Chapitre 4 : Moteurs de Recherche

Internet propose une quantité considérable de contenu numérique. • Trouver une URL répondant à nos besoins implique généralement l'utilisation d'outils de recherche d'informations.

4.1. Définition de la recherche d'information: La recherche d'informations, une discipline de l'informatique, se concentre sur la conception de systèmes visant à faciliter la récupération d'une information précise répondant aux besoins de l'utilisateur parmi un ensemble de documents. En anglais IR (information retrieval)

4. 2. Type d'outils de recherche: pour trouver de l'information sur Internet, on peut utiliser plusieurs outils. Nous citons les suivants

- **Les moteurs de recherche:** C'est une plateforme en ligne qui permet de localiser diverses ressources (pages web, articles de forums, images, vidéos, fichiers, etc.) associées à n'importe quel mot-clé. Certains sites web fournissent un moteur de recherche comme fonctionnalité principale ; dans ce cas, le site lui-même est appelé moteur de recherche (par exemple, http://scholar.google.fr/). Les moteurs de recherche ne se limitent pas à Internet : certains sont des logiciels installés sur un ordinateur personnel (PC). Ces moteurs, appelés moteurs de bureau, fusionnent la recherche parmi les fichiers stockés sur l'ordinateur et la recherche sur le Web. On peut citer parmi eux Exalead Desktop, Google Desktop, Copernic Desktop Search, etc.

- **Les annuaires:** Ce sont des outils de recherche créés par des individus qualifiés qui répertorient et classifient des sites web considérés comme pertinents, plutôt que des robots d'indexation. Parmi ces outils, on peut mentionner Voilà et Yahoo!.

- **Les méta-moteurs de recherche:** Il existe également des métamoteurs, des sites web où une recherche est lancée simultanément sur plusieurs moteurs de recherche, puis les résultats sont combinés pour être présentés à l'utilisateur. Parmi ces métamoteurs, on peut citer MetaCrawler (http://metacrawler.com).

- **La toile invisible:** On entend souvent parler de la toile visible et de la toile invisible sur le web, mais que signifient ces termes ? En réalité, une partie importante des ressources disponibles sur Internet n'est pas indexée par les robots des moteurs de recherche, soit parce que les pages sont protégées par des accès restreints, soit parce qu'elles font partie de bases de données. La Toile visible peut être comparée à celle explorée par les moteurs de recherche tels que Google. La Toile invisible représente donc la partie complémentaire non accessible aux moteurs de recherche traditionnels, comprenant les pages appartenant à des bases de données (comme Medline, Inist, Cismef...) ainsi que toutes les pages à accès restreint (telles que les pages réservées aux professionnels, aux abonnés d'un journal, aux membres d'une association, etc.).

4. 3. Moteur de recherche: Un moteur de recherche est un logiciel, également connu sous le nom de robot ou spider, qui explore automatiquement Internet à intervalles réguliers. En suivant les liens présents sur des millions de pages web, le moteur de

recherche identifie constamment de nouvelles adresses et indexe leur contenu dans d'énormes bases de données. Nous interrogeons ensuite ces bases de données en utilisant des mots-clés.

Le fonctionnement d'un moteur de recherche, comme tout outil de recherche, peut être décomposé en trois processus principaux.

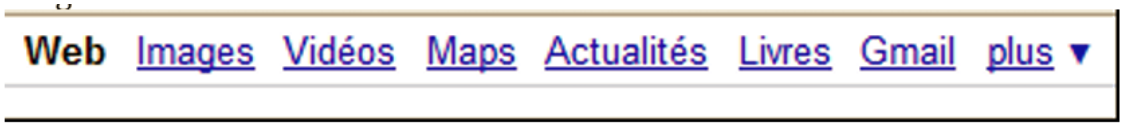
a. ***Robot (Spider):*** Il s'agit d'un robot logiciel qui explore de manière autonome le "Web". Son efficacité est cruciale pour le fonctionnement du moteur de recherche. Il repère les liens des pages pour ensuite les visiter à son tour, ce qui lui permet de parcourir rapidement l'ensemble du site ainsi que ses liens associés. Il examine périodiquement des millions de pages et construit ainsi une base de données des pages déjà visitées.
b. ***Système d'indexation:*** Il analyse les informations collectées, crée un index des mots rencontrés (et des pages correspondantes), puis stocke l'ensemble dans une base de données. Il convertit également certains fichiers qui ne peuvent pas être indexés en raison de leur format. De plus, il utilise des outils d'extraction pour ne récupérer que l'essence des documents. Parmi ces outils, on peut citer Fulcrum, Infoseek, Intelliserv et Livelink.
c. ***Searcher***: Le searcher est l'interface utilisateur principale. Grâce à son interface graphique, l'utilisateur peut poser des questions, sélectionner parmi les options disponibles et lancer une recherche. Un script fait ensuite appel au système d'indexation pour exécuter la requête sur la base de données. Les résultats sont généralement affichés sous forme de page web, intégrant les réponses sous forme de liste.

3. 4. Moteur de recherche Google: Le moteur de recherche Google, fondé en 1998 par Larry Page et Sergey Brin, est le moteur de recherche le plus populaire sur le Web à l'échelle mondiale. Le nom "Google" dérive du terme "Gogol", qui fait référence au nombre 10^100. Ce chiffre a été sélectionné pour symboliser la capacité de Google à traiter d'énormes volumes de données.

- **Classement des résultats (page ranking)**: Google repose principalement sur l'exploitation de la technologie Page Rank. Google analyse de façon très précise les pages Web et fichiers disponibles en igne: 1) il classe toutes ces pages dans son index de plusieurs milliards de pages. 2) le moteur de recherche étudie la popularité de la page, pour lui donner une note, le Page Rank, qui est le résultat d'un algorithme. le Page Rank calcule le nombre de fois qu'un site S est cité par d'autres sites A, B, C, D… mais il prend aussi en compte la popularité des sites A, B, C, D: être cité par des sites qui sont eux mêmes populaires augmente la popularité de S.

3. 5. Recherche sur Google: pour effectuer une recherche sur Google, on suit les étapes ci-dessous.

- Accéder à l'interface de recherche: l'accès se fait à travers un navigateur Web (adresse google.com ou des noms de domaines similaires. L'accès peut se faire
- Spécifier la nature des ressources recherchées (sites Web, images…)



- Spécifier le(s) mot(s)-clé de la recherche (on peut également rechercher par image….) avec des combinaisons et options (dates, langues…)
- Parcourir les résultats et accéder aux ressources.

- **Choix des mots-clés**

Pour trouver des informations sur un sujet, les moteurs de recherche nécessitent des indices, des mots-clés. Le choix des mots-clés est crucial pour la réussite et l'efficacité de votre recherche. Voici quelques règles de base à garder à l'esprit :

- Évitez les termes généraux et privilégiez les termes spécifiques (par exemple : "isolation" plutôt que "rénovation").
- Tenez compte de la manière dont la page recherchée a été rédigée (utilisez les mots les plus susceptibles d'apparaître dans l'article recherché).
- Utilisez des guillemets pour demander au moteur de recherche de ne répertorier que les documents contenant les mots dans l'ordre spécifié.
- Privilégiez les noms plutôt que les verbes, adjectifs, pronoms et adverbes, car ces derniers sont souvent ignorés par les moteurs de recherche.
- La simplicité est essentielle, un seul mot bien choisi peut parfois suffire, car plus il y a de mots, plus la recherche est restreinte.
- L'ordre des mots peut parfois être important pour certains moteurs de recherche, alors commencez par le mot le plus pertinent.

- **Combinaison des mots-clé et options:** pour bien trouver les résultats, il est possible de combiner les mots-clés, ou d'ajouter des options.

  - Les guillemets « » : Cela permet de rechercher une expression exacte. Par exemple, "Le blog du Modérateur" présente les sites où les mots "Le blog du Modérateur" sont présents, mais uniquement dans cet ordre.
  - Le signe moins "-" Cela permet d'exclure un terme. Par exemple, la requête "astuces recherche Google" permet de trouver des pages contenant les mots "astuces" et "recherche", mais exclut celles qui contiennent le terme "Google".
  - Les deux points ".." : deux nombres séparés par deux points permettent de rechercher tous les nombres de la plage spécifiée. Smartphone 200..400 euros liste les téléphones compris entre 200 et 400 euros.
  - AND : exclue les pages ne contenant pas les termes spécifiés. Blog AND Modérateur présente les sites contenant ces deux termes, mais pas ceux contenant uniquement l'un des deux.
  - L'astérisque * est souvent utilisé pour connaître l'intégralité d'une phrase ou d'une expression. Qui vole * vole * permet de retrouver l'expression qui vole un œuf vole un bœuf.
  - Le symbole + permet de tenir compte d'un mot vide (comme par exemple le, la, les, du, avec, de, lettres et chiffres uniques, http, .com, etc) OR : l'opérateur permet de rechercher un terme, ou un autre.

- **Filtrage des résultats (augmentation ou réduction du nombre de résultats)**

Le nombre de résultats obtenus peut être soit assez élevé ou assez réduit (voire même nul) selon

les

• Choisissez plusieurs mots en rapport avec le thème qui vous intéresse (un nombre de 3 mots-clés est une bonne moyenne).

• Cherchez dans les pages retenues des indices qui permettront d'affiner.

• Réduisez le nombre de mots (en ne conservant que les plus importants).