les sociologues, les politiques. Elles donnent des indications à propos de ce qui préoccupe les internautes, et à propos des mots que les internautes emploient pour évoquer ces sujets. Cela peut permettre de se positionner, ou de se repositionner, du point de vue éditorial et/ou commercial et/ou politique pour de nombreux acteurs de l'Internet. Google, en les publiant, peut en effet conduire différents acteurs à prendre certaines décisions stratégiques[6].

Certains éditeurs pourront par exemple choisir de traiter un sujet précis, parce qu'ils savent que ce sujet intéresse particulièrement les internautes et qu'il a donc de grande chance de générer un trafic important, ce qui peut conduire à cette dérive selon certains observateurs inquiets à l'idée que *Tendances de recherche* puisse dicter aux éditeurs leur ligne éditoriale[7]. Un producteur de vêtements pourra quant à lui décider de concentrer ses efforts sur le jogging étant donné le regain d'intérêt pour ce produit mentionné par Google dans son rapport concernant les tendances vestimentaires. Mais l'utilisation de ces informations demeure malgré tout problématique du point de vue épistémologique, dans la mesure où l'on n'a pas accès aux données exactes et où l'on ignore de quel traitement exact elles ont fait l'objet. "L'esprit du temps" pourrait en définitive n'être rien d'autre que "l'esprit de Google".

Les comptes-rendus personnalisés

N'importe quel propriétaire d'un site web a la possibilité d'installer un module Google Analytics, fourni gratuitement par Google pour les sites dont le trafic inférieur à dix millions de pages vues par mois. Ce module lui donnera des informations à propos de la fréquentation de son site et du comportement des internautes sur ses pages. Grâce à cette interface, l'éditeur pourra apprendre quels sont les mots-clés et les expressions qui ont permis d'accéder à son site depuis le moteur de recherche Google. Les chiffres concernant ces fréquentations lui seront donnés en valeur absolue. Notamment, *Google Analytics* fournit des données à propos de la provenance des internautes : combien sont arrivés par des moteurs de recherche, combien d'accès directs, etc. *Google Analytics* donne le pourcentage et le nombre exact. L'outil informe aussi à propos des mots-clés qui ont conduit au site, et communique à l'éditeur, pour chaque mot-clé, combien exactement d'internautes sont arrivés en utilisant celui-là, et quelle est, en moyenne, la durée de la visite sur le site déclenchée, ainsi que la moyenne du nombre de pages vues par les internautes arrivés avec ces mots-clés.

Contrairement au moteur de recherche Google, qui jouit d'une position hégémonique, *Google Analytics*, lancé en 2005 à l'issue du rachat par Google de la société Urchin Software Corporation, a des concurrents crédibles. C'est le cas par exemple des outils comme Xiti, Charbeat, Footprint, GoingUp, Piwik qui proposent des alternatives sérieuses à l'outil proposé par Google. Mais pour que les éditeurs préfèrent sa solution plutôt que celle des autres, Google dispose d'atouts que ses concurrents n'ont pas, notamment les données provenant de son moteur de recherche, dont nous allons expliquer comment et à quel point elles ont pu lui servir d'appâts.

Google a intérêt à ce que les éditeurs choisissent la solution *Google Analytics*, car cela lui permet d'avoir accès à des données concernant le comportement des internautes à l'intérieur des sites, qu'il n'aurait pas sans cela. De plus, *Google Analytics* sert de produit d'appel à *Google Adwords*, le service adressé aux annonceurs désireux de voir leurs liens s'afficher à droite des listes de résultats du moteur de recherche, et qui payent Google pour chaque clic effectué par un internaute sur un de ces liens dits "sponsorisés". Ainsi, sur le tableau d'administration de *Google Analytics*, il est écrit : "Augmentez le trafic ciblé de votre site, utilisez AdWords avec votre compte Google Analytics".

La stratégie a clairement été gagnante. En septembre 2015, selon l'outil W3Techs, *Google Analytics* était utilisé par environ un site web sur deux (52,2% des sites) et la totalité du trafic analysé par cet outil représentait 82,6%. Pour tous ces sites, et pour tout ce trafic, Google proposait donc de communiquer dans le détail les données relatives aux *query logs* les concernant. Cependant, une modification de la politique de Google effectuée progressivement à partir de 2011 est venue changer la donne, et de nombreux éditeurs se plaignent du fait que Google ait ainsi brisé une "entente tacite"[8].

Le repli sur soi

Le 18 octobre 2011, Google a annoncé que désormais, lorsqu'une requête serait effectuée sur une page sécurisée d'un de ses services authentifiés (dont l'adresse commence par https et non http), les informations concernant cette requête ne seraient plus communiquées aux éditeurs utilisant *Google Analytics*. Ces derniers sauraient toujours que des internautes sont venus depuis Google, et connaîtraient leur nombre, mais ne connaîtraient plus les mots-clés employés. Il leur faudrait se contenter d'une liste mensuelle des 1000 requêtes qui auraient amené

le plus grand nombre d'individus sur leurs pages. Autrement dit, à chaque fois qu'un internaute possédant un compte authentifié sur un service de Google (parce qu'il utilise par exemple le réseau social Google+ ou le service de messagerie Gmail) formulerait une requête et se rendrait sur le site d'un éditeur, ce dernier n'apprendrait plus rien à propos de cet internaute, sinon qu'une requête dont les termes seraient "not provided" ("non fournis") lui aurait permis d'accéder à une page en utilisant Google. Or, étant donné le nombre croissant d'internautes à disposer d'un compte authentifié sur un des services de Google, les informations concernant les requêtes étaient de plus en plus "not provided". Cette nouveauté fut présentée comme une manière de protéger la vie privée des utilisateurs de ses services.

En 2013, Google a annoncé que toutes les recherches effectuées sur son moteur le seraient désormais en mode sécurisé, ce qui revenait à étendre la part de "Not provided" aux internautes qui utilisaient le moteur Google, mais n'avaient pas de compte authentifié sur un des autres services de la compagnie. Les données divulguées par Webmaster Tools n'étaient quant à elle pas assez précises. Une fois de plus, il s'agissait de croire Google sur parole, sans moyen de vérifier, ni d'accéder au détail. Un expert explique : "Ces données viennent également de Google et quand on voit la stratégie actuelle du géant de Mountain View, on se dit que les risques de manipulation sont énormes. Les statistiques étant limitées à 90 jours (prochainement extensibles à 1 an selon Google) et 2 000 mots-clés, elles sont de toutes façons quasiment inexploitables, notamment pour les gros sites"[9].

Divulguer les données à ceux qui payent

Après ce mouvement de repli, le seul moyen pour un éditeur d'avoir accès à des données assez précises concernant les requêtes qui avaient conduit vers ses pages était de devenir client de Adwords, et donc de payer Google pour obtenir des informations auxquelles autrefois il avait accès gratuitement. Le client de AdWords savait en effet exactement quels mots-clés avaient conduit un internaute à voir et à cliquer sur son lien sponsorisé, et pouvait à partir de cette information faire des suppositions concernant les mots-clés les plus générateurs de clics pour ce qui est du référencement non payant. Ainsi, au nom du respect de la vie privée des utilisateurs de son moteur, Google avait trouvé le moyen de faire payer pour les données récoltées grâce à son moteur. C'est ce qui fit dire à certains observateurs que Google avait mis un prix sur la vie privée[10,11]. Olivier Andrieu, professionnel de l'optimisation des contenus pour le référencement sur Google, fait remarquer à ce sujet : "Lorsque 100% des requêtes naturelles seront chiffrées et non transmises, la seule solution pour obtenir des informations statistiques sur la façon dont les internautes ont trouvé un site sera donc… d'acheter des publicités Adwords. Et ce jour-là, Google aura gagné son pari : il sera devenu incontournable et aura fermé la porte à ses concurrents sans que personne ne réagisse…"[12].

Enfin, en avril 2014, Google annonça que le "Not provided" s'étendrait désormais aux clics effectués sur Adwords. Cette annonce provoqua un tollé chez les éditeurs, qui utilisaient ces données comme des ressources stratégiques pour savoir sur quels mots-clés se positionner. Beaucoup ne comprirent pas pourquoi Google avait fait cela. Quoi qu'il en fût, cela semblait prouver que Google n'utilisait pas le "Not provided" pour inciter les éditeurs à devenir des clients d'Adwords. En les empêchant d'accéder à ces données, Google pouvait également, sous couvert de protéger la vie privée des internautes, protéger le fonctionnement de son moteur en empêchant que les éditeurs ne s'adaptent aux *tendances de recherche* dans le seul but de capter davantage de trafic et soient prêts pour cela à créer des contenus non pertinents.

En regardant de près, on s'aperçoit que les données relatives aux requêtes n'ont pas été totalement retirées des mains des clients annonceurs. En effet, il se trouve que les données, qui ne sont effectivement plus disponibles depuis l'interface *Google Analytics*, restent disponibles depuis l'interface Adwords. Cela complique l'accès et le traitement des données, mais ne l'empêche pas. L'affirmation selon laquelle Google monnaye la vie privée de ses utilisateurs par le biais du "Not provided" est donc toujours aussi fondée en 2015 qu'elle ne l'était en 2013.

Au final, nous avons vu comment, là encore, Google procédait à une stratégie qui consiste à ne publier qu'une partie des données, sans fournir les données brutes ni communiquer la méthode exacte de traitement. Il y a eu dans les années 2010 un repli sur soi progressif, au nom de la protection de la vie privée des utilisateurs du moteur, qui a rendu hystériques de nombreux éditeurs. En février 2015, 85% des requêtes étaient "not provided" selon NotProvidedCount (outil fournissant des statistiques basées sur un panel de 60 sites).

Conclusion

Les données sont au cœur des stratégies déployées par les géants du web, car nombre d'entre eux financent leur activité grâce à la publicité. Facebook, Google, les sites de presse en ligne et une kyrielle d'acteurs dont les cœurs de métiers sont extrêmement différents souhaitent se financer en vendant des encarts à des annonceurs. Or, ces annonceurs ne sont pas

seulement intéressés par le nombre d'internautes qui verront leurs publicités, mais aussi, et surtout, par le ciblage de ces internautes, car c'est du ciblage que dépendent leurs chances de vendre effectivement leurs produits. C'est pourquoi les données relèvent d'un enjeu stratégique majeur.

Certains auteurs n'hésitent pas à parler d'un changement de paradigme, comme par exemple Yann Moulier-Boutang, qui nomme cela le processus de "pollinisation" : les internautes produisent des données en se déplaçant sur le web, et ce sont ces données et ces déplacements qui permettent à l'ensemble de porter ses fruits, tout comme des abeilles permettent en se déplaçant de transporter les pollens et les spores grâce auxquels un écosystème se maintient et produit[13,14]. L'auteur y voit un changement de paradigme fondamental. D'une économie de l'échange et de la production, nous serions passés à une économie de la pollinisation et de la contribution, centrée sur l'information, et propre à ce qu'il nomme le "capitalisme cognitif"[15].

Les données relatives aux requêtes des internautes constituent un véritable trésor pour le marketing. Google, étant donné la position hégémonique de son moteur, est seul à les posséder. Au lieu de seulement les garder en interne, Google a eu une idée bien plus subtile. La firme publie des informations à propos des données, sans publier les données elles-mêmes. En publiant ces informations, elle peut éventuellement orienter, ou contribuer à orienter, les actions des éditeurs de contenus et des producteurs/vendeurs de biens et services dans certains sens. Les éditeurs de contenus choisiront de traiter tel ou tel sujet parce qu'ils savent que ça intéresse les utilisateurs de Google, les producteurs choisiront de produire telle ou telle marchandise parce qu'ils savent que c'est ce que cherchent les utilisateurs de Google. La firme pourrait même influencer, dans une certaine mesure, la communication politique. Google a ainsi réussi à se faire chambre-écho de la demande de ses utilisateurs, sans pour autant communiquer les précieuses données concernant leurs requêtes. Cependant, les acteurs sont obligés de croire Google à propos de "l'esprit du temps", car il ne dispose ni des chiffres exacts, que Google ne leur communique pas, et ignorent tout de la méthode de traitement qui a permis à Google de publier les différents rapports, car la firme ne dit rien à son sujet. Google a ainsi réussi, sans divulguer les données, à maximiser le pouvoir qu'elle pouvait lui conférer.

A l'échelon individuel, Google a réussi à faire adhérer un maximum d'éditeurs à *Google Analytics*, en leur donnant les données à propos des requêtes qui avaient effectivement conduit des internautes vers leurs pages. Les données étaient ainsi très largement incomplètes, mais communiquées dans le détail. Puis nous avons expliqué comment Google avait peu à peu dissimulé ces données, en dissimulant d'abord celles des utilisateurs authentifiés de ses services, puis, dans un second temps, de tous les utilisateurs de son moteur. Sous couvert de protéger la vie privée de ses utilisateurs, Google pouvait ainsi inciter les éditeurs à devenir des partenaires d'AdWords, c'est-à-dire des clients de Google, qui paieraient pour apparaître dans les listes de liens sponsorisés et obtenir des données susceptibles de remplacer celles que *Google Analytics* ne fournissait plus. Là encore, on voit comment la stratégie qui consiste à ne pas dissimuler totalement les données relatives aux requêtes, mais à les distiller, à les traiter, à les tronquer, à les reprendre après les avoir publiées pendant plusieurs années, s'avère payante stratégiquement pour Google.

Nous pensons que le cas de Google est exemplaire d'une tendance sur le web qui consiste à ne rien cacher tout à fait, sans ne rien dévoiler entièrement. Les données, quand elles sont susceptibles d'avoir une valeur économique, ne sont que très rarement publiées dans le détail, mais elles ne sont pas enfouies pour autant. Au contraire, les acteurs en possession de données sensibles ménagent la chèvre et le chou, en tirant le meilleur du pouvoir qu'est susceptible de leur apporter une publication partielle, sous une forme qui ne peut pas donner lieu à des traitements significatifs, sans pour autant risquer de renoncer à la valeur monétaire qu'ont ces données, pour lesquelles il existe des acteurs prêts à payer très cher et à qui il serait par conséquent dommage de les confier gratuitement.

Guillaume Sire
Institut Français de Presse
Centre d'Analyse et de Recherche Interdisciplinaire sur les Médias
5/7, Avenue Vavin
75006 Paris
France
guillaume.sire@u-paris2.fr
https://noumerika.wordpress.com/

Mai 2017

Notes

1. Sire, Guillaume. Le pouvoir normatif de Google. Analyse de l'influence du moteur sur les pratiques des éditeurs. *Communication & Langages*, n°188, 2016, p. 91-105.
2. Un marché multiversant est une plate-forme réunissant plusieurs groupes d'agents distincts sur des versants qui leur sont propres, chacun de ces groupes ayant la particularité d'avoir potentiellement intérêt à interagir avec les agents du groupe d'au moins un autre versant. Ainsi, la présence d'agents sur le versant n°2 rend plus attractif le bien vendu sur le versant n°1, l'inverse étant également possible. Il s'agit de ce que les économistes nomment "des effets de réseaux croisés".
3. Battelle, John. *The search: how Google and its rivals rewrote the rules of business and transformed our culture.* Nicholas Brealey Publishing, 2005.
4. Argenton, Cédric ; Prüfer Jens. Search engine competition with network externalities. *Journal of Competition Law and Economics*, 2012, vol. 8, n°1, p. 73-105.
5. Lui, Catherine; Metaxas, Panagiotis; Mustafar, Eni. On the predictability of the U.S. elections through search volume activity, [en ligne], 2011 (consulté le 9 mai 2017). <http://cs.wellesley.edu/~pmetaxas/e-Society-2011-GTrends-Predictions.pdf>
6. Voir, par exemple, l'influence de Google sur la presse : Sire, Guillaume. *Google, la presse et les journalistes. Analyse interdisciplinaire d'une situation de coopétition*, Bruxelles, Bruylant/Concurrences, coll. "Sciences Politiques", 2015.
7. McBride, Kelly. SEO Makes It Too Late for Truth for "Ground Zero Mosque". *Poynter* [en ligne], 2011 (consulté le 9 mai 2017). <http://www.poynter.org/2010/seo-makes-it-too-late-for-truth-for-ground-zero-mosque/105201/>
8. Lavoie, Samuel. Google (not provided) : le mot-clé utilisé pour trouver cet article n'existe plus. *Adviso*, [en ligne], 2013 (consulté le 9 mai 2017). <http://www.adviso.ca/blog/2013/10/02/google-not-provided/>
9. Andrieu, Olivier. Bientôt 100% de (not provided) ? Quand Google-Big Brother veut écraser le Web… *Abdondance*, [en ligne], 2013 (consulté le 9 mai 2017).

   <http://www.abondance.com/actualites/20130924-13176-bientot-100-de-not-provided-quand-google-big-brother-veut-ecraser-le-web.html>
10. Charlton, Graham. Google's keyword data apocalypse: the experts' view. *Econsultancy*, [en ligne], 2013 (consulté le 9 mai 2017) <https://econsultancy.com/blog/63460-google-s-keyword-data-apocalypse-the-experts-view/>.
11. Sullivan, Danny. Google Puts A Price On Privacy. Search Engine Land, [en ligne], 22 octobre 2011 (consulté le 9 mai 2017) <http://searchengineland.com/google-puts-a-price-on-privacy-98029>
12. Cfr note 7.
13. Moulier Boutang, Yann. Inescapable Google ? Organization of kowledge, Economic valua in cognitiv capitalism, and collective intelligence. *Conférence Society of Query, stop searching, start questioning*, Université d'Amsterdam, 2009.
14. Moulier Boutang, Yann. *L'abeille et l'économiste*, Paris, Carnets Nords, 2010.
15. Moulier Boutang, Yann. *Le capitalisme cognitif : La Nouvelle Grande Transformation*. Editions Amsterdam, coll. "Multitude/Idées", 2007. Il est intéressant de rapprocher le postulat de Yann Moulier-Boutang, selon lequel nous aurions changé de capitalisme, des conclusions récentes du rapport rendu au gouvernement français le 18 janvier 2013 par le membre du Conseil d'État Pierre Collin et l'inspecteur des finances Nicolas Colin. Ce rapport préconise en effet qu'une nouvelle fiscalité soit progressivement mise en place afin de s'adapter aux modèles d'affaires d'entreprises comme Google dont le modèle économique est essentiellement basé sur la collecte et le traitement de données relatives au comportement des internautes. Ainsi, si nous exprimons les conclusions de Collin et Colin dans le langage de Moulier-Boutang, nous dirions que l'environnement fiscal, hérité de l'ancien paradigme capitalistique, n'est pas adapté à l'économie de la pollinisation.