CHAPITRE III : LA RECHERCHE D'INFORMATION SUR LE WEB

3.1 INTRODUCTION :

Un moteur de recherche (Searchbot en anglais) est une machine spécifique (matérielle et logicielle) chargée d'indexer des pages web afin de permettre une recherche à l'aide de mots-clés dans un formulaire de recherche.

Des robots (logiciels), appelés spiders (en français araignées) doivent parcourir la toile en suivant récursivement les liens des millions de pages Web et indexent le contenu dans de gigantesques bases de données afin de permettre leur interrogation.

Aucun moteur de recherche ne peut parcourir la totalité des pages en une journée (ce processus prend généralement plusieurs semaines), chaque moteur adopte donc sa propre stratégie,

Lorsque l'utilisateur d'un moteur de recherche remplit le formulaire, il spécifie les mots qu'il cherche (éventuellement ceux qu'il ne souhaite pas) grâce aux opérateurs booléens "et", "ou", "non" ... (symbolisés par +, -,...), la requête est envoyée au moteur de recherche qui consulte ses bases de données pour chacun des mots.

Selon une étude faite en Aout 2007 par Comscore , le nombre de recherches pendant ce mois sur les différents moteurs de recherches a donné :

- Google : environ 36 milliards de recherches, soit 60 % des 61 milliards de recherches sur Internet
- Yahoo : 8,5 milliards de recherches, soit 14 % du total
- Baidu : « le Google chinois », : 3,3 milliards de requêtes, soit 5,4 % du total,
- Bing, remplaçant de Live Search, (Microsoft) : 2,1 milliards de recherches, soit 3,4 %.

3.2 ANNUAIRE ET MOTEUR DE RECHERCHE :

Un **annuaire** est un outil de recherche qui recense un certain nombre de sites au travers de fiches descriptives comprenant, en règle générale, le titre, l'adresse (l'URL) et un bref commentaire. Chaque site est inscrit dans une ou plusieurs catégorie(s) - on parle également de rubrique(s) -. Ces outils peuvent ainsi être considérés comme les « pages jaunes » du Web.

Lorsqu'un mot-clé est saisi dans le formulaire proposé, l'annuaire effectue une recherche sur les occurrences de ce terme dans ses fiches descriptives de site, et non pas dans le contenu des pages du site en question. Il s'agit là de la différence la plus notable avec les moteurs de recherche.

Exemples d'annuaires : Yahoo, Looksmart, …etc.

Le moteur de recherche fonctionne sur un concept radicalement différent de celui de l'annuaire. Des robots logiciels (appelés crawlers ou spiders) scrutent le Web, vont de page en page (en fait de lien en lien) et sauvegardent au fur et à mesure le contenu texte des documents rencontrés, constituant ainsi un "index", c'est-à-dire une collection plus ou moins grande de pages Web.

Lorsque l'internaute saisit un mot clé dans le formulaire proposé, le moteur va en rechercher les occurrences dans son index, c'est-à-dire dans le contenu (le texte) des pages Web sauvegardées au préalable. Une fois identifié le "lot" de pages contenant le terme demandé.

Exemples de moteurs de recherches : Google, Altavista, …

Le moteur de recherche effectue donc ses recherches sur des **pages Web**, alors que l'annuaire effectue ses recherches sur les **sites Web**.

Les métamoteurs sont des outils de recherche qui interrogent non leur propre base de données, mais celles de plusieurs moteurs de recherche simultanément et affichent à l'internaute une synthèse pertinente. Exemple : Ixquick, Scroogle .

3.2 PRINCIPE DE FONCTIONNEMENT :

Le fonctionnement d'un moteur de recherche se décompose en trois étapes principales : l'exploration, l'indexation et la recherche.

1. **L'exploration ou crawl** : le web est systématiquement exploré par un robot d'indexation suivant tous les hyperliens qu'il trouve et récupérant les ressources jugées intéressantes. Un moteur de recherche est d'abord un outil d'indexation, c'est-à-dire qu'il dispose d'une technologie de collecte de documents à distance sur les sites Web, via un outil que l'on appelle robot ou bot. Un robot d'indexation dispose de sa propre signature (comme chaque navigateur web). Googlebot est le user agent (signature) du crawler de Google

2. **L'indexation** des ressources récupérées consiste à extraire les mots considérés comme significatifs du corpus à explorer. Les mots extraits sont enregistrés dans une base de données organisée comme un gigantesque dictionnaire inverse ou, plus exactement, comme l'index terminologique d'un ouvrage, qui permet de retrouver rapidement dans quel chapitre de l'ouvrage se situe un terme significatif donné. Les termes non significatifs s'appellent des mots vides. Les termes significatifs sont associés à une valeur de poids. Ce poids correspond à une probabilité d'apparition du mot dans un document.

3. **La recherche** correspond à la partie requêtes du moteur, qui restitue les résultats. Un algorithme est appliqué pour identifier dans le corpus documentaire (en utilisant l'index), les documents qui correspondent le mieux aux mots contenus dans la requête, afin de présenter les résultats des recherches par ordre de pertinence supposée. Les algorithmes de recherche font l'objet de très nombreuses investigations scientifiques. Les moteurs de recherche les plus simples se contentent de requêtes booléennes pour comparer les mots d'une requête avec ceux des documents. Mais cette méthode atteint vite ses limites sur des corpus volumineux. Les moteurs plus évoluées utilisent les probabilités pour mettre en perspective le poids des mots dans une requête avec ceux contenus dans les documents. Cette formule est utilisée pour construire des vecteurs de mots, comparés dans un espaces vectoriel. Pour améliorer encore les performances d'un moteur, il existe de nombreuses techniques, la plus connue étant celle du PageRank de Google qui permet de pondérer une mesure de cosinus en utilisant un indice de notoriété de pages. Les recherches les plus récentes utilisent la méthode dites d'analyse sémantique latente qui tente d'introduire l'idée de co-occurrences dans la recherche de résultats (le terme "voiture" est automatiquement associé à ses mots proches tels que "garage" ou un nom de marque dans le critère de recherche).

Des modules complémentaires sont souvent utilisés en association avec les trois processus de bases du moteur de recherche, comme le correcteur orthographique : il permet de corriger les erreurs introduites dans les mots de la requête.

3.3 INFRASTRUCTURE NECESSAIRE :

Selon certaines sources le moteur de recherches Google est capable de traiter près de 40.000 requêtes par secondes, et possède un index de plus de 1000 milliards de pages web. De telles performances, nécessite une infrastructure matérielle et logicielle gigantesque.

Pour stocker les données et répondre aux requêtes, Google avait le choix entre des très gros serveurs ou un grand nombre de PC traditionnels. Voici une comparaison des coûts de deux solutions étudiées, qui explique pourquoi Google a choisi la seconde (source : conférence Google en 2004) :

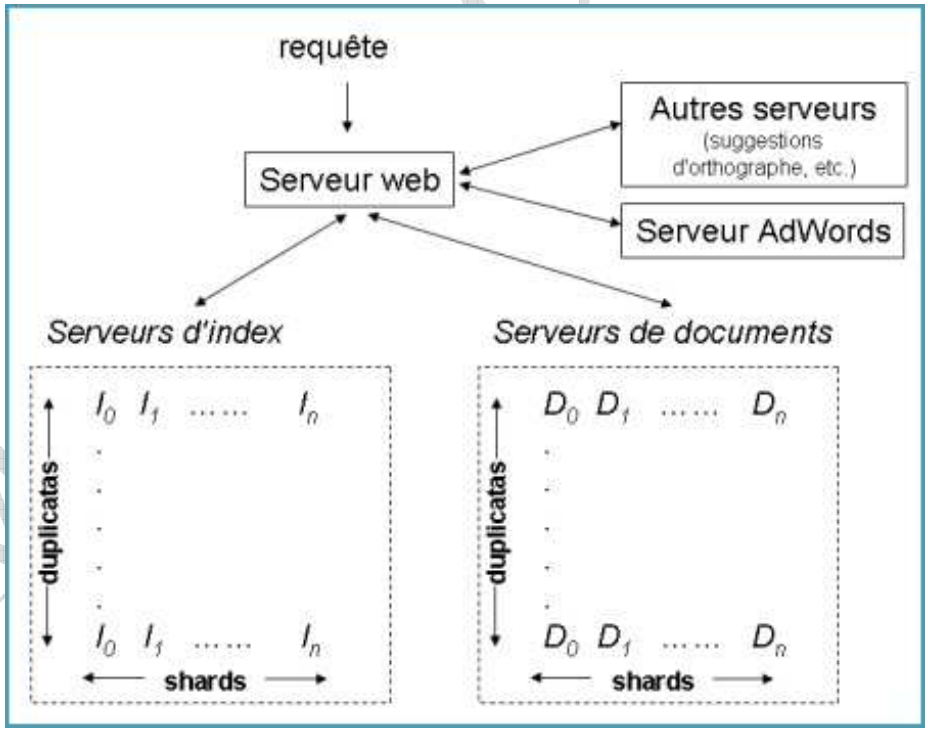
Serveur IBM eServer xSeries 440
- 8 processeurs Xeon de 2 GHz
- 65 Go de RAM
- 8 To de disque
- Couût : environ 758 000 $

Rack de 88 machines
- 176 processeurs Xeon de 2 GHz (88 x 2)
- 176 Go de RAM (88 x 2)
- 7 To de disque
- Coût : environ 278 000 $

Chaque jour dans les data centers de Google, plusieurs machines tombent en panne. Tous les développements d'applications sont donc conçus pour être tolérants aux pannes.

L'index de Google est découpé en petits bouts afin qu'ils puissent être stockés sur chaque machine. Chacun de ces bouts est appelé un *shard*. La répartition des documents en shards se base entre autres sur le PageRank . Chaque shard est dupliqué pour être sur plusieurs machines.

Google accorde beaucoup d'importance au temps de réponse à chaque requête. Pour ne pas excéder 0,5 seconde, Google déploie des data centers dans le monde entier afin de rapprocher les serveurs des utilisateurs.



Infrastructure technique du moteur Google.

3.4 PERFORMANCES D'UN MOTEUR DE RECHERCHE :

Parmi les performances espérées d'un moteur de recherche, on peut citer notamment :

1. **Simplicité, pertinence** : L'usage d'une simple boîte de recherche dans laquelle l'internaute va pouvoir s'exprimer naturellement est souvent possible. Parallèlement la notion de recherche avancée ou de recherche multicritère n'est que très rarement utilisée par le visiteur et ne doit donc pas être présentée comme mode de recherche par défaut. Le visiteur va mesurer la pertinence et la performance de la solution par correspondance entre l'offre produit qui lui est présentée et sa demande ainsi que la rapidité de la réponse du moteur de recherche. Le temps de réponse ne devra pas dépasser les 0,5 sec. Avant toute chose, le moteur de recherche se doit d'être pertinent et doit pouvoir répondre à tout type de question posée par l'internaute : question sur un article, question sur un service, …etc.

2. -**La navigation à facettes** : La navigation à facettes (faceted navigation) est l'une des avancées majeure des outils de recherche de nouvelle génération. Elle s'appuie sur une recherche simple, sans mention de critères. En réponse, le moteur va présenter une vue de la répartition de l'ensemble des réponses selon diverses catégories ainsi que le nombre d'occurrences. L'internaute choisit alors la facette qui l'intéresse, et va ainsi restreindre les résultats à ceux qui correspondent à son choix..

3. **Tolérance orthographique** : Les sources d'erreur dans la recherche d'un article ou d'un produit sont multiples (faute d'orthographe, recherche de référence mal saisie...) Le moteur de recherche se doit de présenter une réponse en adéquation avec la saisie du visiteur. Des algorithmes de recherche et la construction de dictionnaires propres aux sites permettent de mettre en place une stratégie de réponse qui offre à l'internaute la possibilité de reformuler sa question ou lui propose des articles ou produits correspondants à sa recherche..

4. **Support du langage naturel** : Le support du langage naturel permet à l'internaute de s'exprimer dans son propre vocabulaire, sans avoir à se soucier des données du site et de leurs modélisations. Il pourra par exemple obtenir des réponses précises à une question du type : "une location d'un appartement F3 de moins de 15000 DA". Ce type de module est une aide à la recherche qui réduit considérablement la notion de multicritère lorsque le modèle de données est très détaillé et précis. Ce module est capable d'extraire de l'expression recherchée l'ensemble des critères spécifiés et de répondre avec énormément de précision.

5. **L'autocomplétion** : Cette fonction permet de compléter en live la saisie de l'internaute dans la boîte de recherche. Cela permet de lui suggérer des recherches avant même qu'il ait terminé de l'exprimer.