

M3-moteurs-generalistes

Les moteurs généralistes

- Duopole [Google](#) [1] / [Bing](#) [2] avec une avancée commerciale et technologique considérable de la part du moteur appartenant à la société Alphabet.
- On peut noter quelques initiatives autour du respect de la vie privée comme [DuckDuckGo](#) [3] et [Qwant](#) [4], dont on peut également mettre en avant les efforts différenciés de visualisation des résultats sous forme de colonnes.
- D'autres moteurs se positionnent comme des acteurs du Web éco-responsables en finançant des projets écologiques ou solidaires comme [Lilo](#) [5] ou [Ecosia](#) [6]. S'ils peuvent constituer une alternative à Google pour des recherches simples, ils sont en revanche trop limités pour des recherches professionnelles.

Etat des lieux en 2020

Les parts de marché des principaux moteurs de recherche sont à peu près équivalentes quels que soient les pays, avec une prépondérance notable de Google, en dehors de la Russie et de la Chine, dans lesquelles les moteurs nationaux respectifs, [Yandex](#) [7] et [Baidu](#) [8] sont majoritairement utilisés.

L'hégémonie de Google (Stat France entre juin 2019 et juin 2020)

Source: [StatCounter Global Stats - Search Engine Market Share](#) [9]

Comment fonctionnent les moteurs ?

4 éléments fondamentaux

- Le « robot » ou « bot » ou encore « spider » ou « crawler » qui parcourent le Web afin d'identifier de nouvelles pages et les mises à jour de celles figurant déjà dans sa base de données ;
- Le « parser » ou analyseur qui extrait les contenus pour les stocker dans une immense base de données. Un module d'indexation crée et met constamment à jour plusieurs index. Une fois « parsé », le document est découpé en mots qui constituent l'index du moteur. Le module d'indexation repère la position des mots dans la page (notamment grâce aux balises de titre) et analyse également les liens présents sur la page ainsi que leurs points d'ancrage dans la page. Les index ainsi générés sont très volumineux et stockés dans d'immenses data centers ;
- Le formulaire par lequel l'internaute formule sa requête ;
- Les pages de réponse en provenance du serveur du moteur constitués de plusieurs types : les résultats « organiques », les annonces (référencement payant) et parfois un encadré baptisé « Knowledge Graph ».

La collecte par les robots est-elle exhaustive ? S'effectue-t-elle en temps réel ?

Pour information, GoogleBot, le robot de Google « crawle » environ 20 milliards de sites par jour.

Malgré ce chiffre énorme et devant la surabondance de contenus sur le Web, les robots ne peuvent pas visiter toutes les pages tous les jours et priorisent leurs visites en fonction de plusieurs critères :

- Fréquence de réactualisation du site,
- Indice de popularité,
- ...

Selon ce principe les pages provenant de sites d'actualité sont ainsi souvent intégrés quasiment en temps réel.

Les robots suivent l'architecture des sites :

- Page d'accueil,
- Suivi des liens internes et externes à cette page afin de « crawler » les contenus associés.

Attention donc à site très volumineux et/ou une arborescence très complexe qui ne permettent pas aux robots de prendre en compte toutes les pages d'un site. Par ailleurs, un webmaster peut bloquer l'accès à certaines parties d'un site ou à certaines pages grâce au fichier « robots.txt ».

L'indexation se fait elle sur tous les termes contenus dans une page ?

Les contenus multimedia posent actuellement encore des problèmes et leur indexation se fait essentiellement à partir du nom du fichier et des métadonnées associées.

Comment sont présentés les résultats ?

Même si le nombre de résultats peut atteindre plusieurs milliers voire millions de résultats, Google n'affiche que les 1000 premiers. Chaque réponse comprend la balise titre, l'adresse URL, des extraits du contenu comportant les mots de la requête, et le cas échéant, le « sitemap ».

Que contient une page de résultats ?

- Liens organiques ou liens naturels ou encore liens bleus => proviennent de l'index du moteur ;
- Liens sponsorisés : ils sont en tête et précédés de la mention « Annonce » Ce bon positionnement résulte de l'achat préalable de mots-clés parmi lesquels, pour les plus utilisés un système d'enchères permet à l'annonceur le plus offrant de figurer dans la position la plus élevée ;
- « Knowledge graph », encadré à droite affichant des données provenant de plusieurs sources (pas toujours indiquées) dont Wikipedia. Peut aider à trouver de nouveaux contenus grâce aux « recherches associées » ou à désambiguïser une recherche ;
- Les résultats locaux => concerne des résultats liés à la position géographique de l'internaute ;
- « Rich snippets » ou extraits enrichis => informations supplémentaires sur le contenu de la page. Ils accompagnent généralement des sites commerciaux et/ou grand public. Ces extraits enrichis proviennent des microdonnées, éléments invisibles aux navigateurs et encapsulés par le créateur de la page dans le code source d'une page HTML.
- « Featured snippets » : Pour certaines requêtes, Google affiche en tête des résultats des « featured snippets », ou extraits optimisés : il s'agit d'extraits provenant de

Wikipedia ou de tout autre site considéré comme le plus pertinent. Cet extrait est censé répondre à la question de l'internaute, sans même qu'il ait à quitter Google pour se rendre sur la page d'où provient l'extrait. Cette fonctionnalité correspond à des recherches assez basiques, liées à la vie quotidienne.

- « PAA » ou « People Also Ask » (« Autres questions posées »). Cet encadré, bien souvent situé sous le « featured snippet », reprend des questions similaires à la requête de base, posée par d'autres internautes, et propose un « featured snippet » adapté à chaque question.

Les featured snippets et le knowledge graph ne sont pas calculés par les mêmes algorithmes mais font partie de la « position zéro » de Google, c'est-à-dire avant les liens organiques.

La stratégie de Google est donc de faire rester l'internaute le plus longtemps possible sur la page de résultat sans qu'il ait besoin de cliquer sur un lien organique.

Les algorithmes de classement

Il existerait plus de 200 algorithmes utilisés dans le cadre du SEO (Search Engine Optimization) soit le référencement naturel.

Les principaux critères de tri utilisés par les moteurs :

- Les contenus => le nombre d'occurrences trouvées dans une page par rapport à la requête saisie par l'internaute, pondéré en fonction du nombre total de mots dans la page (densité des mots). La position du ou des termes de la requête dans certaines zones de la page a également une importance en privilégiant la balise <title>, l'url et la balise h1. La proximité des mots de la requête dans une même page est aussi un critère de pertinence de même qu'une mise en exergue particulière (balise par exemple). Enfin, toutes les pages web n'étant pas crawlées de la même manière, les pages dotées d'url courtes mais comportant des mots-clés recherchés par les internautes, les pages provenant d'un site disposant d'un « sitemap » ou dont l'architecture ne repose pas sur une arborescence trop profonde et complexe seront mieux référencées que d'autres.
- La « popularité » et la mesure d'audience : le « Pagerank » est une subtile et opaque promotion mêlée de la notion de popularité (nombre de liens pointant vers la page auditée, dits aussi « backlinks ») et de notoriété, puisque la provenance de ces liens compte également dans la note finale, Google partant du principe qu'un certain nombre de sites sont incontournables dans leur domaine. Il est calculé selon un barème allant de 1 à 10. Plus les sites pointant vers la page auditée ont eux-même un « Pagerank » élevé, plus la page verra son propre « Pagerank » augmenter. À côté du « Pagerank » présent depuis les débuts de Google en 1998, Google a développé aussi le principe du « trustrank » (basé sur la fiabilité des liens, destinée à éviter les tricheries) et de l'« authorank » (lié à l'expertise et à la réputation de l'auteur d'un site).

Autres critères liés à l'architecture des sites :

- L'ancienneté du site est privilégiée.
- Site « responsive » et actualisé, formats « AMP » et protocole HTTPS
- « Mobile first » => déploiement par Google depuis fin 2017 d'un nouvel index qui concerne les versions mobiles des sites. A terme, c'est la version mobile d'un site qui servira de référence et non la version ordinateur. Ce bouleversement bénéficiera d'une transition assez longue.
- La présence sur les réseaux sociaux => une page partagée sur les réseaux sociaux est en général bien classée dans les résultats de Google.
- Le « rankbrain » => algorithme élaboré par Google en fin 2015, basé sur des techniques d'intelligence artificielle. L'objectif est d'aider le moteur à mieux comprendre la requête et l'intention de l'internaute. Le système modélise des requêtes sous forme de vecteurs mathématiques. Des fonctionnalités d'autoapprentissage augmentent l'intelligence du système et intègrent le sens des mots ou phrases en toutes langues. « Rankbrain » serait devenu le 3ème critère de pertinence majeur pour Google après les contenus et les liens. L'algorithme « Rankbrain » illustre une évolution de l'indexation chez Google : des chaînes de caractères (« strings ») aux concepts (« things »).
- Pour être bien classée, une page doit comporter non plus seulement une succession de mots-clés, mais des contenus variés et de qualité, organisés de façon structurée et non pas simplement juxtaposés.

Facteurs liés à l'internaute

- Personnalisation des résultats en fonction des préférences ou des requêtes antérieures de l'internaute,
- Localisation de l'internaute et surtout du mobinaute.
- Les microdonnées
- Éléments encapsulés dans le code HTML 5 et destinés à apporter du contenu sémantique, mais invisibles pour les internautes. Le standard de microdonnées schema.org a été élaboré en 2011 suite à un accord avec Google et d'autres moteurs. Ce schéma de microdonnées a pour objectif d'apporter aux moteurs une meilleure « compréhension » des contenus grâce aux différentes balises incorporées, et d'offrir une meilleure visualisation des réponses pour les internautes. Ce schéma est critiqué pour avoir une vocation « commerciale » et d'autres extensions sont à l'étude pour le secteur non marchand.

Ce qu'il faut retenir

Les moteurs généralistes et notamment Google, sont optimisés pour guider les internautes dans leurs recherches au quotidien, ce qui biaise les résultats, lesquels sont largement basés sur la notion de popularité. De ce fait, ils ne mettent pas en avant la nouveauté que l'on recherche lorsque l'on est dans une position de veille.

Si leurs algorithmes évoluent régulièrement (à titre d'exemple il y en a eu plus de 1600 en 2016), ils sont assez opaques.

- Attention à la navigation en étant connecté à son compte Google car les recherches précédentes sont mémorisées.
 - Quelques liens utiles à ce sujet :
 - Désactivation de l'enregistrement des recherches par Google quand on est pas connecté à son compte ⁽¹⁰⁾
 - Vérifier la configuration des données enregistrées lorsque l'on est connecté à un compte Google : ⁽¹¹⁾Google ⁽¹¹⁾History ⁽¹¹⁾
 - En savoir + sur le ⁽¹²⁾Pagerank ⁽¹²⁾
 - Le PageRank : quoi de neuf en 2020 ? ⁽¹³⁾
 - Mes pistes pour un bon référencement Google en 2020

Se servir de la syntaxe de Google pour des requêtes avancées

Les 5 opérateurs indispensables chez Google

Disponibles directement à partir du formulaire classique de saisie de Google ou par la grille sur le formulaire « recherche avancée », qui offre parfois moins de souplesse.

- Les guillemets pour rechercher des expressions exactes ou imposer l'orthographe d'un mot ;
- L'opérateur ET (ou AND ou encore +) est implicite chez Google et de nombreux moteurs.
- L'opérateur OR pour rechercher un mot ou un autre (inclusif c'est-à-dire soit un terme, soit l'autre, soit les 2, Google ne permettant pas la version exclusive).

Deux syntaxes possibles pour les mêmes résultats :

- marché automobile France OR Allemagne
- marché automobile (France OR Allemagne)

Opérateur intéressant dans la recherche de personnes car permettant la combinaison « prénom nom » et « nom prénom ».

Attention, chez Google, le OR (contrairement au ET) désactive la recherche « floue » et la prise en compte des variations d'un terme (singulier/pluriel, masculin/féminin ...)

=> nourriture pour chiens ou chats => nourriture chien OR chiens OR chat OR chats

- Le signe - (équivalent à SAUF pour exclure un mot)

Permet de réduire l'ambiguïté d'un terme, par exemple dans le cas d'une homonymie. Comme pour le OR, le signe - désactive chez Google la prise en compte des variations d'un terme. A utiliser avec parcimonie.

Exemple : jaguar -automobile -véhicules

- Site : pour cibler un domaine ou un site spécifique

Très utile pour délimiter un périmètre à une recherche lié au nom de domaine ou à l'extension de domaine (ex. : .edu, .gouv.fr, .org).

=> Google peut ainsi quasiment jouer le rôle d'un moteur de recherche interne, à l'intérieur des pages d'un site ... du moins les pages préalablement « crawlées » par le moteur.

Astuce : il vaut mieux saisir le nom de domaine sans les www, cela permet de prendre en compte davantage de pages du site concerné.

Cette syntaxe possède son contraire : -site :.com afin d'éliminer des résultats les pages avec extension .com.

Dernière utilisation : servir à faire apparaître le nombre de pages « crawlées » par Google à partir d'un site donné en saisissant la syntaxe site :nom-de-domaine-concerné, sans ajouter de mots-clés. Exemple : site :urfist.chartes.psl.eu

- Filetype : pour filtrer sa recherche par format

Cet opérateur recherche les pages créées dans un format de fichier particulier. Le format PDF est particulièrement utile pour concentrer sa recherche sur des documents à valeur ajoutée : articles, études, etc. Ces types de documents longs sont en effet assez mal classés par Google.

Le format PPT ou PPTX est intéressant pour trouver des présentations ou des cours, et de ce fait identifier des experts. Le format XLS ou XLSX peut aider à retrouver des tableaux de données. A noter que dans la recherche simple, il est possible de choisir des formats de fichier qui ne sont pas proposés par la liste déroulante du formulaire de recherche avancée.

En revanche, l'opérateur ne fonctionne plus avec les formats images (JPG, PNG, GIF, etc). Le filtrage concernant ces formats d'images peut s'effectuer à partir d'un menu déroulant dans le formulaire « recherche avancée d'images », choisir « type de fichiers ».

Liste des opérateurs de syntaxe avancée de Google : <https://support.google.com/websearch/answer/2466433> [14]

Utiliser les moteurs académiques

Ceux-ci sont étudiés pour indexer des ressources académiques dont certaines font partie du Web invisible, appelé aussi Web profond ou "Deep Web", parce-qu'il n'est pas pris en compte par les moteurs généralistes.

URL source: https://urfst.chartes.psl.eu/m3-moteurs-generalistes?f%5B0%5D=field_sujet_principaux%3A947&f%5B1%5D=field_domaines_disciplines%3A1026&f%5B2%5D=field_sujet_secondaire%3A2686&f%5B3%5D=field_sujet_principaux%3A950&f%5B4%5D=field_domaines_dis

Liens

[1] <https://www.google.fr/> [2] <https://www.bing.com/> [3] <https://duckduckgo.com/> [4] <https://www.qwant.com/?l=fr> [5] <https://www.lilo.org/> [6] <https://www.ecosia.org/> [7] <https://yandex.com/> [8] <https://www.baidu.com/> [9] <https://gs.statcounter.com/search-engine-market-share/all/france> [10] <https://www.google.fr/history/optout?hl=fr&fg=1> [11] https://myactivity.google.com/?restrict=ytw&hl=fr&utm_source=udc&utm_medium=r&utm_campaign= [12] <https://www.webrankinfo.com/dossiers/pagerank/formule> [13] <https://www.reacteur.com/2020/05/le-pagerank-quoi-de-neuf-en-2020.html> [14] <https://support.google.com/websearch/answer/2466433>