Mickaël MARCHAL
Nadia TEA
Promo 2007

EFREI
ECOLE D'INGENIEURS DES TECHNOLOGIES DE L'INFORMATION ET DU MANAGEMENT

Les moteurs de recherche

Comment indexent-ils l'information, et comment la restituent-ils ?

Sommaire

I. Préambule

"Internet est comme l'union de toutes les bibliothèques du monde entier, où malheureusement, quelqu'un s'est amusé à renverser tous les livres exposés sur les étagères"

Notre monde tend à s'informatiser de plus en plus notamment par l'entremise d'Internet. En effet, ce dernier s'est inséminé dans notre quotidien et dans notre vie professionnelle. Dès que l'on se pose une question surfer sur la toile nous permet d'obtenir de promptes réponses bien qu'il existe des milliards de sites. Grâce à des outils rapides et simples d'utilisation que sont les moteurs de recherche, nous sommes à même de trouver ce que nous cherchons au travers d'un petit champ de texte, comme l'itinéraire pour se rendre à notre lieu de vacances ou à un rendez-vous d'affaires, pour appeler le plombier, trouver un tutoriel. Internet nous offre une multitude de possibilités et les moteurs de recherche nous permettent d'y accéder plus facilement. Mais que se cache-t-il derrière cette interface sobre, cet outil devenu indispensable? Il s'agit certainement là d'une question que seuls peu de gens se posent et qui pourtant parait très intéressante. Quels sont leurs principes ? Comment sont indexées toutes ces pages Internet ? Comment les moteurs de recherche interprètent-ils nos requêtes? Comment les moteurs obtiennent les résultats en fonction des demandes?

L'objectif de ce mémoire est d'offrir un aperçu des différentes techniques utilisées ou pouvant être utilisées par les moteurs de recherche pour indexer les pages, et les outils mis à disposition de l'internaute pour saisir ses requêtes. Nous tenterons de répondre à cette problématique par une approche d'ingénieurs : comprendre aussi bien les aspects « business » que scientifiques, car nous sommes conscients que les enjeux de la recherche sur le net dépassent largement le cadre de la théorie scientifique.

II. Principe des services de recherche

II.1 Petit historique de la recherche sur Internet

Le premier moteur de recherche apparait en 1990, crée par Adam Emtage, étudiant à Mc Gill (Québec). Ce moteur, dénommé **Archie**, comportait les principes de base du moteur de recherche : on remplissait une base de données, que le moteur faisait correspondre aux requêtes des utilisateurs. Le Web de l'époque comportait seulement quelques centaines de sites, et Archie resta un projet universitaire.

Mais le saut technologique le plus important fut introduit par **Wanderer** (« le Vagabond ») en 1993 par Matthew Gray. Il fut le premier moteur à déployer des **robots d'indexation** (spiders). L'idée de base, qui était de mesurer la croissance du Web, fut rapidement remaniée pour arriver au premier moteur de recherche à indexation automatique (Bot search) Ce moteur a d'ailleurs causé un certain nombre de problèmes, car il retournait plusieurs centaines de fois par jour sur certains sites et les ralentissait.

En octobre 2003, le successeur d'Archie fait son apparition : **Aliweb** (Archie-like indexing the web). Ce moteur repose sur la **soumission manuelle** de sites. Le moteur se basait sur les **mots clés** et les descriptions fournies au moment de l'inscription pour effectuer la recherche.

Le premier moteur intelligent fut **Excite** (1993). Construit par six étudiants de Stanford, il se base sur **l'analyse statistique** des mots.

Enfin, en 1994, c'est la naissance de **Yahoo**, le premier « grand » service de recherche, crée également par des étudiants de Stanford. Mais à la différence des outils de l'époque, Yahoo se base sur un **annuaire**, pas sur un moteur de recherche. Les résultats sont **sélectionnés** et indexés **par l'homme**. En quelques mois, Yahoo devient le plus important portail du Web.

Les années 1995-1997 voient l'apparition des grands moteurs de recherche (**Excite, Hotbot, Lycos**…). **Altavista**, crée par un français et jugé efficace et rapide, deviendra la star des moteurs de recherche du moment jusqu'aux années 2000, détrôné par Google.

De son coté, **Inktomi** développe la première activité de recherche destinée aux **entreprises**. C'est la première fois que les moteurs de recherche ciblent les professionnels.

Enfin, c'est en 1998 que naît **Google**, crée par Sergei Brin et Larry Page, encore une fois étudiants de Stanford. Google va littéralement révolutionner le monde de moteurs de recherche grâce à sa simplicité et son efficacité. L'interface dépouillée se charge instantanément sur les connexions bas-débit de l'époque, et la technologie d'indexation est inédite : Google se base sur le **nombre de liens pointant sur une page** pour en déterminer sa **pertinence**.

Vers 2001-2002, **l'éclatement de la bulle internet** fait disparaitre les premiers moteurs de recherche, et seuls les plus grands survivent. C'est l'ère moderne de la recherche internet.

II.2 Les enjeux d'aujourd'hui et de demain

Selon le cabinet d'études Nielsen Netratings, environ **70% des visites** d'un site web proviennent d'un **moteur ou service de recherche**, le reste provient de « bonnes adresses » données par un proche, ou de la publicité. Quand on sait que la Toile est devenue le vecteur principal des échanges commerciaux entre entreprises (B to B), et un canal majeur dans la vente et les services aux particuliers (B to C), on comprend mieux l'**enjeu considérable de la recherche d'informations sur le net**.

Toujours selon Nielsen Netratings, 1,2 milliards de recherches ont été effectuées par les américains au mois de Mai 2004. Ce chiffre constitue une augmentation de 30% par rapport à l'année précédente. L'essentiel des recherches est effectué sur une petite dizaine de moteurs. Nombre d'entre eux sont hautement symboliques et sont détenus par des multinationales parmi les plus importantes du monde : **Microsoft, Google, AOL TimeWarner**…

Ces sociétés se livrent une **guerre sans merci** pour gagner quelques parts de marché sur leurs concurrents. Pendant longtemps, le cheval de bataille principal était le nombre de sites référencés. Ainsi, les géants de la recherche ont mené une véritable « **course à l'indexation** », récoltant des dizaines de millions de sites par mois. Mais depuis l'éclatement de la bulle internet en 2002-2003, le concept du « **plus gros** » s'est évanoui face au concept du **« plus large** ». Dorénavant, les moteurs de recherche se diversifient et permettent de trouver des images, du son, des vidéos, des livres, et ainsi de suite. Il est maintenant possible d'élargir le champ d'une recherche à de **nombreux supports ou médias**.

Aujourd'hui **Google** possède une avance confortable sur ses concurrents grâce à tous ses services annexes : Google Images, Google Groups, Google Books, Google Suggest, Froogle, etc. Il possède aujourd'hui **40 à 80%** des parts de marché selon les pays. Mais le nouveau virage entamé par les moteurs de recherche pourrait bien renverser la tendance.

Qui ne s'est jamais senti perdu devant les milliers, parfois les millions de réponses trouvées pour une simple requête ? L'indexation à tout prix permet une **réponse exhaustive** mais bien souvent **inexploitable**. Les internautes expérimentés l'ont bien compris et ciblent dorénavant mieux leurs requêtes. Ainsi, selon Onestat, les requêtes complexes (de 4 mots et plus) deviennent de plus en plus fréquentes, alors que les requêtes simples (1 ou 2 mots) diminuent fortement. Le moteur de recherche, initialement défini comme « un outil simple et universel d'accéder à l'information sur le net » **perd peu à peu en efficacité et en précision**.

Voilà l'enjeu pour les prochaines années : les moteurs de recherche devront se différencier par le concept du « **plus efficace** ». Le moteur devra répondre de la façon la plus précise possible aux attentes les plus simples de l'internaute. Cela passe par deux approches :

- **Indexer plus précisément**, mieux cibler le contenu de la page et mieux déterminer sa pertinence,
- Fournir à l'utilisateur des moyens de **questionner le moteur plus précisément** et plus intuitivement.

Ces deux points font largement appel aux concepts dérivés de l'Intelligence Artificielle, et en particulier à linguistique et à la reconnaissance des formes.

II.3 Les familles d'outils de recherche

Bien que les différents moteurs de recherche arborent une interface d'utilisation simple, ils possèdent une puissance impressionnante. En effet, des milliers de pages Internet sont créées chaque jour et doivent être répertoriées pour être référencées. De plus, 70% des visites de sites proviennent des moteurs de recherche. C'est pourquoi tous ces sites nouvellement créés cherchent un bon référencement auprès de nos outils de recherche préférés tels que Google, Yahoo, Alta Vista. des billions de pages sont alors à stocker. Seulement pour parvenir à les faire correspondre aux demandes de l'utilisateur, il est nécessaire de leur associer certaines critères afin de les retrouver facilement. Pour cette raison, les moteurs de recherche utilisent des bases de données considérables, et pour faire face aux millions de requêtes envoyées chaque minute, ils doivent impérativement posséder une bande passante immense. Pour gérer tous ces éléments; les différents moteurs de recherche aménagent, répertorient, classent, recherchent tous ces sites Internet différents. Aussi on peut distinguer trois types de moteur de recherche : l'annuaire, le moteur de recherche à proprement dit ou search engine et enfin le métamoteur.

II.3.1 L'Annuaire

L'annuaire (ou directory) est en fait une liste de liens subdivisés en catégories suivant une structure en arbre, accompagnée d'une brève description. Bien que ce procédé fût pionnier en la matière, il tend à disparaître. En effet, le fait de devoir sélectionner les catégories dans lequel on recherche suppose que l'on sache exactement où chercher. Et on peut se demander où se positionne le site qui appartient à plusieurs catégories. Mais à cette question, les moteurs utilisant ce procédé vous répliqueront qu'ils se trouvent dans toutes celles susceptibles de correspondre. Néanmoins, on doit lui reconnaître un gros avantage, celui de mettre en quelque sorte dans le contexte, ainsi les recherches dans la base de données sont diminuées, en plus d'obtenir des résultats plus pertinents.

Quelques annuaires : Yahoo, Voilà, …

II.3.2 Le moteur de recherche (Search engine)

Le plupart des moteurs préfèrent désormais chercher directement les résultats dans la base de données grâce à des requêtes spécifiques basées sur les critères entrés. C'est le type de service le plus utilisé actuellement, et c'est celui que nous détaillerons dans les prochains chapitres.

Quelques moteurs de recherche : Google, MSN Search, Lycos, Altavista, Excite…

II.3.3 Le Métamoteur

Certains moteurs ont opté pour une solution plus économique, puisqu'ils utilisent les bases de données des autres moteurs. Ainsi les métamoteurs rassemblent plusieurs moteurs de recherche. L'un des avantages évidents de ce procédé pourrait être d'obtenir des résultats plus pertinents, puisque la recherche s'étend sur un plus grand nombre de sites indexés, sites figurant sur tel moteur, mais pas sur un autre. Néanmoins, la redondance de sites affichés peut-être un inconvénient gênant. De même que l'augmentation considérable de résultat qui peut engendrer un délai d'attente supérieur. De plus, le fait d'envoyer différentes requêtes à différents serveurs rallonge également le temps de réponse.

Quelques métamoteurs : Infospace, Askjeeves, MyWay, Websearch.com…

C'est ici que s'achève notre tout d'horizon des services de recherche sur le Web. Nous allons maintenant étudier plus en détail les moteurs de recherche eux-mêmes, d'un point de vue plus technique et plus fonctionnel.

III. L’indexation des pages

Sans sa base de données de 11 milliards de pages, Google ne serait rien. C’est cette constatation qui démontre la criticité de l’indexation des pages web par un moteur de recherche.

Avant de pouvoir lancer des requêtes sur le serveur, il est indispensable de remplir la base de données, puis de la mettre quotidiennement à jour. Pour cela, il existe plusieurs moyens. Le plus simple est d’acheter une base de données à un autre moteur de recherche, comme l’a fait AOL en 2002, qui a acheté la base de données de Google pour plusieurs millions de dollars. Mais cette pratique reste marginale, car la plupart des moteurs constituent-eux même leur base de données. Seulement avant d'insérer effectivement une page, il est nécessaire de vérifier sa provenance pour éviter les faux sites (ou spams). A cet effet, des crawlers ou logiciels équivalents ont été créés : ils contrôlent si les domaines recensés dans la base de données existent réellement. Et il est indispensable d'extraire les paramètres tels que les mots clés du site qui vont permettre de le retrouver selon les requêtes effectuées.

Seulement, arriver à déterminer la pertinence des critères est un acte difficile à programmer. Pour cette raison, la part humaine est un élément indispensable. De ces faits, l'indexation de site Internet ne se résume pas à une simple insertion dans la base de données. il existe deux procédés : soumettre une page directement à un moteur de recherche ou attendre qu'un logiciel d'un moteur détecte cette page et le répertorie dans sa base de données.

III.1 La soumission directe d'une page Internet

Sur la plupart des moteurs de recherche, il est possible de soumettre son site, en indiquant son adresse au moteur. Certains moteurs demandent à cette occasion une participation financière au soumetteur, trouvant ainsi une source de revenu. De plus, vu que l'indexation du site est payante, la proportion de faux sites ou spams devient quasi-inexistante. Le soumetteur pour sa part gagne l'avantage d'un ajout assuré dans la base de données. D'autres moteurs doivent examiner plus attentivement les soumissions. Certains moteurs comme Yahoo préfèrent analyser les candidatures et le reste du procédé manuellement. Cela a des avantages évidents : la certitude de l'insertion de sites de qualité; ainsi que la bonne adéquation des paramètres par rapport au site proposé et donc des réponses aux requêtes plus pertinentes. Cependant cette solution nécessite une grande main d'œuvre, un délai d'attente de plusieurs semaines et 70% des sites proposés sont refusés dans le cas de Yahoo. Certains programmes tels que Directory engine permettent de cataloguer automatiquement une page internet en développant une architecture structurée par argument permettant une intervention humaine minimum.

III.2 L’Insertion automatique des pages

Certains moteurs comme Google préfèrent ou complètent l'indexation des nouvelles pages par des algorithmes mathématiques plus complexes, mais moins coûteuse en main d'oeuvre et

plus courte en durée. Pour cela, ces moteurs utilisent des programmes automatisés appelés Spiders, Crawlers, Bots ou encore Robots (l'équivalent français, peu utilisé, est « robot d'indexation »). Ces spiders parcourent sans interruption les pages déjà indexées, naviguant de lien en lien à la recherche de nouveaux liens et en recense les pages. Ensuite des logiciels tel que ICE (Intelligence Concept Extraction) permettent d'établir les rapports entre les termes que les spiders ou crawlers ont trouvé déterminants dans ces pages, les mots clés et les autres paramètres.

III.3 Le principe du Crawler

Le crawler est un logiciel d'analyse structurelle, syntaxique et sémantique de page Web (parser en anglais). Pour chaque page, il extrait les éléments jugés significatifs et pertinents, afin de se constituer une base de mots-clés relatifs à la page analysée. Lorsque le spider détecte des liens vers d'autres pages, il les garde en mémoire dans une base de données contenant des adresses restant à analyser. Une fois la page analysée, le spider regarde dans sa base de données la prochaine page à visiter, et ainsi de suite.

Un crawler est donc capable de voyager de lien en lien, de page et page, et donc de site en site sans aucune intervention humaine. Il utilise le concept de l'hyperlien, fondement même du Web.

Mais le robot n'agit pas pour autant à l'aveuglette. Lorsqu'il arrive sur une page, il détermine tout d'abord s'il « connait » la page, autrement dit si il l'a déjà indexé. Si c'est le cas, il fera un passage plus rapide, se contentant de relever les modifications effectuées depuis sa dernière visite. Le robot doit aussi déterminer s'il est autorisé ou non à indexer la page. Cela se fait au moyen de directives standards mises au point par Google, et contenues dans un fichier intitulé « robots.txt ». Ce fichier permet de limiter ou modifier la façon dont les moteurs de recherche référencent un site. Il est ainsi possible de préserver les fichiers sensibles de la divulgation.

Les statistiques concernant les robots sont jalousement tenues secrètes. Mais selon toute vraisemblance, les spiders les plus puissants sont capables d'analyser des centaines de milliers de pages par jour.

III.4 L'analyse d'un site

C'est l'étape la plus complexe, et par conséquent, c'est celle dont les secrets sont le mieux gardés. L'analyse d'un site diffère d'un moteur à l'autre, mais les grandes lignes de ce processus sont connues et communes à tous les moteurs de recherche.

III.4.1 Extraction des métadonnées

La première étape consiste à extraire les métadonnées du fichier analysé. Ces métadonnées sont des informations renseignant sur la nature du document. Ainsi, il s'agit souvent :

- De l'extension du fichier

- De la date de création et de dernière modification
- De la taille du fichier
- Du nom de fichier et de l'adresse URL à laquelle il se trouve

Les autres métadonnées dépendent du type de fichier. Pour une image, il s'agira des dimensions de celle-ci, pour une vidéo, de sa longueur, pour un fichier MP3 on cherchera a extraire les tags ID3 (données renseignant sur l'auteur, l'album… d'un fichier mp3).
Dans le cas d'une page Web (fichier HTML ou équivalent), il s'agira d'extraire le titre de la page (balise <title>), et les données contenues dans les balises <meta>. Par exemple :

```
<title>Bienvenue sur le site de Yé Mistikrik ?, l'association de théâtre de
l'EFREI !</title>

<meta name="keywords"
content="theatre,théatre,efrei,association,etudiante,asso,mistikrik,ye" />

<meta name="description" content="Yé Mistikrik? est l'association de
théatre de l'EFREI. Venez découvrir sur ce site notre assoce, la troupe,
les pièces que nous avons joué et les photos !" />
```

La première ligne montre le titre du document HTML. La deuxième fournit des mots clés en rapport avec le site (ici le site de l'association de théâtre de l'EFREI – http://assos.efrei.fr/theatre/index.php). La dernière contient une brève description du contenu du site. On trouve aussi d'autres balises <meta> renseignant sur l'auteur, le logiciel utilisé pour générer la page, etc.

III.4.2 Les métadonnées ne suffisent pas !

Idéalement, ces données devraient suffire à déterminer les mots-clés associés à la page. Cela fut vrai les premières années de la recherche internet. Mais malheureusement, de nombreux abus on fait perdre toute crédibilité à ces métadonnées. Certaines personnes n'hésitaient pas à indiquer des mots-clés très demandés mais sans relation avec leur page, simplement pour s'attirer plus de visiteurs. Ainsi, et paradoxalement, on ne peut pas faire confiance à l'auteur pour la classification des pages. Il faut « voir au delà » et utiliser des méthodes bien plus poussées afin de déterminer la vraie signification d'une page.

III.4.3 L'analyse structurelle globale

Une page HTML contient de nombreux éléments, mais la plupart d'entre eux n'ont qu'un rapport lointain avec le contenu de la page. Les menus, bannières, publicités… doivent être éliminés par le spider pour ne garder que le contenu d'une page, d'où il pourra extraire les mots-clés.

Pour cela, le robot se base sur des règles statistiques. En effet, la plupart du temps, les images plus longues que larges situées en tête de page sont des bannières publicitaires. L'analyse de l'URL de l'image sera un poids de plus pour déterminer la pertinence d'une image.
De même, les menus de navigation sont composés de liens comportant peu de texte encadré par de nombreuses balises, et sont souvent situés à gauche de la page.

```
<h2>Le Site...</h2>
        <a href="/index.php">Accueil</a>
        <a href="/mika/cv.php">Consulter mon CV</a>
        <a href="/contact.php">Crédits et Contact</a>
        <a href="/design/index.php">Webdesign</a>
        <a href="/labo/index.php">Web applications</a>
        <a href="/extras/index.php">Extras</a>
```

(exemple de structure typique d'un menu - http://www.lesitedemika.org)

Par ce procédé, le robot arrive dans la plupart des cas à déterminer la structure globale d'une page, et à faire abstraction des éléments non significatifs.

III.4.4 L'analyse sémantique

Une fois le contenu du site isolé, le robot peut alors l'analyser pour en extraire les mots-clés. Il est à noter que l'objectif d'un spider n'est pas de comprendre le contenu d'une page, mais d'en déterminer les éléments importants.

Pour cela, plusieurs opérations sont réalisées :

- Tout d'abord, le robot détermine la langue dans laquelle est rédigée la page. Il peut le faire grâce aux métadonnées, mais il est plus sûr pour lui d'utiliser des algorithmes de reconnaissance des langues. Ceux-ci se basent sur des statistiques pour déterminer une langue à partir de mots reconnaissables, ou de la fréquence de certains mots ou expressions.

Je m'appelle **Mickaël MARCHAL**, j'ai 21 ans, j'habite en Seine-Saint-Denis (France) et je suis actuellement en deuxième année de cycle ingénieur à l'EFREI, où je prépare mon diplôme d'ingénieur en Technologies de l'Information et du Management. Je suis donc un **futur ingénieur informaticien**, passionné comme il se doit par les technologies et le développement.

→ Langue reconnue : Français

- Ensuite, le robot retire les articles de liaisons et les mots communs de la langue. En français, des mots comme « de », « un », « le », « ou » seront retirés.

Je m'appelle **Mickaël MARCHAL**, j'ai 21 ans, j'habite en Seine-Saint-Denis (France) et je suis actuellement en deuxième année de cycle ingénieur à l'EFREI, où je prépare mon diplôme d'ingénieur en Technologies de l'Information et du Management. Je suis donc un **futur ingénieur informaticien**, passionné comme il se doit par les technologies et le développement.

→ Les mots de liaison ont été supprimés (en gris)

- Le robot utilise aussi la structure HTML du document pour juger de la pertinence des mots. Ainsi, une phrase en gras, ou un texte écrit plus gros sera considérée comme « important » par un spider. De même, les liens, ou les mots en majuscules verront leur importance s'accroitre. Mais si tout le site est en gras, les poids des mots restera le même.

Je m'appelle **Mickaël MARCHAL**, j'ai 21 ans, j'habite en Seine-Saint-Denis (France) et je suis actuellement en deuxième année de cycle ingénieur à l'EFREI, où je prépare mon diplôme d'ingénieur en Technologies de l'Information et du Management. Je suis donc un **futur ingénieur informaticien**, passionné comme il se doit par les technologies et le développement.

→ Les mots en gras, en majuscules, et les liens ont été accentués

- Le robot va ensuite parcourir les liens situés au fil du texte et déterminer leurs mots-clés. Si des mots clés sont identiques à ceux trouvés dans le texte, il y a de fortes chances que le texte analysé soit corrélé avec ces mots clés.

Je m'appelle **Mickaël MARCHAL**, j'ai 21 ans, j'habite en Seine-Saint-Denis (France) et je suis actuellement en deuxième année de cycle ingénieur à l'EFREI, où je prépare mon diplôme d'ingénieur en Technologies de l'Information et du Management. Je suis donc un **futur ingénieur informaticien**, passionné comme il se doit par les technologies et le développement.

→le lien « EFREI » renvoie à http://www.efrei.fr, référencé avec les mots clés « ingénieur », « informatique », « management », « école », etc. Ces mots deviennent alors plus importants (en rouge).

- La répétition de mots (et encore plus, d'expressions) au fil d'un texte est aussi un indicateur d'importance (toutefois moindre).

Je m'appelle **Mickaël MARCHAL**, j'ai 21 ans, j'habite en Seine-Saint-Denis (France) et je suis actuellement en deuxième année de cycle ingénieur à l'EFREI, où je prépare mon diplôme d'ingénieur en Technologies de l'Information et du Management. Je suis donc un **futur ingénieur informaticien**, passionné comme il se doit par les technologies et le développement.

→les mots « ingénieur » et « technologies » sont accentués (en bleu)

Au final, le spider aura déterminé les mots les plus importants comme étant :

- Mickaël
- Marchal
- « Mickaël Marchal » (car les deux mots sont tous les deux en gras côte à cote, l'expression *Mickaël Marchal* est accentuée)
- EFREI
- Ingénieur
- Technologies
- Information
- Management
- Futur
- « Futur ingénieur informaticien »
- Etc.

III.4.5 La pondération des mots clés

Une fois les principaux mots clés extraits, ils sont pondérés en fonction de leur rareté sur la toile. Plus un mot est rare sur le web, plus l'importance qui lui est accordée sur le site analysé sera grande, et inversement.

Ainsi, dans notre exemple, « Mickaël », « Marchal », et « EFREI » auront une plus grande importance que « futur », « technologies », « information ».

III.4.6 Les limites de l'indexation

Voici les principes de base de l'analyse faite par les moteurs de recherches. Mais bien sûr, les mécanismes intimes des robots ne sont pas connus, et de nombreuses hypothèses circulent à ce sujet : ce sont les techniques de référencement. Certaines paraissent logiques et vérifiables ; d'autres tiennent plus des recettes de grand-mère. Une chose est sure : le référencement n'est pas une science exacte, mais est une discipline sur laquelle il faut se pencher pour faire connaître son site sur le net.

Voici d'ailleurs quelques conseils généraux à appliquer pour un bon référencement :

1. Contenu
2. Nombreux liens externes
3. Liens externes de qualité
4. Bons titres
5. S'inscrire dans les grands annuaires
6. Pas de frames, et une bonne adresse de site
7. Un site toujours disponible
8. Interconnexion de vos pages
9. Pas de site sans texte
10. Mises à jour régulières du site

Mais il faut aussi se rappeler que la multitude des contenus, des mises en page ou des informations ne font pas non plus de l'indexation une science exacte. Les moteurs de recherche commettent de nombreuses erreurs de référencement, entrainant des lapsus plus ou moins loufoques. Ainsi, en tapant le mot « failure » (échec) sur Google.com, le premier résultat renvoyé est la biographie officielle de George W. Bush, sur le site de la Maison Blanche !

III.5 Cas particulier : Google et le PageRank

Lors de son lancement en 1998, Google a introduit un nouveau concept révolutionnant le petit monde des moteurs de recherche : le PageRank.

Le PageRank est une méthode inventée par Google pour mesurer l'importance relative des pages du web, que l'on appelle souvent la popularité. Elle est basée sur la topologie du web, c'est-à-dire sur l'étude des liens entre les pages.

L'idée principale est que si une page A fait un lien vers une page B, alors c'est que la page A juge que la page B est suffisamment importante pour mériter d'être citée et d'être proposée aux visiteurs. Ce lien de A vers B augmente le PageRank de B.

Deux idées supplémentaires mais essentielles viennent la compléter :

- l'augmentation du PageRank de la page B est d'autant plus importante que le PageRank de la page A est élevé. En d'autres termes, il est bien plus efficace d'avoir un lien depuis la page d'accueil de Google que depuis une page d'un site personnel.

- l'augmentation du PageRank de la page B est d'autant plus importante que la page A fait peu de liens. En d'autres termes, si la page A juge qu'il n'y a qu'une page qui mérite un lien, alors il est normal que le PageRank de la page B augmente plus que dans le cas où de nombreuses pages obtiennent un lien.

Le PageRank est donc un moyen assez puissant de déterminer la popularité d'une page – autrement dit, sa qualité. Le PageRank est une note donnée sur 10 à chaque page. Les « petits sites » ont souvent un PR entre 0 et 2, les sites de moyenne fréquentation ont un PR généralement situé entre 3 et 6. Les PR supérieurs sont réservés aux gros ou très gros sites : seuls les géants comme Google, Yahoo, Amazon, la NASA ou Microsoft peuvent prétendre au PageRank 10.

Il faut toutefois noter que PageRank n'influe pas sur les mots clés du site, mais sur l'ordre de classement du site lui-même sur Google, lorsque la requête saisie contient les mots-clés du site.

En recherchant « http » sur Google (mot présent sur quasiment tous les sites de la planète), les premiers résultats sont les sites à PageRank élevé : Microsoft, puis le W3C (organisme de standardisation du web, donc du protocole http), Altavista, CNN, Yahoo, etc.

La quête du PageRank élevé est l'objectif numéro un des spécialistes en référencement, quand on connait les parts de marché de Google. Augmenter le PR d'une unité peut amener des centaines de milliers de nouveaux visiteurs sur un site !

IV. Traitement des requêtes

Indexer, c'est bien, mais encore faut-il restituer les fruits de ce référencement aux utilisateurs. Il faut surtout fournir à ces derniers les outils pour interroger le moteur de recherche de la façon la plus complète possible.

IV.1 Langage naturel ou mots clés ?

Vu la richesse de la langue française, le nombre de questions possibles pour une simple interrogation est trop important pour arriver à interpréter la demande, sans oublier le nombre d'erreurs de l'utilisateur (fautes d'orthographes, de frappes, de français) qui croît énormément. Il devient alors quasiment impossible d'analyser et traiter la demande. Le moteur de recherche *ask.com* fait une exception, puisqu'il a tenu à utiliser ce procédé. Toutefois, l'utilisation du langage naturel n'a pas permis d'augmenter la popularité de ce moteur, qui reste très peu utilisé. En effet, le langage naturel comporte également des inconvénients pour l'utilisateur puisqu'il doit réfléchir à une question adéquate et qui peut être prise en compte correctement. D'autre part, effectuer des recherches sur de simples mots-clés ouvre de nombreuses possibilités d'affinement de la recherche et ainsi offre une plus grande panoplie de réponses. Cela permet d'être plus rapide, puisque le moteur impose une syntaxe simple et facile à parser ou interpréter.

Gageons qu'avec le temps et les progrès de l'Intelligence Artificielle, le langage naturel s'imposera tôt ou tard pour les moteurs de recherche. Mais pour le moment, les ressources nécessaires pour le déploiement d'un tel système à grande échelle en font un outil peu utilisable et peu rentable.

C'est pourquoi, la quasi-totalité des moteurs utilisent ce procédé. Pour un souci d'efficacité, la plupart des moteurs offre de nombreuses fonctionnalités pour affiner les recherches. Ainsi, selon les moteurs, il est possible de chercher des sites selon une extension de domaine bien définie, dans une langue particulière, des images, etc... mais il s'agit là d'options avancées.

IV.2 Principales différences entre moteurs

Voici un tableau comparatif répertoriant les différences entre les principaux moteurs de recherche par mots-clés :

Nombre de pages
C'est le nombre total de pages indexés dans le moteur de recherche

Contenu
Indique le type d'informations cataloguées

Version française
Disponibilité d'une version française

Recherche avancée
Indique si des fonctions de recherche avancées sont disponibles

Opérateur standard
Recherche les pages contenant tous les mots insérés (AND) ou seulement un seul d'eux (OR

Recherche de phrases
Spécifie si on a la possibilité et comment chercher des phrases composées de plusieurs termes

Pluriel/Singulier
Indique si les mots sont cherchés aussi dans leur forme plurielle

Mots ignorés
Indique si la recherche ignore les articles, certaines particules ou des mots très communs (par ex.: web, Internet,etc.)

Champs de recherche
Reporte la possibilité d'utiliser des paramètres afin de préciser la recherche

Recherche d'images, sons, Java, etc.
Indique la possibilité de spécifier files multimédia

Recherche par langue
Indique s'il est possible de sélectionner uniquement les pages web écrites dans une langue prédéterminée

Filtre pour les enfants
Permet d'exclure des résultats des recherches les sites web à caractère pornographique, violent, etc.

Regroupement de résultats
Pour chaque site trouvé qu'une seule page est reportée

Personnalisation des résultats des recherches
Indique la possibilité de personnaliser la façon de laquelle les résultats des recherches sont affichés

Mémorisation des préférences personnelles
Indique s'il est possible de mémoriser des paramètres personnels

Fonctions spéciales
Indique les fonctions particulières

Catalogue sites
Indique s'il existe une archive des sites subdivisés par arguments

Temps d'insertion dans les sites
C'est le temps qu'il faut compter depuis l'enregistrement avant d'apparaître dans le moteur de recherche

	AltaVista	AllTheWeb (FAST)	HotBot	Google
N° de pages (millions)	1100	2100	3300	3000
Contenu	Web, Usenet, Images, MP3, Audio, Video	Web, News, Photos, Videos, Audio, FTP, PDF, SWF	Web, Usenet, News	Web, Usenet, News, PDF (22 millions), DOC, XLS, PPT, RTF
Version française	OUI	NON	OUI	OUI
Recherche avancée	OUI	OUI	OUI	OUI
Opérateur standard	AND	AND	AND	AND
Recherche de phrases	Oui, en utilisant les guillemets (")	Oui, en utilisant les guillemets (")	Oui, en utilisant les guillemets (")	Oui, en utilisant les guillemets (")
Pluriel/ singulier	Oui, utilisant "*"; par ex.: "auto*" trouve autos, automobile, etc.	NON	OUI (seulement anglais)	OUI (seulement anglais)
Mots ignorés	OUI	OUI	OUI	OUI
Champs de recherche	Oui: langue des pages, title, keyword, url, link, site, image. Ex.: "image:foto.jpg"	Oui: langue, filtre de mots et domaines,date de mise à jour et dimension des pages	Oui: position géographique, type de fichier, date, url, etc. Utiliser les "meta words" dans le texte de la recherche ou sinon les options	Oui: title, keyword, url, link, site, image. Ex.: "image:photo.jpg"
Recherche par langue	OUI	OUI	OUI	OUI
Filtre pour les enfants	OUI	OUI	OUI	NON
Regroupement de résultats	OUI	OUI	OUI	OUI
Personnalisation des résultats des recherches	Oui: beaucoup de possibilités	Oui: beaucoup de possibilités	Oui: visualise 10/25/50/75/100 sites par page; description brève, complète ou seulement l'adresse	Oui: beaucoup de possibilités
Mémorisation des préférences personnelles	OUI	OUI	OUI	OUI
Fonctions spéciales	-	Recherche rapide images, personnalisation interface	Recherche sur Lycos, Google et AskJeeves en option, personnalisation/skin interface	Traduction des pages web;Bouton "J'ai de la chance" amène directement au 1er résultat,version cache des pages, autres...
Annuaire de sites	OUI	OUI	OUI	OUI
Temps d'insertion dans les sites	7-45 j.	9-14 jours	20-30 j.	15-30 j.

IV.3 Mots clés

Dès la formulation de la requête, on peut exclure les sites comportant un certain mot ou rechercher les sites contenant deux mots juxtaposés et non dispersé dans le texte grâce à des opérateurs décrits dans le tableau ci-dessous :

Tableau des opérateurs

Opérateur Booléen	Résultat
+nom1 +nom2	Renvoie les documents contenant les 2 mots cherchés
nom1 + nom2 ou **nom1 nom2**	Renvoie les documents contenant un des 2 mots cherchés (ou les 2)
+nom1 -nom2	Renvoie les documents ne contenant pas le mot qui suit l'opérateur -
nom*	Renvoie les documents contenant les mots proches du mot cherché
guillemets	Renvoie les documents contenant la phrase entière
Essentiel (+)	Le symbole "+" classifie un mot comme essentiel
Exclusion (-)	Le symbole "-" exclut un mot de la recherche

	Fast	Altavista	Excite	Google	Voilà	Webcrawler
ET	+	+	+ ou AND	Par défaut ou +	+ ou ET ou AND	+ ou AND
OU	non	par défaut	par défaut ou OR	non	par défaut ou OR ou OU	par défaut ou OR
PROCHE	non	NEAR ou ~	non	non	PROCHE ou NEAR	non
SAUF	-	AND NOT ou !	- ou AND NOT	-	- ou NOT ou SANS ou AND NOT	- ou NOT
PHRASE EXACTE	" "	" "	" "	" "	" "	" "
TRONCATURE	non	*	non	non	non	non
ACCENTS PRIS EN COMPTE	oui	oui	oui	oui	indiff.	oui
MAJUSCULES	indiff.	indiff.	indiff.	indiff.	indiff.	indiff.
MOTS ORDONNES	indiff.	oui	indiff.	oui	indiff.	indiff.

Source: Le Monde Interactif

Comme l'utilisateur est toujours susceptible de mal orthographier un mot ou faire une faute de frappe, les moteurs de recherche proposent selon les cas d'effectuer la recherche sur les synonymes des mots ou des mots ayant une orthographe proche.

Pour étendre la recherche, certains moteurs comme AltaVista proposent de traduire les mots que l'utilisateur a entrés avant de lancer la requête.

IV.4 Résultats trouvés en fonction des requêtes lancées

Tous les types de moteurs de recherche sans exception doivent être capables d'analyser la demande de l'utilisateur, puis d'effectuer des requêtes afin de lui fournir les résultats les plus pertinents possibles. Pour cela, ils composent des requêtes selon les mots ou expressions entrées par l'utilisateur. La première recherche consiste à chercher dans les mots-clés des pages indexées les mots recherchés dans le même ordre. Certains moteurs vont jusqu'à rechercher dans le contenu des pages. Ensuite les moteurs recherchent également les mots mais dans un ordre différent, certains autorisent les singulier/ pluriel au même titre, si bien que si on a entré un mot au pluriel, il effectuera aussi une requête avec le mot au singulier. Des associations de mots sont aussi possibles, comme par exemple sur Google, en saisissant *"avenant installation logiciel"*, on trouve des résultats avec les mots en gras *"avenant installer logiciel"*. Ainsi les moteurs analysent d'abord les mots ou expressions entrés et en extraient les données par exemple le fait de rechercher deux mots juxtaposés et non comme deux mots distincts, le fait d'ôter les requêtes reportant à un certain mot, etc...

IV.4.1 Recherches avancées

La plupart des moteurs offrent des fonctionnalités supplémentaires grâce aux recherches avancées. Mais cela ne fait que rallonger la requête lancée puisqu'il s'agit de faire appel à d'autres paramètres déjà associés aux pages indexées.

IV.4.2 Ordre des résultats lors du rendu des réponses

Etant donné la masse colossale des pages indexées, le nombre de résultats est important, de l'ordre de la centaine voir plus pour chaque requête. Même la recherche de ses prénom et nom juxtaposés peut rapporter plus d'une centaine de pages, bien que de moins en moins pertinents. Personnellement, j'ai eu la surprise de trouver que mon identité était associée à une maison de thé-restaurant (eh oui, mon nom(prénom) sont déjà pris comme nom de domaine. Peut-on faire un procès pour çà?). Pour cette raison, les utilisateurs ne lisent plus les résultats à partir de la troisième page de vingt résultats. D'où l'importance de positionner les résultats les plus pertinents en tête de liste. Aussi certains moteurs commercialisent l'emplacement des sites selon les mots clés et l'on trouve de colossales enchères pour certains d'entre eux. Pour les autres pages, le degré de pertinence (et le PageRank) prennent tout leur sens : le nombre de critères satisfaits par la page indexée par rapport à la recherche lancée, et la côte de popularité qui permet de déterminer sans part humaine du côté du moteur, le degré d'adéquation des résultats aux demandes des utilisateurs.

Un site est plus côté quand d'autres sites pointent vers lui, mais il est encore plus côté si lui contient des liens pointant vers ces sites qui le pointent. Ainsi, les partenariats peuvent être très intéressants.

Certains moteurs vont jusqu'à personnaliser la liste. En effet, le moteur peut enregistrer les sites les plus visités par l'utilisateur avec sa durée et peut ainsi être capable de graduer l'appréciation d'un site par cet utilisateur. Il est possible alors d'instaurer des degrés d'appréciation sur ce site avec des commentaires et de supprimer des sites automatiquement des listes de résultats. Les sites les plus appréciés par l'utilisateur se positionnent en tête de liste. L'on peut se rendre compte rapidement compte d'un inconvénient majeur : vu que ces données sont stockées sur l'ordinateur de l'utilisateur, celui-ci ne pourra bénéficier uniquement de ces améliorations sur un de ses pc et ceux qui partagent leur ordinateur hésiteront simplement à les utiliser.

V. Conclusion

Ce mémoire ne fait qu’effleurer le monde complexe et impitoyable des moteurs de recherche. Nombre de secrets sont bien gardés, et nombre d’incertitudes planent sur les algorithmes utilisés par Google et ses pairs. Mais, une chose est certaine : les principes de base sont bien là. Comprendre le fonctionnement, même succinctement, d’un moteur de recherche, permet de mieux entrevoir les possibilités et la puissance qu’ils nous offrent en tant qu’utilisateurs, mais aussi en tant qu’entrepreneurs.

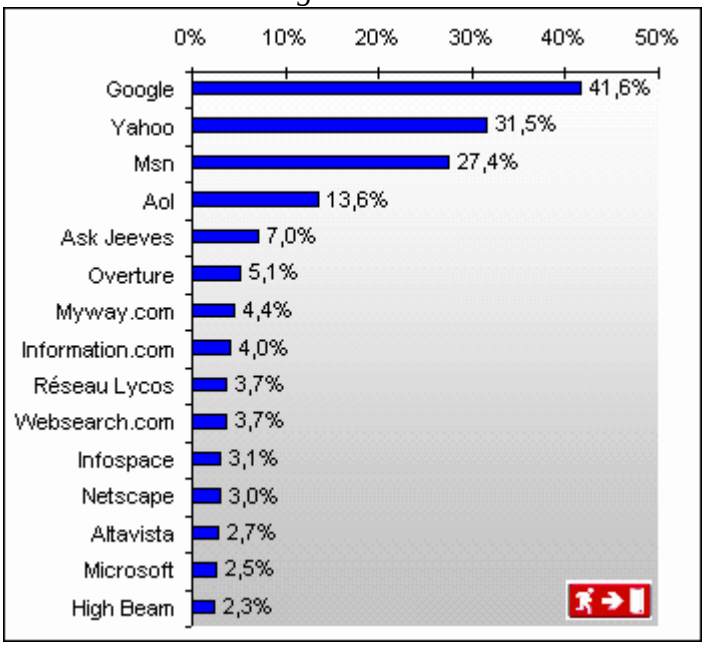
VI. Annexes

VI.1 Quelques Chiffres

Comment arrive-t-on sur un site Web ?
Etude CommerceNet/Nielsen Media - Juillet 1997

71,0%	*Par les moteurs de recherche*
9,8%	*Conseillé par amis ou collègues*
8,5%	*Journaux quotidiens ou périodiques*
8,4%	*Lien sur un autre site*
8,1%	*Par hasard, en surfant*
3,6%	*Signalé à la TV*
3,3%	*Guides sur les sites web*

Parts de marché des moteurs de recherche aux USA
Etude Nielsen Netratings - Juillet 2004



Parts de marché des moteurs de recherche en France
Etude Indicateur.com - Février 2006

Moteur	PDM
1. Google	72,20
2. MSN	6,37
3. Yahoo	5.74
4. Voila	5,07
5. Aol	1,54
6. Free	1,2
7. Club Internet	0,44
8. Altavista	0,34
9. 9Online	0,19
10. Lycos	0,14

IV.2 Bibliographie / Webographie

- http://www.lesmoteursderecherche.com/
- http://docs.abondance.com/portails.html
- http://www.webrankinfo.com/
- http://www.dsi-info.ca/moteurs-de-recherche/langages/operateurs-logiques.html
- http://www.bius.jussieu.fr/web/recherch.html
- http://www.asktibbs.com/php/article.php3?id_article=11
- http://www.uhb.fr/ccb/moteurs.htm
- http://www.commentcamarche.net/utile/recherch.php3
- http://searchenginewatch.com/
- http://www.indicateur.com
- http://www.search-marketing.info/search-engine-history/
- http://www.answers.com/topic/archie-search-engine
- http://www.owil.org/lexique/r.htm

- PC Expert n°161 (février 2006)
- Redesign Web 2.0 : Conduite de projet – Kelly Goto & Emily Cotler – Ed. Eyrolles
- Projet de site Web 2eme édition – Nicolas Chu – Ed. Eyrolles