Enjeux et défis du web profond | Le Net Expert Informatique

Enjeux et défis 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…

Profond comme le web

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 tailles, des pages éphémères, mal construites, à très faible trafic ou volontairement rendues inaccessibles par leurs créateurs aux moteurs traditionnels.
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
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 la 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.

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 les concurrences de souverainetés nationales. La Chine avec son 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 à pré-installer sur leurs machines les Google Apps et à utiliser Google comme moteur par défaut sous Android…

Le moteur face aux mégadonnées

La course à l'indexation des contenus du web profond apparaît comme l'une des composantes stratégiques de la guerre des moteurs. Si la géopolitique des données impose désormais aux nations de définir des politiques claires de stockage et de préservation des données numériques, elle commande également une vision à long terme de l'adressage des contenus. La production mondiale de données dépassera en 2020 les 40 Zo (un zettaoctet est égal à dix puissance vingt et un octets). L'évolution de cette production est exponentielle: 90% des données actuelles ont été produites durant les deux dernières années. Les objets connectés, la géolocalisation, l'émergence des villes intelligentes connectées et de l'information ubiquitaire contribuent au déluge de données numériques. La collecte et l'exploitation des mégadonnées (le terme officiel français à utiliser pour big data) induiront le développement de moteurs polyvalents capables d'indexer toutes les bases de données publiques quelle que soient leurs tailles et leurs contenus.

Le moteur de recherche doit être considéré aujourd'hui comme une infrastructure de puissance stratégique au service des nations technologiques. Qu'attend l'Europe pour développer le sien?

_____

[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 »
How Google May Index Deep Web Entities

Cet article vous plait ? Partagez !
Un avis ? Laissez-nous un commentaire !

Source : http://www.huffingtonpost.fr/thierry-berthier/enjeux-et-defis-deep-web_b_7219384.html
Par Thierry Berthier