Les moteurs de recherche

Culture numérique – Lille 3

Des ressources qui n'existent que quand on les demande...

Prenons l'exemple de l'URL suivante :

`http://www.univ-lille3.fr/etudes/orientation-emploi/`.

Rappelons que la partie `etudes/orientation-emploi` désigne une ressource sur le serveur web `www.univ-lille3.fr`. Il est possible que ce soit un document composé par une personne du service des études puis enregistré sur les disques durs de ce serveur web pour le mettre à disposition des internautes. Mais à vrai dire, c'est un processus de conception à la mise en ligne de ressources aujourd'hui de plus en plus rare. Dans le web moderne, de plus en plus souvent, ces ressources sont composées par des programmes informatiques, à partir d'éléments pris dans de nombreuses sources de données. Ces programmes sont par exemple des outils de publication web, systèmes de gestion de contenu (CMS en anglais), des wiki, des moteurs de blogs...

Mais un autre exemple évident de la génération automatique de ressources est celui des moteurs de recherche. Lorsque vous appuyez sur le bouton de recherche après avoir saisi vos mots clefs, le document qui apparaît dans votre navigateur a évidemment été construit juste pour vous, au moment de votre demande.

Un annuaire de toutes les ressources

Le web est un immense ensemble de ressources reliées entre elles. On pouvait imaginer à ses débuts parcourir cet ensemble et trouver son chemin vers la ressource souhaitée. On a donc commencé à construire des annuaires et des répertoires à l'image de ce qui peut se faire dans des bibliothèques. Tim Berners Lee, inventeur du web, a même maintenu une liste de serveurs web à cette époque. Mais cet idéal a

rapidement été abandonné. La taille du web a grandi tellement vite qu'il est devenu impossible de consigner les adresses de toutes les ressources, ou même seulement les plus importantes. C'est alors que sont entrés en jeu les moteurs de recherche.

Comment fonctionne un moteur de recherche aujourd'hui

Comment fonctionne un moteur de recherche ? C'est à la fois simple dans certains principes généraux et complexe pour de nombreux détails importants. C'est à la fois connu dans sa généralité et bien caché dans ses détails. Nous nous contentons ici de simples généralités.

Les moteurs de recherche construisent constamment, car le web évolue sans cesse, un index. L'index, c'est comme dans un livre, un moyen d'aller directement à une page à partir d'un mot. Pour construire un tel index, il faut avoir lu toutes les pages du livre et consigné pour tous les mots, la liste des pages où ils se trouvent. Les moteurs de recherche téléchargent toutes les ressources du web en permanence pour extraire la liste des mots qu'on y trouve et garder l'énorme liste des URLs où ces mots se trouvent. Ce ne sont pas des hommes qui parcourent le web pour eux, mais des programmes, appelés des robots. Les robots sont les clients des serveurs web les plus nombreux et réguliers... et de loin !

Mais afficher simplement la liste de ces ressources quand l'internaute saisit quelques mots dans le formulaire de recherche n'est pas satisfaisant. La liste est bien trop longue. Le deuxième ingrédient du moteur de recherche est le programme qui permet d'interroger cet index, simplement en lui donnant quelques mots, et qui construit une liste, présentée par ordre d'importance, d'URLs désignant les ressources où ces mots se trouvent.

La magie des moteurs de recherche tient dans les détails qui permettent à l'ensemble de fonctionner tels que l'existence d'un index à jour, la forme de l'index qui permet d'y retrouver extrêmement rapidement les pages associées à un mot, ou encore l'ordre d'importance dans lequel les résultats de l'interrogation de l'index apparaissent.

L'avance technologique des grands moteurs de recherche se cache dans les détails de la construction de l'index mais surtout du programme qui permet de l'interroger et de la détermination de l'ordre des URLs affichées en retour. Ces détails sont protégés par de nombreux secrets industriels.

Collecte de données d'usage

Mais un avantage qui rend la mise en concurrence des grands moteurs de recherche actuels presque impossible tient à un dernier paramètre. C'est la disponibilité d'énormes quantités de données d'usage, parfois personnalisées. En effet le résultat (l'ordre d'apparition des ressources) des requêtes au moteur dépend aujourd'hui fortement de ce qu'ont fait leurs utilisateurs : sur quels liens ont-ils cliqué ? À l'inverse des ressources du Web derrière les URLs, ces données d'usage ne sont pas publiques, mais sont tout aussi cruciales pour générer des réponses aux requêtes dans un ordre pertinent.

En conséquence, les moteurs de recherche collectent sans cesse des données à propos de vos recherche. La tendance actuelle est de rendre les réponses personnalisées, ce qui entraîne une collecte de données personnelles rendue possible à la fois par les techniques de cookies et l'utilisation de comptes chez ces opérateurs de recherche.

Modèle économique du moteur de recherche

Pour une institution qui veut être visible sur internet, if faut assurer sa présence dans l'index. Mais cela n'est pas suffisant : il faut être en haut de la liste et donc apparaître important aux yeux du moteur de recherche.

De bonnes pratiques en matière de conception de pages web peut y contribuer. Puisque toute la chaîne de traitement est automatique, les ressources que le moteur analyse et indexe doivent être parfaitement intelligibles par la machine. Il est donc très important d'écrire correctement ses pages web dans ce but de traitement automatisé autant que dans le but de se faire comprendre de ses lecteurs humains. Parfois des conseillers un peu charlatans tentent de se faire passer pour des gourous qui vont propulser des sites en première page des résultats de recherche.

Il faut s'en méfier car pour le moteur de recherche, une des premières sources de revenu est de vendre ces places. Cela se traduit littéralement par des *ventes de mots*. Une deuxième source de revenu est liée à la collecte des données personnelles des utilisateurs. Tirer des informations à l'insu ou non de ses usagers n'est pas une pratique réservée aux moteurs de recherche. De nombreux autres acteurs du web fonctionnent sur ce même principe.

Aller plus loin

Cette petite introduction des moteurs de recherche est volontairement très succinte et parcellaire. Des éléments techniques essentiels ne sont pas mentionnés comme :

- les pré-traitements des textes et la sélection du vocabulaire, le traitement des majuscules, des accents etc...
- le calcul du score de pertinence sur lequel repose cet ordre d'affichage des réponses, et bien-sûr
- l'un des algorithmes les plus connus qu'est PageRank utilisé par Google.

Nous vous invitons à suivre les cours d'option transversale en licence, les options de master sur les humanités numériques, ou les prochains cours de culture numérique qui aborderont sans doute ces questions beaucoup plus précisément.