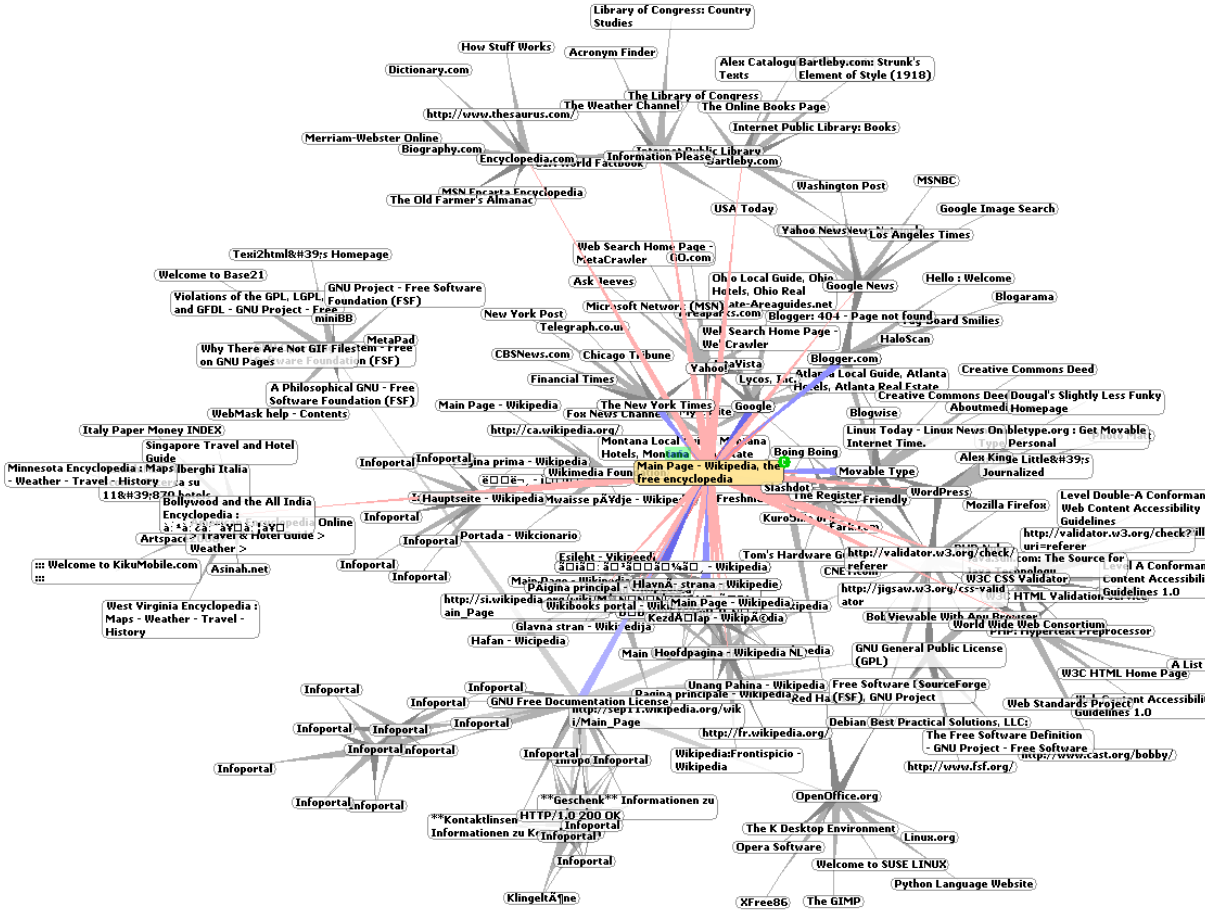
Les moteurs de recherche

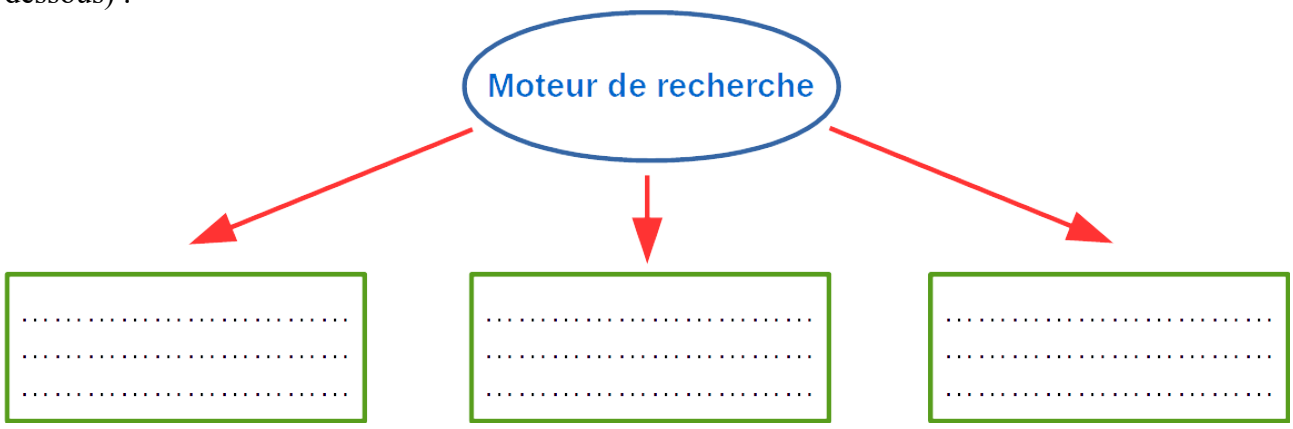
(sources : wikipedia.org, infowebmaster.fr, neocamino.com)

Le WEB serait difficilement utilisable sans les moteurs de recherche. Il est en effet constitué de milliards de documents liés entre eux par des hyperliens. La première idée a été de mettre en place des annuaires mais elle a été très vite abandonnée face à la très forte production de contenu.



Aperçu du web autour du site Wikipedia

Le but des moteurs de recherche est de fournir aux utilisateurs les bonnes ressources, classées par ordre de pertinence. Pour cela ils réalisent trois tâches successives (à remettre dans l'ordre ci-dessous) :



(les pages explorées sont sauvegardées et classées, des robots explorent le web, les mots clés sont recherchés dans les pages sauvegardées)

Les robots

Les moteurs de recherche utilisent des robots, appelés spiders. Ces araignées explorent la toile, le WEB, de liens en liens. On dit qu'ils rampent. Ce processus est appelé le crawl. Les robots, spiders, sont aussi dénommés crawlers.

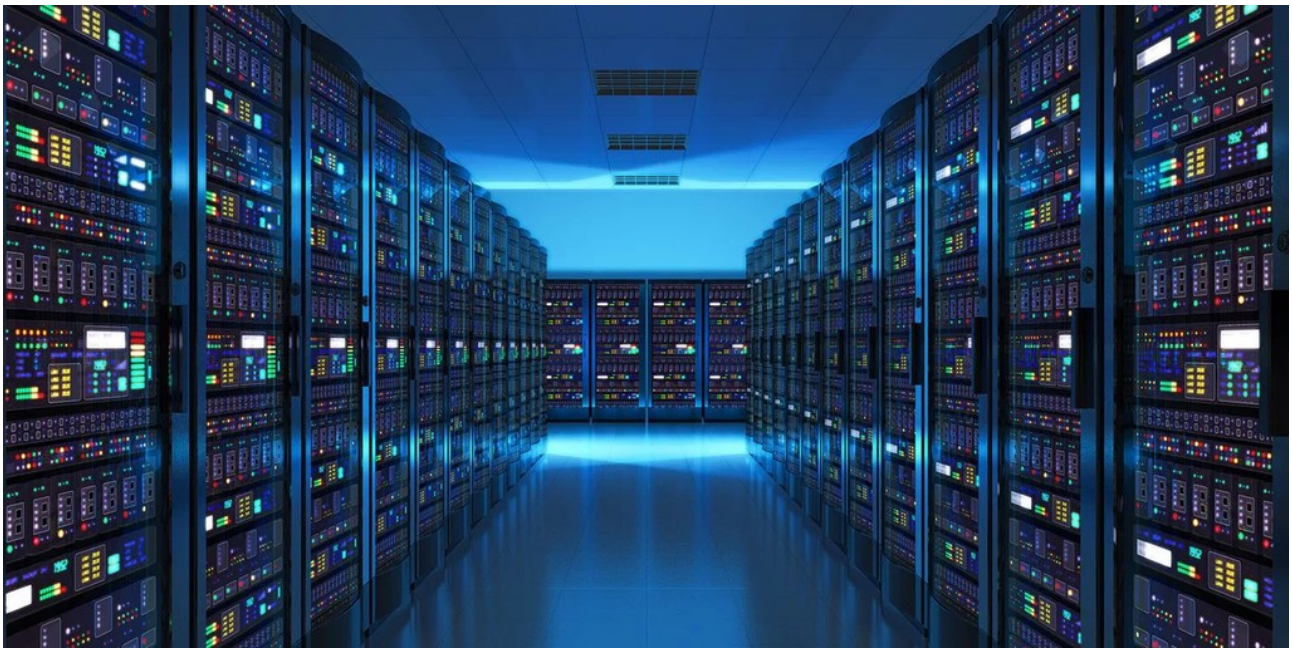
Les principales fonctions d'un robot est d'identifier les hyperliens, d'en extraire les adresses et de se déplacer de page en page. Une page WEB étant codées en HTML, un type de balise particulier doit être repéré. Son attribut donne l'adresse de l'hyperlien. En analysant le code source d'une page (ctrl-U dans un navigateur), identifier et compléter l'exemple de balise concernée.

<..... .....="http://home.cern/fr" >

Balise Attribut

L'indexation

Les pages explorées sont téléchargées et enregistrés dans d'énormes data-centers.



Un des data-centers de Google

Le contenu des pages est analysé en termes de mots clés. Les pages sont ensuite classées selon leur pertinence, leur degré de confiance, ... A chaque mot clé va alors correspondre une liste d'adresses de pages WEB. C'est un index.

Les index ne listent pas tous les mots. Les stops words (mots vides) désignent des petits mots très souvent utilisés comme "le", "la", "du", "à", …, mais non significatifs. Les mots significatifs sont associés à un poids dépendant de leur occurrence dans la page, obtenu grâce à la formule du TF-IDF (Term Frequency-Inverse Document Frequency).

Toutes les pages ne seront pas sauvegardées. Certaines pages provenant de sites illégaux ou de mauvaise réputation sont blacklistées.

La recherche

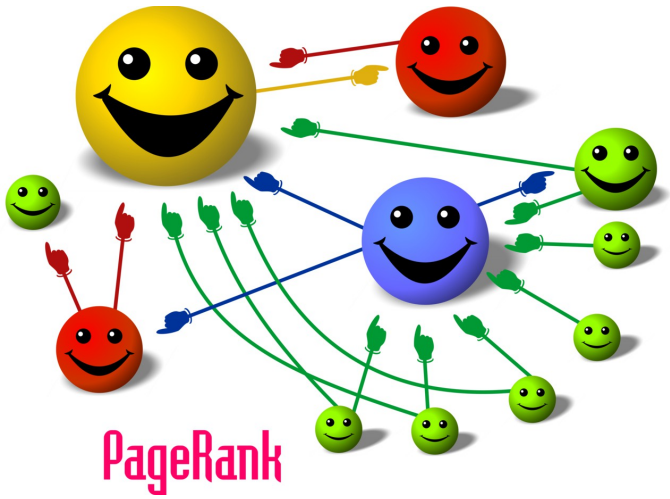
Lorsqu'un internaute effectue une recherche, il y a souvent des millions de pages qui possèdent les mots recherchés avec une bonne occurrence. Pour cette raison, les moteurs doivent classer les résultats. Deux groupes de critères influent le classement des résultats (popularité, pertinence, à compléter) :

……	……...
le mot-clé est-il présent dans le titre ? le mot-clé est-il présent dans l'URL ? le mot-clé est-il présent dans le contenu ? y'a t-il des synonymes du mot recherché dans le contenu ? … etc ...	est-ce que la page reçoit beaucoup de liens ? ces liens proviennent-ils de pages elles-mêmes populaires ? les pages faisant des liens ont-elles la même thématique ? les sites qui font des liens vers cette page sont-ils dans la même langue ? sont-ils des sites de confiance ? … etc ...

En plus de ces deux principaux groupes, des critères alternatifs ont fait leur apparition. Par exemple, le moteur de recherche Google base maintenant ses résultats selon la localité du visiteur et l'historique des précédentes recherches effectuées par l'internaute.

Pour améliorer encore les performances d'un moteur, il existe de nombreuses techniques. La plus connue est celle du ***PageRank*** de Google qui calcule un indice de notoriété de pages.

Enfin, les moteurs de recherche préparent à l'avance les résultats des requêtes les plus populaires : "Facebook", "Youtube", "Vidéo", "TV", "Jeux", ... Ainsi, ils donnent directement les résultats sans nécessairement avoir à chercher dans l'index. Un analyse des requêtes les plus populaires est donnée par Google à l'adresse ci-dessous.



https://trends.google.fr

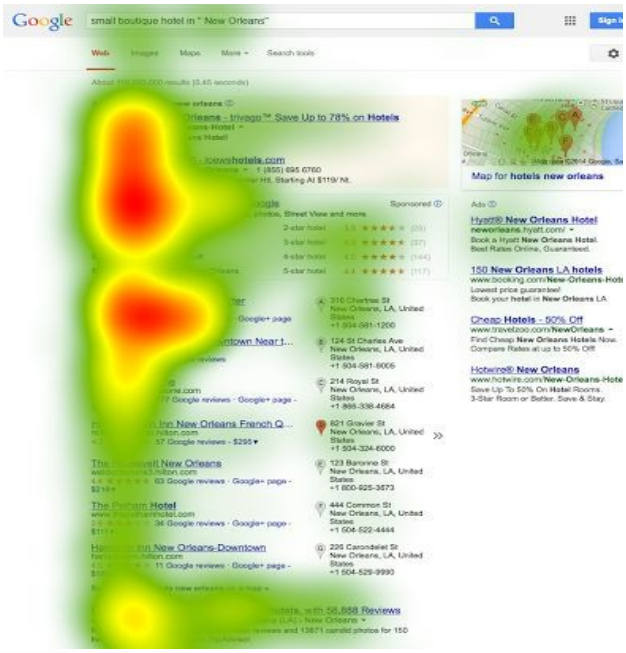
Que peut-on remarquer au sujet de l'occurrence depuis 2004 des mots ci-dessous ? En donner une interprétation.

Facebook : .....
.....
.....
Jouet :.....
.....
.....
Football :.....
.....
.....

Business is business !!

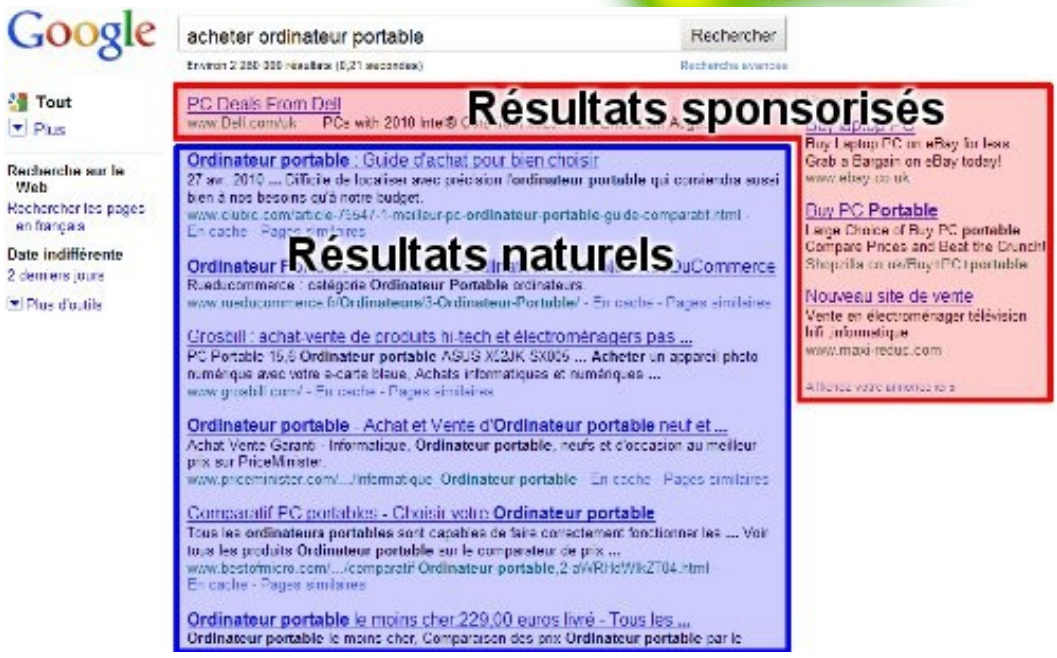
62 % des internautes ne dépassent pas la première page des résultats d'une recherche sur internet. D'où l'importance d'y figurer !

L'image de droite montre où se porte le regard de l'internaute sur une page Google. Vous aurez :

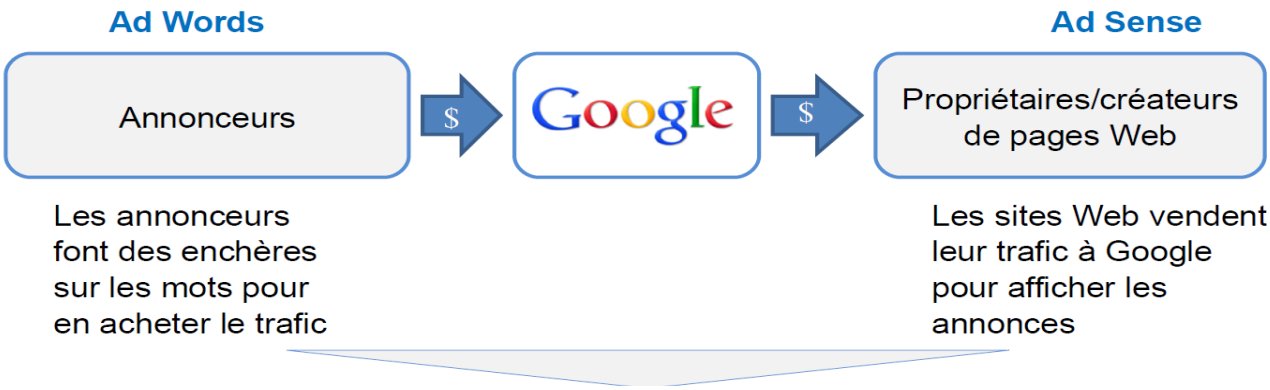


100% de visibilité si vous êtes en 1er, 2e ou 3e position
85% de visibilité pour la 4e position
60% de visibilité pour la 5e position
50% de visibilité pour la 6e et 7e position
30% de visibilité pour la 8e et 9e position
et 20% de visibilité si vous êtes en 10e position

Ainsi, les moteurs de recherche vont proposer des liens sponsorisés avant les résultats dits naturels de la requête de l'internaute. De plus, des publicités seront aussi présentes sur les sites clients.



Les mots clés entrés dans les moteurs de recherche par les millions d'internautes constituent un trafic de données très important. Une pratique est de vendre aux enchères les mots clés pour en acheter le trafic sur une certaine durée.



Petite discussion

Une araignée se déplace-t-elle réellement sur la toile ? Expliquer.

.....
.....
.....
.....

La méthode est rapide. Pourquoi ?

.....
.....
.....
.....

Il n'est normalement pas possible pour les robots de trouver une page orpheline, qui ne reçoit aucun lien. Pourquoi ?

.....
.....
.....
.....

Les robots ne peuvent tout indexer et mettre à jour rapidement. Selon le moteur de recherche et la méthodologie qui lui est associé, une page peut être revisitée quelques heures après sa publication ou plusieurs mois après. Pourquoi ? Quelles en sont les conséquences ?

.....
.....
.....
.....