Comment fonctionne un moteur de recherche

Introduction

Internet est complexe

La recherche n'y est pas une science exacte.

Plusieurs problèmes se posent en effet :

- L'information est **abondante** : plusieurs millions de pages
- Elle est **hétérogène**

  En effet, divers sont
  - Les sources,
  - Les supports,
  - Les contenus,
  - Les modes d'accès à l'information.
- Elle n'est **pas structurée**,
- Le renouvellement est continuel et, par conséquent, le **contenu est éphémère**.
- Il est **multilingue**.
- Les contenus ne sont **pas toujours fiables**.

Les besoins sont divers

Chaque utilisateur a des attentes, des priorités différentes, en fonction des projets qu'il développe.

Les moteurs de recherche

Aujourd'hui, la plupart des internautes utilisent quasi exclusivement les moteurs de recherche pour trouver de l'information.

Qu'est-ce qu'un moteur de recherche ?

C'est un outil automatique pour collecter et indexer un grand nombre de pages web ; il recense des sites et pas annuaires.

Fonctionnement du moteur :

- **Le robot d'exploration** : collecte le contenu de milliards de pages dans une base de données structurée en champs (le texte de la page, le titre de la page, l'adresse url + des infos variées comme la provenance géographique, des liens externes...)
  Le stockage et la réactualisation se font à vitesse du robot et pas en temps réel

- **L'indexation automatique** : il construit un index de la base de données par explorations successives. L'index contient tous les mots significatifs des pages visitées
  Le mode d'indexation et la notion de « mot significatif» sont différents selon le moteur.
  Généralement, ce sont tous les mots, sauf les mots vides (articles, conjonctions…).
  Certains indexent aussi les métadonnées (pas Google) : c'est-à-dire les infos cachées dans l'en-tête de la page, tels que les mots clés.
  Souvent, il s'agit seulement d'une indexation morphologique : les mots sont considérés comme des chaînes de caractères il ne s'agit donc pas d'un traitement linguistique.
- **L'interrogation de l'index** : elle se fait par l'entrée d'un ou 2 mots clés. Chaque page contenant les mots est considérée comme pertinente, ce qui entraîne un grand nombre de réponses
  La concurrence entre les moteurs vient de la pertinence des premiers résultats.

La collecte au cœur du système

- La page d'accueil des sites soumis est explorée par le moteur.
- Celui-ci suit également les liens internes et externes du site.

Attention

- Il s'agit rarement de toutes les pages et elles sont rarement explorées en même temps. Cela entraîne le fait qu'en ce qui concerne les sites très volumineux, il n'y a aucune garantie que la totalité des pages soit collectée.
- La mise à jour des index est variable d'un jour à plusieurs semaines si bien que certaines pages ne sont parfois plus disponibles dans l'intervalle.
- Pour certains sites, la mise à jour est « partiale » : on se préoccupe d'abord des sites les plus populaires et les plus mouvants.
- Les index sont répartis sur plusieurs machines, ce qui entraîne une disparité dans le comptage des résultats.

Volume de l'indexation

En 1998, aucun moteur n'indexait plus de 16% du web et la couverture globale des 11 moteurs était d'environ 40%

Google et Yahoo ont progressé mais l'exhaustivité est impossible

En 2005, la couverture était de :

- Google : 76,16 %
- Yahoo : 69,32 %
- MSN : 61,9 %
- AskTeoma : 57,62 %

Pour en savoir plus: http://www.cs.uiowa.edu/~asignori/web-size/size

Tendance à la normalisation des principales syntaxes d'interrogation

Grandes constantes

- Par défaut, tous les termes sont recherchés
- « + » force le passage pour un mot considéré comme non significatif
- « - » exclut un terme
- « OR » : indique que l'un ou l'autre terme est recherché
- La troncature est peu répandue
- Une combinaison est possible des termes
- Il existe parfois des options plus précises

Algorithmes de pertinence

Variété

Selon le moteur de recherche, il existe des variantes :

- Dans la requête :
  - La position des mots n'est pas toujours neutre.
  - Il est parfois fait référence à l'internaute (si son profil est connu ou en fonction de requêtes antérieures)
- Dans les pages de résultats :
  - La densité des mots-clés (nombre d'occurrence /nombre de termes, de pages) varie.
  - Leur présence dans le titre ou le premier tiers de la page peut seule être prise en considération.
  - La mise en exergue du texte (gras, taille des caractères) peut être déterminante.
  - La présence des mots clés dans l'adresse peut avoir une influence
  - La proximité des mots clés également.
- Dans la base de données du moteur :
  - La rareté des mots peut compter : la pondération est plus importante que celle des mots communs
  - La popularité des pages est un élément déterminant, c'est-à-dire l'indice de clic ou indice de popularité

Popularité comme mesure de pertinence est apparue dans les années 1980 et elle a fait le succès de Google

Concrètement, cela signifie que le nombre et la qualité des liens pointant sur la page (la mesure de sa popularité), selon cette conception, détermine la pertinence d'une ressource ; ce qui s'exprime selon la formule : « le plus important, c'est qui vous connaît » cf. la mesure de la crédibilité de l'auteur scientifique en fonction du nombre de ses publications.

L'**avantage** de ce fonctionnement est de donner une meilleure visibilité aux sites incontournables dans un domaine de recherche

Son **inconvénient** est de pénaliser les nouveaux venus