A. Introduction

Les moteurs de recherche sont aujourd'hui des outils incontournables pour trouver simplement et rapidement des informations pertinentes sur le Web. Ils sont également omniprésents dans les activités de veille et notamment pour :

- le ciblage, afin de tester le plan de veille, identifier de nouveaux mots-clés, prendre rapidement connaissance d'un domaine peu ou pas connu, etc. ;
- le sourcing, pour trouver de nouvelles sources d'information ;
- la collecte, pour identifier manuellement des informations potentiellement pertinentes ;
- le traitement, pour trouver des informations permettant de recouper et qualifier celles qui ont été collectées automatiquement.

Il est donc important dans le cadre de la veille de bien comprendre le fonctionnement des moteurs de recherche, de savoir lesquels utiliser et dans quels cas, mais aussi et surtout, d'appliquer les bonnes pratiques qui permettront d'en tirer le meilleur parti.

B. Panorama des moteurs de recherche web

1. Introduction

La vocation d'un moteur de recherche web est de rendre accessibles et « trouvables » des pages web ainsi que les ressources qui leur sont associées (documents, images, vidéos, podcasts…).

Devenu incontournable, au point d'être, pour certains, le seul moteur de recherche sur le Web, Google est en fait l'arbre qui cache la forêt. En effet, Google, du moins dans son utilisation la plus populaire, n'est que le représentant d'une catégorie de moteurs de recherche web : les moteurs de recherche généralistes. Il en existe d'autres : métamoteurs, moteurs de recherche verticaux ou spécialisés…

Mais avant de dresser un panorama des moteurs de recherche web disponibles et des méthodes pour les utiliser au mieux, il est nécessaire de s'arrêter quelques instants sur leurs grands principes de fonctionnement afin de mieux cerner leurs forces et leurs faiblesses, notamment dans le cadre d'une activité de veille.

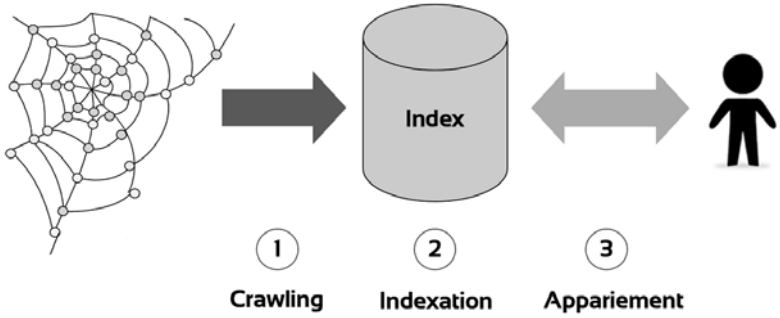
Bien souvent, lorsque l'on évoque les moteurs de recherche web, les annuaires sont également cités. Ces derniers sont d'excellents outils de recherche dans la mesure où ils référencent des sources qui ont été identifiées à la suite d'un traitement manuel. Cependant, leur mode de fonctionnement est très différent des moteurs de recherche et ne seront pas présentés ici. Très utiles à la veille, ils seront abordés dans le chapitre suivant consacré au sourcing.

2. Fonctionnement d'un moteur de recherche sur le Web

a. Grands principes

Un moteur de recherche web repose généralement sur trois grandes fonctions :

- Le **crawling**, qui consiste à identifier et collecter des données sur les contenus disponibles sur le Web et que l'on souhaite rendre accessible via une recherche.
- L'**indexation**, qui vise à extraire les informations les plus importantes de ces contenus.
- Et l'**appariement**, qui permet de présenter une liste de résultats jugés pertinents par rapport à une recherche.



Les trois grandes fonctions d'un moteur de recherche web

Crawling

Pour remplir pleinement leur rôle, les moteurs de recherche doivent, dans un premier temps, identifier les pages web et les ressources qu'ils souhaitent rendre accessibles.

Cette identification repose sur des logiciels, nommés **crawleur** ou robot d'indexation, qui explorent les réseaux de liens formés par les liens hypertextes contenus dans les pages web.

À partir d'une page initiale, généralement la page d'accueil qui a été soumise au référencement par le responsable du site, le crawleur enregistre dans une base de données tous les éléments constitutifs des pages qu'il a parcourues durant son exploration :

- URL ;
- titre ;
- contenu textuel ;
- images ;
- hyperliens ;
- documents associés ;
- etc.

Il faut noter qu'un crawleur ne visite pas uniquement les pages des seuls sites soumis au référencement. Il peut en effet explorer des liens externes, c'est-à-dire des liens qui pointent vers un nom de domaine différent de celui du site initial. De ce fait, même les sites web qui ne sont pas soumis au référencement peuvent finalement être pris en compte par le moteur de recherche.

Il existe un protocole d'exclusion qui permet à un responsable de site web d'interdire aux crawleurs de référencer tout ou partie des ressources. Ces instructions sont placées dans un fichier dédié à cet effet et nommé ***robots.txt****.*

Indexation

Sur la base des données recueillies par les crawleurs, le moteur de recherche va procéder à l'indexation de contenus, c'est-à-dire qu'il va extraire de ces contenus de nouvelles données afin de construire un index inversé. Cet index inversé est une structure de correspondance entre des mots et leur position dans les contenus. Et c'est cet index inversé qui sera utilisé comme référence pour les recherches dans les contenus.

Sur le Web, pour les ressources digitales de type image, vidéo ou encore podcast, ce sont généralement les textes et/ou les métadonnées associés qui sont pris en compte pour l'indexation. Si les transcriptions des contenus des vidéos ou des podcasts sont disponibles, ce sont elles qui sont indexées.

Appariement

Pour trouver une information, les moteurs de recherche permettent aux utilisateurs d'exprimer leur besoin au travers d'une interface dédiée. Bien souvent, il s'agit d'un champ de recherche dans lequel il est possible de taper une requête constituée de mots-clés accompagnées ou non d'opérateurs de recherche (pour plus de précision sur les opérateurs de recherche voir la section ci-après consacrée aux bonnes pratique de recherche à adopter).

Il existe d'autres types d'interfaces permettant d'effectuer par exemple des recherches exploratoires (navigation dans des arborescences de mots-clés, expressions, catégories, etc.), des recherches par similarité (utilisation d'un exemple de contenu recherché comme requête afin de trouver des contenus similaires), etc. Dans tous les cas, une requête est générée et utilisée pour interroger l'index inversé.

À partir de cette requête, le moteur de recherche va parcourir son index de manière à identifier les contenus qui répondent le mieux au besoin exprimé. Pour y parvenir, le moteur de recherche va opérer un certain nombre de calculs dont la vocation est, de manière simplifiée, de mesurer l'écart entre les contenus indexés et la requête. On parle d'indice de pertinence.

À titre d'exemple, voici quelques critères pris en compte pour calcul l'indice de pertinence d'un contenu par rapport à une requête :

Exemple de critères pour le calcul de l'indice de pertinence

Il existe bien d'autres types de critères de pertinence comme :

- la popularité d'une page, critère qui a fait le succès de Google à ses débuts face au moteur de recherche qui était alors le numéro un et qui aujourd'hui a disparu, Altavista ;
- les préférences de l'utilisateur ;
- etc.

En pratique, ces critères ne sont généralement pas dévoilés dans le détail par les moteur de recherche dans la mesure où ils constituent le cœur de leur fonctionnement et de leur potentiel avantage concurrentiel aux yeux des utilisateurs.

Grâce aux résultats de ces calculs, le moteur de recherche va pouvoir présenter à l'utilisateur les contenus qu'il juge être les plus pertinents par rapport à sa requête.

b. Caractéristiques importantes

Par-delà ces grands principes, il est important de mettre en lumière un certain nombre de caractéristiques inhérentes aux moteurs de recherche web qui peuvent avoir un impact sur leur utilisation dans le cadre d'activités de veille.

Un périmètre d'indexation limité

Il est très important de noter que toutes les ressources web existantes ne sont pas prises en compte par les moteurs de recherche, et ce pour différentes raisons :

- le contenu se trouve derrière un formulaire d'interrogation nécessitant la réalisation d'un certain nombre d'actions (remplir des champs, valider le formulaire, etc.) ;
- le contenu est protégé ou non accessible (c'est le cas pour la plupart des contenus du Web social) ;
- le propriétaire a spécifié dans le fichier « robot.txt » qu'il ne souhaitait pas qu'une ou plusieurs pages de son site soient visitées par un robot ;
- etc.

Une optimisation des contenus qui influence directement les résultats d'une recherche

Sur le Web, quels que soient les producteurs de contenus (entreprises, médias, institutions publiques, experts…), la recherche de visibilité à tout prix est devenue quasiment vitale pour des raisons de rentabilité, de notoriété ou bien encore de crédibilité.

C'est ainsi qu'au fil du temps sont apparues diverses techniques visant à améliorer le positionnement d'un contenu ou d'un site web dans les résultats des moteurs de recherche web. Connues sous le terme de SEO (*Search Engine Optimization*), ces techniques ont, par nature, une influence directe sur la liste des résultats obtenus à la suite d'une recherche sur le Web.

Il faut donc bien garder en tête que, dans certains cas, les premiers résultats proposés à la suite d'une recherche peuvent être plus pertinents par rapport au travail important fourni en matière de SEO que par rapport au contenu lui-même.

Cela veut donc dire qu'il est important de ne pas se contenter de la première page de résultats dans le cadre d'une activité de veille.

Des fonctionnalités d'assistance à la recherche très diverses

Pour rechercher une information, les moteurs de recherche web proposent une interface web permettant aux internautes d'effectuer une requête composée de mots-clés.

Bien que très largement répandu, ce mode de recherche d'information est loin d'être optimal et les problèmes potentiellement nombreux :

- Utilisation de trop peu de mots-clés pour décrire précisément son besoin.
- Fautes d'orthographe ou de frappe.
- Emploi de mots polysémiques (par exemple, souris : animal, souris d'ordinateur, verbe sourire conjugué, pièce de viande d'agneau, etc.).
- Utilisation d'acronymes ou un jargon particulier à un métier.
- Etc.

Pour réduire le risque d'obtenir des résultats non pertinents et adresser ces problèmes, certains moteurs de recherche ont mis en place un certain nombre de traitements linguistiques et proposent des fonctionnalités d'aide à la formulation des requêtes : correction automatique, suggestions, recherche phonétique, utilisation d'opérateurs spécifiques, formulaire de recherche avancée, etc.

D'autres proposent des fonctionnalités d'affinage des résultats : filtrages des résultats suivant différents critères (dates, types de contenus, sources…), suggestion de mots-clés ou de requêtes complémentaires, classement des résultats, etc.

Ces fonctionnalités ne sont pas toujours présentes et surtout peuvent être très différentes d'un moteur de recherche web à l'autre. Il est donc très important, avant d'utiliser un moteur de recherche, de connaître l'étendue des fonctionnalités d'assistance à la recherche proposées et surtout d'en comprendre le fonctionnement.

Des algorithmes de calcul d'indice de pertinence pouvant créer des bulles de filtres

La notion de pertinence d'une information est intimement liée au jugement même de la personne sur cette information. Ce jugement n'est pas lié à la requête effectuée pour trouver l'information mais bien à son besoin concret en matière d'information.

Or, comme nous l'avons vu, le calcul d'indice de pertinence s'appuie généralement sur la requête pour évaluer la pertinence d'un contenu. Dit autrement, si la personne n'a pas correctement exprimé son besoin au travers de sa requête, ou l'a fait de manière partielle, il est évident que le calcul de l'indice de pertinence n'aura que très peu de sens.

C'est la raison pour laquelle, un certain nombre de moteurs de recherche prennent en compte d'autres critères comme le profil de l'utilisateur avec notamment ses centres d'intérêt. Ces derniers peuvent être déclarés explicitement par l'utilisateur et/ou détectés automatiquement en fonction des recherches déjà effectuées et des résultats consultés.

Dans certains cas, ces critères liés au profil sont prépondérants dans le calcul de l'indice de pertinence. Si bien qu'il est très facile de créer une bulle de filtre autour de l'utilisateur et donc d'impacter fortement les résultats de ses recherches.

Il est donc très important de « dépersonnaliser » les moteurs de recherche que nous utilisons notamment lorsque nous faisons de la veille sur des sujets, des domaines et avec des points de vue qui peuvent être très différents à chaque fois. Ces aspects seront abordés par la suite dans la section de ce chapitre consacré aux bonnes pratiques.

3. Moteurs de recherche généralistes

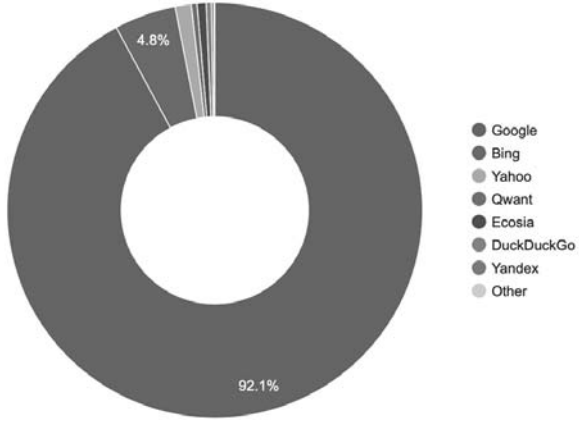
a. Panorama

Les moteurs de recherche dits généralistes sont des moteurs de recherche dont la vocation est d'indexer n'importe quels types de pages web sur n'importe quelle thématique.

Le moteur de recherche généraliste le plus connu et le plus utilisé à travers le monde est sans conteste Google. Les deux seuls pays où Google n'est pas le moteur de recherche numéro 1 des internautes sont la Russie et la Chine avec respectivement les moteurs Yandex et Baidu comme leaders.

En France, les moteurs de recherche généralistes les plus utilisés sont :

- Google : 92 % de part de marché fin 2022 ;
- Bing : 4,79 % ;
- Yahoo : 1,25 % ;
- Ecosia : 0,70 % ;
- Qwant : 0,45 % ;
- DuckDuckGo : 0,40 %.



Parts de marché des moteurs de recherche en France en octobre 2022
(source StatCounter – Infographie WebRankinfo)