Internet. Géopolitique de la donnée

Maîtriser la donnée : enjeux et défis géopolitiques. Moteurs de recherche et web profond

mercredi 3 juin 2015, par Thierry BERTHIER

Pourquoi la donnée peut-elle être considérée comme une ressource qui, une fois exploitée, créé de la valeur et de la puissance ? L'auteur répond clairement en présentant successivement comment les moteurs de recherche sont des vecteurs de puissance, mais aussi les défis du web profond et la dépendance excessive de l'Union européenne.

LE VOLUME MONDIAL des données numériques pourrait atteindre en 2020 les 40 Zo (un Zetaoctet est égal à dix puissance vingt-et-un octets). L'évolution de cette production est exponentielle puisque 90% des données actuelles ont été produites durant les deux dernières années et que cette tendance s'accélère. Les activités humaines et les systèmes automatisés sont à l'origine du déluge de données qui transforme la texture même de l'environnement. L'information ubiquitaire (produite partout, tout le temps, pour tous) émerge des villes intelligentes, des objets et bientôt des corps connectés. Elle fait le lien entre les espaces physique et numérique qui s'intriquent de plus en plus sous l'effet des interactions algorithmiques. Souvent comparée à un nouveau pétrole, la donnée peut être considérée comme une ressource qui, une fois exploitée, créé de la valeur et de la puissance. La collecte et l'interprétation des mégadonnées apparaissent alors comme les défis technologiques et stratégiques majeurs pour des nations en concurrences.

Le moteur de recherche, vecteur de puissance et de souveraineté nationale

La numérisation des activités humaines, de l'économie, du savoir, du corps biologique, engendre un flux continu de données à la fois hétérogènes et souvent peu structurées qui doivent être stockées et traitées par des infrastructures physiques toujours plus puissantes. Les systèmes de télécommunication, les centres de stockage (Data Center) et les centres de traitement des données enregistrent une forte croissance d'activité. La consommation en électricité de l'ensemble des « Data Centers » dépasse actuellement celle de la France. Une économie de rupture fondée sur l'exploitation de la donnée émerge et vient modifier les équilibres traditionnels. **En réduisant les effets de l'aléatoire, la donnée produit de la connaissance et de la richesse**. Elle permet de comprendre et de maîtriser des phénomènes complexes qui échappaient jusqu'à présent au contrôle humain. Si l'industrie de la donnée est fortement dominée par les États-Unis, l'Asie se donne aujourd'hui les moyens de rattraper une partie de son retard, la Russie et la Chine cherchent quant à elles à conserver une forme de souveraineté et de contrôle sur leurs données nationales. Les concurrences et duels géopolitiques se projettent ainsi naturellement sur le terrain informationnel et les données incarnent plus que jamais les vecteurs de puissance d'économies tournées vers la connaissance. Les données internationales ouvertes (accessibles) représentent des gisements particulièrement attractifs pour les grands acteurs mondiaux du numérique, qu'ils soient étatiques, industriels ou institutionnels. **L'indexation de ces données par les moteurs de recherche apparaît alors comme le premier défi à relever face au déluge des données ouvertes**.

La carte ci-dessous spatialise la hiérarchie de l'Internet.

La carte au format pdf haute qualité

Vers une guerre des moteurs ?

Les grandes nations technologiques ont pris en compte depuis longtemps les enjeux stratégiques de l'indexation des contenus numériques. Peu bruyante, une « guerre » des moteurs de recherche a bien lieu aujourd'hui, épousant scrupuleusement les contours des conflits et des concurrences de souverainetés nationales.

Ces chiffres montrent que seulement **quelques moteurs se partagent l'essentiel du marché mondial de l'indexation**. Cette très forte concentration sur les plus grands moteurs est renforcée par le principe bien connu du « gagnant qui rafle tout » puisque la qualité des services de référencement augmente toujours avec la taille (et le rang) des moteurs de recherche. **Jamais autant d'information n'a été détenue et exploitée par aussi peu d'acteurs** et certaines nations (dont la France) affichent une dépendance forte à un unique moteur. La Chine avec le moteur Baidu a su construire très tôt son indépendance informationnelle face au géant américain. Aujourd'hui, plus de 500 millions d'internautes utilisent quotidiennement Baidu à partir d'une centaine de pays. La Russie utilise massivement le moteur de recherche Yandex qui ne laisse que peu de place à Google sur le secteur du référencement intérieur russe puisqu'il détient plus de 60% des parts du marché national. En 2014, Vladimir Poutine a souhaité que son pays développe un second moteur de recherche exclusivement contrôlé par des capitaux russes et sans aucune influence extérieure. Plus récemment, en février 2015, le groupe Yandex a déposé une plainte contre Google en Russie pour abus de position dominante sur les smartphones Android. Yandex reproche en effet à Google de bloquer l'installation de ses applications de moteur de recherche sur les smartphones fonctionnant sous Android. Les constructeurs sont contraints aujourd'hui de pré-installer sur leurs machines les Google Apps et d'utiliser Google comme moteur par défaut sous Android… Ce type de concurrence et de duel intervient de manière récurrente lorsqu'il est question du gisement des données nationales ouvertes et bien référencées. Désormais, il en est de même avec les données du web profond, moins accessibles car non adressées par les grands moteurs de recherche.

Les enjeux du web profond

Le web profond (Deep Web) désigne le sous-ensemble d'internet qui n'est pas indexé ou mal indexé par les grands moteurs de recherche comme Google, Yahoo ou Bing,…On sait que cet ensemble de données reste difficilement mesurable mais qu'il occupe un espace très supérieur à celui de l'ensemble des sites web bien indexés par les moteurs classiques. Certaines études avancent un ratio de 80% de Deep Web contre 20% de web de surface à l'image de la partie immergée d'un iceberg… Le contenu du Deep web

demeure hétérogène. On y trouve de grandes bases de données, des bibliothèques volumineuses non indexées par les moteurs en raison de leur taille, des pages éphémères, mal construites, à très faible trafic ou volontairement rendues inaccessibles aux moteurs traditionnels par leurs créateurs. D'après une étude récente de la Darpa, l'agence américaine en charge des projets de défense, **plus de 60 millions de pages à vocation criminelle ont été publiées depuis deux ans dans les profondeurs du web**. Les moteurs de recherche classiques, Google en tête, utilisent des algorithmes d'indexation dérivés du puissant Pagerank qui s'appuient sur une mesure de popularité du site ou de la page. Cette approche qui a fait le succès de Google va de fait exclure les pages à faible trafic, éphémères ou furtives. Ce sont précisément ces pages qui sont utilisées par les acteurs de la cybercriminalité pour diffuser de l'information tout en restant sous les radars des grands moteurs. Lorsque cette information concerne une activité criminelle, c'est dans le Dark Web qu'elle sera dissimulée et rendue accessible aux seuls clients potentiels via des outils d'anonymisation spécialisés comme Tor. Le web profond réunit donc de la donnée légitime, souvent de haute qualité lorsqu'il s'agit de bases de données scientifiques volumineuses peu ou mal indexées par les moteurs. Il réunit de la donnée sécurisée accessible seulement par mot de passe mais aussi de la donnée clandestine issue de trafics et d'activités criminelles. Cet ensemble informationnel hétérogène intéresse depuis longtemps les grands acteurs du numérique, chacun avec une motivation spécifique. L'accès au web profond constitue un élément stratégique du dispositif global de lutte contre la cybercriminalité qui reste l'une des grandes priorités de l'administration américaine. Les efforts pour obtenir des capacités de lecture du web profond se sont concrétisés avec le développement en 2014 du moteur de recherche Memex tout droit sorti des laboratoires de la Darpa.

Memex, le moteur Darpa

Dans son communiqué officiel publié le 9 février 2014 [1], l'agence Darpa décrit Memex comme « le moteur qui révolutionne la découverte, l'organisation et la présentation des résultats de recherche en ligne. Le programme Memex imagine un nouveau paradigme, où il est possible d'organiser rapidement et intelligemment un sous-ensemble de l'internet adapté à l'intérêt d'une personne ». Le moteur est construit autour de trois axes fonctionnels : 1 - l'indexation de domaines spécifiques, 2 - la recherche de domaines spécifiques et 3 - la mise en relation de deux premiers axes. Après plus d'un an d'utilisation en phase de test par les forces de l'ordre américaines, Memex a permis de démanteler un réseau de trafiquants d'êtres humains. Durant la finale du Super Bowl, Memex a servi pour détecter les pages associées à des offres de prostitution. Ses outils d'analyse et de visualisation captent les données invisibles issues du web profond puis tracent le graphe des relations liant ces données. De telles fonctionnalités s'avèrent très efficaces pour cartographier des réseaux clandestins de prostitution en ligne. D'après les récents communiqués de la Darpa, Memex ne traite pour l'instant que les pages publiques du web profond et ne doit donc pas être associé aux divers outils de surveillance intrusifs utilisés par la NSA. A terme, Memex devrait offrir des fonctionnalités de *crawling* du Dark Web intégrant les spécificités cryptographiques du système Tor. On peut raisonnablement imaginer que ces fonctions stratégiques faisaient bien partie du cahier des charges initial du projet Memex dont le budget est estimé entre 15 et 20 millions de dollars… La Darpa n'est évidemment pas seule dans la course pour l'exploration du web profond. Google a parfaitement mesuré l'intérêt informationnel que représentent les pages non indexées par son moteur et développe de nouveaux algorithmes spécifiquement adaptés aux profondeurs du web.

Google et le défi des profondeurs

Le web profond contient des informations provenant de formulaires et de zones numériques que les administrateurs de sites souhaitent maintenir privés, hors diffusion et hors référencement. Ces données, souvent très structurées, intéressent les ingénieurs de Google qui cherchent aujourd'hui à y avoir accès de manière détournée. Pour autant, l'extraction des données du web profond demeure un problème algorithmiquement difficile et les récentes publications scientifiques des équipes de Google confirment bien cette complexité. L'Université de Cornell a diffusé un article remarquable décrivant une infrastructure de lecture et de copie de contenus extraits du web profond [2], [3]. L'extraction des

données s'effectue selon plusieurs niveaux de *crawling* destinés à écarter les contenus redondants ou trop similaires à des résultats déjà renvoyés. Des mesures de similarités de contenus sont utilisées selon les URL ciblées pour filtrer et hiérarchiser les données extraites. Le système présenté dans l'article est capable de traiter un grand nombre de requêtes sur des bases de données non adressées par le moteur de recherche classique de Google [4]. **A moyen terme, les efforts de Google permettront sans aucun doute de référencer l'ensemble du web profond publiquement accessible**. Le niveau de résolution d'une requête sera fixé par l'utilisateur qui définira lui même la profondeur de sa recherche. Seuls les contenus privés cryptés ou accessibles à partir d'une identification par mot de passe demeureront (en théorie) inaccessibles à ce type de moteurs profonds. En tant que leader mondial du référencement de contenus, Google ne peut faire l'impasse sur les données du web profond. Là encore, ce sont ses algorithmes qui feront la différence avec ceux de ses concurrents. Le niveau d'intelligence artificielle embarquée dans le moteur permettra de prendre l'ascendant et de conserver l'avantage, toujours selon le principe du « gagnant qui rafle tout ».

Une Union européenne dépendante

L'absence d'un grand moteur de recherche européen provoque une dépendance fonctionnelle à Google et écarte tout espoir de souveraineté nationale en la matière. Ainsi, plus de 92 % des requêtes françaises en 2013 et 2014 ont été effectuées sur le moteur américain. **La France, qui fait partie du groupe des pays les plus « Google-dépendants » du monde**, avait pourtant lancé en mars 2008 le programme de recherche et développement Quaero avec l'objectif initial de construire un Google européen. Achevé cinq ans plus tard pour un coût total de 198 millions d'euros, Quaero a dû revoir son ambition à la baisse : point de moteur de recherche disruptif opérationnel en fin de programme mais 35 prototypes assez prometteurs qui ont donné lieu à plusieurs déploiements industriels et commerciaux. L'entreprise publique Oséo a investi près de cent millions d'euros dans le projet, rejointe par 32 entreprises partenaires dont Orange et Technicolor. **Le défi technologique de construction d'un concurrent de Google ne peut raisonnablement se concevoir aujourd'hui qu'à une échelle européenne**, à l'image des grands programmes de recherche mutualisés du LHC ou du Human Brain Project de l'EPFL en Suisse. Un projet de grand moteur pourrait émerger d'une volonté politique européenne affirmée visant à rééquilibrer la souveraineté informationnelle communautaire. Le degré actuel de dépendance de l'UE aux infrastructures d'indexation américaines créé **une fragilité systémique et des vulnérabilités incompatibles avec ses ambitions stratégiques et géopolitiques**. Le refus d'une vassalisation technologique passe par la construction de Data Centers et d'infrastructures de traitement de la donnée. La Chine et la Russie se sont d'ores et déjà engagées vers une indépendance informationnelle partielle. Que fait l'UE ?

P.-S.

Chaire de Cybersécurité & Cyberdéfense Saint-Cyr – Thales

Notes

[1] La présentation du moteur Memex par l'agence Darpa
http://www.darpa.mil/newsevents/releases/2014/02/09.aspx

[2] « Google's Deep-Web Crawl » - publication de l'Université Cornell
http://www.cs.cornell.edu/~lucja/publications/i03.pdf

[3] « Crawling Deep Web Entity Pages » - publication de recherche, Google
http://pages.cs.wisc.edu/~heyeye/paper/Entity-crawl.pdf

[4] « How Google May index Deep Web Entities »
http://www.seobythesea.com/2015/04/how-google-may-index-deep-web-entities/