

Recherche d'information sur Internet 1

Trouver l'adresse d'une ressource Internet

Deux grandes familles d'outils de recherche sur Internet : répertoire et moteur de recherche

> différence entre répertoire et moteur de recherche

- Les répertoires, appelés également annuaires ou guides web, sont des listes d'adresses de sites classés par grandes catégories.
- Les moteurs recherchent à la demande, grâce à leurs robots, les mots contenus dans les pages des sites web.

Les répertoires

Les répertoires sont des listes de sites classés manuellement par catégories thématiques et souvent commentées par des rédacteurs.

- Certains annuaires sont généralistes, comme l'Open Directory. Ces annuaires tendent à avoir la couverture la plus étendue.

<http://www.dmoz.org>

- Certains sont sélectifs, comme les signets de la BNF ou de la BPI. Ils sont fondés sur l'évaluation du contenu des sites par des professionnels pour retrouver facilement des informations fiables.

<http://signets.bnf.fr/>

<http://www.bpi.fr/signets.php>

- Certains sont spécialisés et traitent d'un seul domaine, comme l'annuaire le Point du FLE.

<http://www.lepointdufle.net/>

> Interrogation des répertoires

Deux modes d'interrogation sont possibles :

- soit parcourir l'arborescence des catégories thématiques depuis la page d'accueil.
- soit saisir un mot-clé dans le formulaire de requête

> Avantages et inconvénients des annuaires

Avantages :

- La navigation est très simple et constitue un guidage très efficace.
- Les sites référencés dans l'annuaire sont le fruit d'une sélection, les sites trouvés sont donc de bonne qualité et bien centrés sur le thème cherché.
- Les annuaires donnent accès à une liste réduite de sites, souvent les plus connus, ce qui vous évite de vous noyer dans une masse d'informations

Inconvénients :

- L'annuaire étant construit manuellement, cela entraîne que le nombre de sites référencés dans l'annuaire est réduit et ne suit pas la croissance du Web et la mise à jour de l'annuaire n'est pas toujours très bonne (nouveaux sites, sites disparus, etc.).

Les moteurs de recherche

> Principe

La plupart des moteurs sont généralistes et multilingues car ils cherchent des ressources dans différentes langues.

Il existe plusieurs moteurs reconnus sur Internet :

Google : <http://www.google.fr/>

Yahoo! : <http://fr.search.yahoo.com/>

Live Search : <http://www.live.com/>

> Fonctionnement des moteurs de recherche

Un **robot** parcourt automatiquement et à intervalle régulier les pages du Web **de liens en liens** pour enregistrer le contenu et l'adresse de chaque document et l'indexer dans une base de données.

L'efficacité des réponses du moteur dépend des algorithmes d'indexation des pages et de sa capacité à trier et à afficher les résultats.

- Les moteurs ne couvrent ni la totalité ni les mêmes parties du Web.
- Les moteurs délivrent donc des résultats différents et ils ne classent pas de la même façon.
- Les moteurs sont en concurrence économique et utilisent chacun leur technologie avec leurs propres algorithmes de calcul.

> Interrogation d'un moteur

L'utilisateur saisit un ou plusieurs mots clés dans la zone de requête. La recherche se fait alors sur le texte de toutes les pages que le robot a enregistrées.

Le moteur affiche la liste des pages contenant le ou les mots clés.

> Présentation des résultats

La liste des résultats est présentée sous la forme de pavés.

- titre du document
- un extrait du texte dans lequel apparaissent en gras le ou les mots-clés tapés
- adresse du document
- la taille du document
- des liens commerciaux

The screenshot shows a Google search interface with the query 'paris en images'. The search results are displayed as a list of links (pavés) on the left and a sidebar with commercial links on the right. The main results include:

- Paris en images - banque images Paris collection photo Paris**: La Parisienne de Photographie vous propose avec Paris en images une découverte interactive des collections de photos de Paris à travers une banque d'images ...
- Galerie des collections - collection photo noire et blanc photos ...**: La Parisienne de Photographie vous propose une découverte interactive de la collection de photos de Paris à travers le site Paris en Images.
- Paris en images**: Voici quelques images de Paris! Nous espérons que vous prendrez du plaisir à les regarder comme j'ai eu du plaisir à les prendre et à me promener dans cette ...
- Action Speculand à Paris en images - Attac France**: Action Speculand à Paris en images, article publié le 5/07/2008 auteur-e(s) : Wilfried Maurin. Photos prises au cours de la première action Speculand à Paris ...

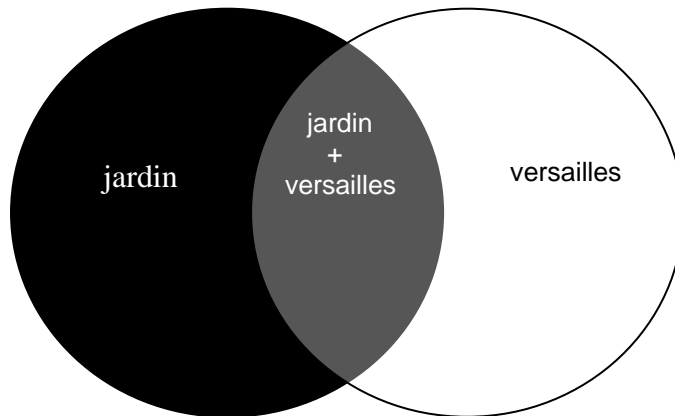
The sidebar on the right contains commercial links such as 'Images Paris', 'Photos Paris', and 'Images De Paris'.

> Formuler une requête simple dans un moteur

Dans de nombreux cas, il suffit de saisir un ou plusieurs mots-clés dans la zone de requête. Les termes choisis doivent être séparés par un espace. Les mots clés doivent être choisis dans la langue des documents que l'on recherche.

L'opérateur ET

Pour rechercher des documents contenant les deux termes à la fois, Google transforme implicitement l'espace entre deux mots par l'opérateur booléen ET sans qu'il soit nécessaire de le saisir.



Choix des mots clés

Google recherche (la plupart du temps) les mots tels qu'ils ont été saisis, par conséquent :

- les singuliers comme les pluriels constituent autant de critères déterminants
- les mots sont à saisir dans la langue du document cherché

Casse des mots et accents

De façon générale, les moteurs sont insensibles à la casse des caractères. La situation est plus contrastée en ce qui concerne les accents : parfois, les moteurs ne retournent pour un mot clé accentué que les mots contenant l'accent, mais pour une requête non accentuée, ils retournent les mots avec ou sans accent.

Types des mots

Les moteurs de recherche indexent indifféremment la quasi-totalité des mots d'une page Web : noms communs, noms propres, verbes, adjectifs, adverbes. Généralement, seuls les articles et pronoms sont ignorés.

Google ignore les mots vides que sont, dans toutes les langues :

- les articles (le / the)
- les prépositions (de / of)
- les mots de liaison (avec / with)

En effet, ces mots à faible poids sémantique sont rarement déterminants et produiraient du "bruit".

Inclusion d'un mot

Pour forcer l'inclusion d'un mot, il suffit de le faire précéder par un + (collé au mot) ou du mot SAUF

Exclusion d'un mot

A l'inverse, pour exclure un mot des résultats, il suffit de le faire précéder d'un -

Choix des mots clés

Combien de mots clés ?

On procédera par étapes pour affiner éventuellement sa recherche à l'aide de plusieurs mots clés. On commence par 1 ou 2 mots clés et on affine si besoin en ajoutant des mots clés supplémentaires.

Synonymes

Il est important d'explorer la terminologie du domaine de recherche, pour récupérer les synonymes (très rares sont les moteurs travaillant sur les concepts). De façon générale, les premiers documents intéressants récupérés permettent de valider, compléter ou revoir ses mots clés.

Identifier des synonymes :

- Utiliser un dictionnaire de synonymes : <http://elsap1.unicaen.fr/cgi-bin/cherches.cgi>
- S'inspirer des mots clés suggérés par des moteurs à contextualisation comme Exalead : <http://www.exalead.fr/search>

Outils de suggestion de mots clés : Google propose sur son moteur de recherche une nouvelle fonction d'aide à la formulation de requêtes. Baptisée *Google Suggestions*, son principe est d'éviter à l'utilisateur de taper un mot-clé en entier. Le moteur va afficher, en cours de frappe, dans un menu déroulant placé sous le champ de recherche des suggestions qui correspondent au mot-clé en train d'être tapé. En outre, les propositions d'orthographe alternatif (« Essayez avec cette orthographe ») sont désormais affichées en cours de frappe dans le menu déroulant de Google Suggestions, et non plus lors de l'affichage des résultats.

Classement des résultats

Les résultats sont classés par pertinence, calculée automatiquement selon :

- la structuration de la page
- mesure de notoriété de la page

Structuration des mots dans la page

- Il analyse comment les deux mots tapés se présentent dans les pages Web de sa base :
 - sont-ils l'un à côté de l'autre ?
 - sont-ils proches l'un de l'autre ?
 - combien de fois apparaissent-ils ?
 - où se trouvent-ils dans la page (dans le titre, le corps du texte ou dans les liens de la page ?)
- Il attribue un score en fonction de ces critères à chacune des pages trouvées.

Notoriété

Google combine ce score avec un indice de notoriété : le PageRank pour classer les résultats.

La notoriété d'une page est calculée en fonction du nombre de liens qui dans d'autres pages, pointent vers cette page. La notoriété des pages où se trouvent ces liens est également prise en considération.