2

Fonctionnement des moteurs de recherche et des annuaires

Avant d'y référencer votre site, savez-vous ce que l'outil de recherche que vous utilisez au quotidien a « dans le ventre » ? La réponse à cette question n'est pas si évidente. En effet, bien que les moteurs de recherche tels que Google, Yahoo! ou encore Live Search semblent très simples d'utilisation, leur fonctionnement « sous leur capot » est en réalité très complexe et élaboré. De même, le processus « humain » d'indexation dans les annuaires doit être connu pour mieux y référencer son site. Nous vous proposons dans ce chapitre une analyse globale du fonctionnement des moteurs de recherche, des annuaires ainsi que des processus qui sont mis en œuvre pour traiter les documents, stocker les informations les concernant et restituer des résultats suite aux requêtes des utilisateurs. Bien maîtriser le fonctionnement d'un outil de recherche permet de mieux appréhender le référencement et l'optimisation de son site.

Comment fonctionne un moteur de recherche ?

Un moteur de recherche est un ensemble de logiciels parcourant le Web puis indexant automatiquement les pages visitées. Trois étapes sont indispensables à son fonctionnement :

- la collecte d'informations (ou *crawl*) grâce à des robots (ou *spiders* ou encore *crawlers*) ;
- l'indexation des données collectées et la constitution d'une base de données de documents nommée « index » ;
- le traitement des requêtes, avec tout particulièrement un système d'interrogation de l'index et de classement des résultats en fonction de critères de pertinence suite à la saisie de mots-clés par l'utilisateur.

Comme nous l'avons vu dans les pages précédentes, les pages de résultats des moteurs de recherche affichent deux principaux types de contenu : les liens « organiques » ou « naturels », obtenus grâce au crawl du Web et les liens sponsorisés.

Nous allons nous concentrer ici sur les techniques utilisées par les moteurs de recherche pour indexer et retrouver des liens naturels. Nous n'aborderons pas le traitement spécifique des liens sponsorisés, qui seront quant à eux étudiés au chapitre 8 de cet ouvrage.

Technologies utilisées par les principaux portails de recherche

En dehors des trois leaders du marché (Google, Yahoo! et Microsoft Live Search), de nombreux moteurs n'utilisent pas leurs propres technologies de recherche mais sous-traitent cette partie auprès de grands moteurs. En fait, il n'existe que peu de « fournisseurs de technologie » sur le marché : Google, Yahoo!, Microsoft, Teoma (Ask.com), Wisenut (abandonné aujourd'hui) et Gigablast sont les principaux aux États-Unis, comme sur le plan mondial. En France, les acteurs majeurs sont Exalead, Mirago et Voila, qui côtoient d'autres noms moins connus comme Antidot, Deepindex, Seekport, Misterbot, Megaglobe ou encore Dir.com (la liste n'est pas exhaustive). Voici un récapitulatif des technologies utilisées par les différents portails de recherche en 2007.

Tableau 2-1 Technologies de recherche utilisées par les principaux portails de recherche francophones en 2007

Sites web / Technologies de recherche	Google	Yahoo!	Live Search	Antidot	Exalead	Mirago	Teoma	Voila
Google	X							
Yahoo!		X						
Live Search			X					
Mozbot.fr	X							
Orange								X
Ask.com France							X	
AOL.fr	X							
Free	X	X	X				X	
Neuf	X							
Club Internet	X							
Exalead					X			
Lycos		X						
La Poste				X				
Ujiko (Kartoo)		X						

Tableau 2-2 Technologies de recherche utilisées par les principaux portails de recherche anglophones en 2007

Sites web / Technologies de recherche	Google	Yahoo!	Live Search	Antidot	Exalead	Mirago	Teoma	Voila
Google	X							
Yahoo!		X						
Live Search			X					
AllTheWeb		X						
A9 (Amazon)			X					
AltaVista		X						
Ask.com							X	
Eurekster		X						
Exalead					X			
Hotbot	X						X	
Mozbot.com	X							
Lycos							X	

Mise à jour

Les données de ce tableau, valables à la fin 2007, peuvent fluctuer en fonction des contrats signés d'une année sur l'autre. Une mise à jour de ces informations est disponible à l'adresse : *http://docs.abondance.com/portails.html.*

Principe de fonctionnement d'un moteur de recherche

Plusieurs étapes sont nécessaires pour le bon fonctionnement d'un moteur de recherche : dans un premier temps, des robots explorent le Web de lien en lien et récupèrent des informations (phase de crawl). Ces informations sont ensuite indexées par des moteurs d'indexation, les termes répertoriés enrichissant un index - une base de données des mots contenus dans les pages - régulièrement mis à jour. Enfin, une interface de recherche permet de restituer des résultats aux utilisateurs en les classant par ordre de pertinence (phase de ranking).

Les crawlers ou spiders

Les spiders (également appelés agents, crawlers, robots ou encore bots) sont des programmes de navigation visitant en permanence les pages web et leurs liens en vue d'indexer leurs contenus. Ils parcourent les liens hypertextes entre les pages et reviennent

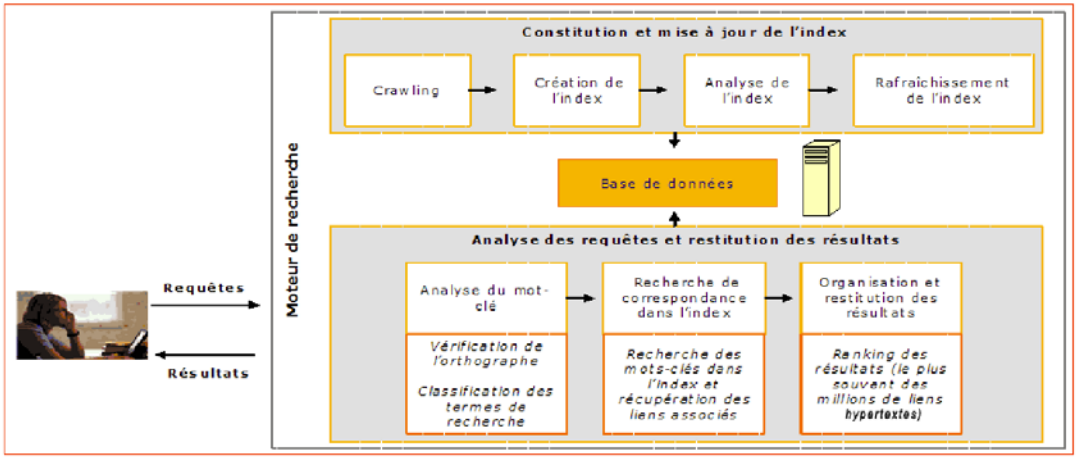

Figure 2-1

Les différentes étapes du fonctionnement des moteurs de recherche

périodiquement visiter les pages retenues pour prendre en compte les éventuelles modifications.

Un spider est donc un logiciel très simple mais redoutablement efficace. Il ne sait faire que deux choses :

- lire des pages web et stocker leur contenu (leur code HTML) sur les disques durs du moteur ;
- détecter les liens dans ces pages et les suivre pour identifier de nouvelles pages web.

Le processus est immuable : le spider trouve une page, la sauvegarde, détecte les liens qu'elle contient, se rend aux pages de destination de ces liens, les sauvegarde, y détecte les liens, etc. Et cela 24h/24... L'outil parcourt donc inlassablement le Web pour y détecter des pages web en suivant des liens... Une image communément répandue pour un spider serait celle d'un internaute fou qui lirait et mémoriserait toutes les pages web qui lui sont proposées tout en cliquant sur tous les liens qu'elles contiennent pour aller sur d'autres documents, etc.

Parmi les spiders connus, citons notamment Googlebot de Google, Yahoo! Slurp de Yahoo, Henri Le Robot du moteur Mirago ou encore le plus récent, MSNBot de Microsoft Live Search.

Mais parcourir le Web ne suffit pas. En effet, lorsqu'un spider arrive sur une page, il commence par vérifier qu'il ne la connaît pas déjà. S'il la connaît, il contrôle si la version découverte est plus récente que celle qu'il possède déjà... En cas de réponse positive, il supprime l'ancienne version et la remplace par la nouvelle. L'index se met ainsi automatiquement à jour.

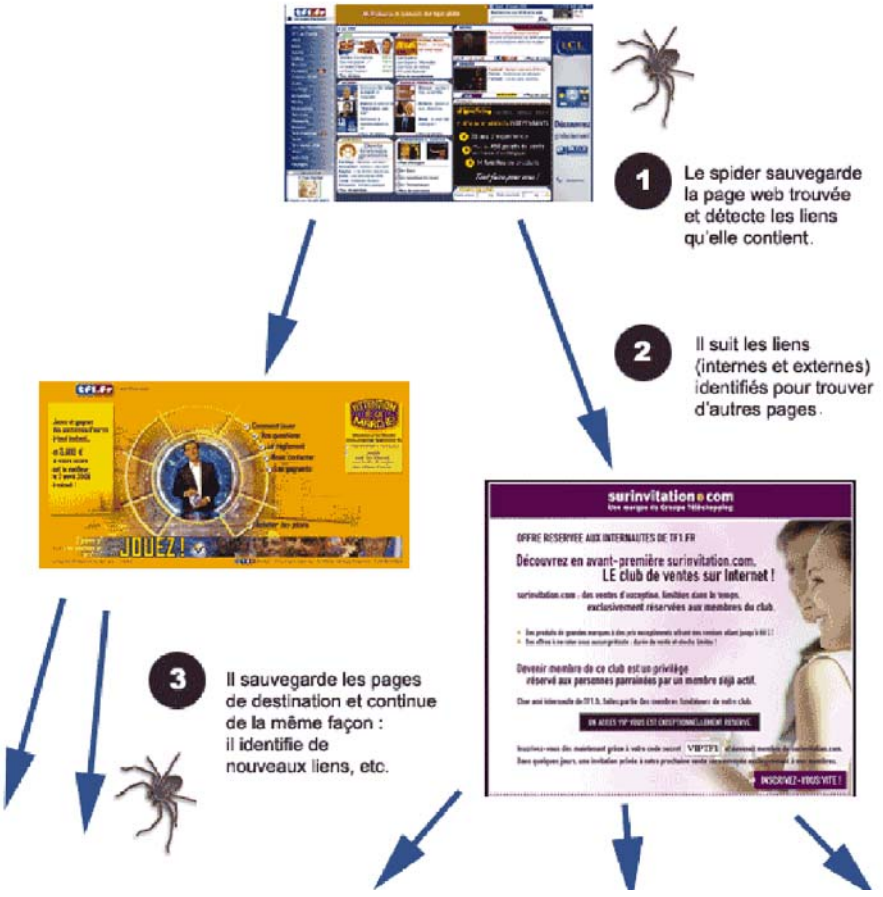

Figure 2-2

Principe de fonctionnement d'un spider

> **Le fichier robots.txt pour indiquer aux spiders ce qu'ils ne doivent pas faire**
>
> Le fichier `robots.txt` est utilisé par les webmasters pour indiquer aux spiders les pages qu'ils souhaitent indexer ou non (voir chapitre 7).

De la « Google Dance » au « Minty Fresh »...

Il y a quelques années de cela, les mises à jour des index des moteurs étaient mensuelles. Chaque mois, le moteur mettait à jour ses données en supprimant un ancien index pour le remplacer par un nouveau, mis à jour pendant 30 jours par ses robots, scrutant le Web à la recherche de nouveaux documents ou de versions plus récentes de pages déjà en sa possession. Cette période avait notamment été appelée la « Google Dance » par certains

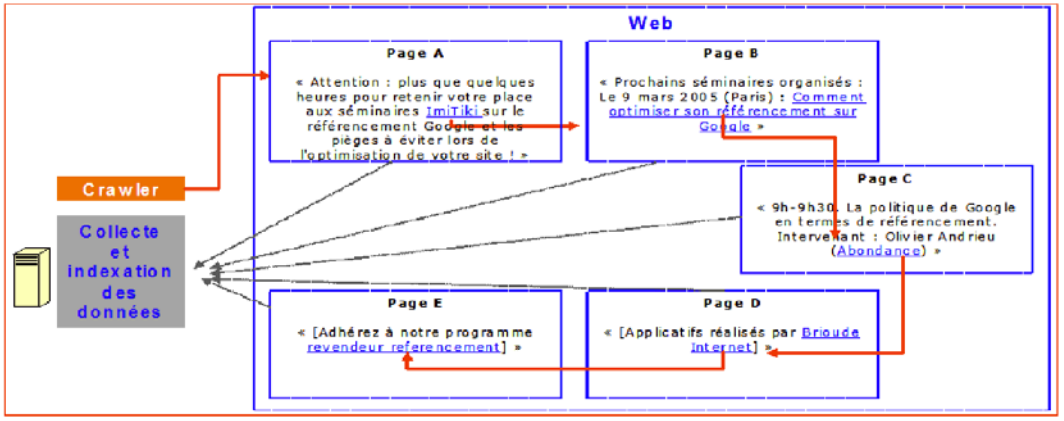

Figure 2-3

Processus de crawl (ou crawling) des robots en suivant les liens trouvés dans les pages web

Quels critères de décision ?

Pour savoir si une page est plus récente qu'une version déjà sauvegardée, le moteur de recherche va jouer sur plusieurs facteurs complémentaires :

- la date de dernière modification du document fournie par le serveur ;
- la taille de la page en kilo-octets ;
- le taux de modification du code HTML du document (son contenu) ;
- les zones modifiées : charte graphique ou contenu réel. Ainsi, certains moteurs pourront estimer que l'ajout d'un lien dans un menu de navigation ne constitue pas une modification suffisante pour être prise en compte... Ils sauront différencier « charte graphique et de navigation » avec « contenu réel » et ne prendre en compte que la deuxième forme de modifications...

webmasters. Pour l'anecdote, elle fut d'ailleurs longtemps indexée (c'est le cas de le dire) sur les phases de pleine lune... On savait, à cette époque, que lorsque la pleine lune approchait, un nouvel index était en préparation chez Google... Nous verrons plus loin que l'expression « Google Dance » désigne désormais tout autre chose.

Ce système de mise à jour mensuelle des index n'a plus cours aujourd'hui. La plupart des moteurs gèrent le crawling de manière différenciée et non linéaire. Ils visitent plus fréquemment les pages à fort taux de renouvellement des contenus (très souvent mises à jour) et se rendent moins souvent sur les pages « statiques ». Ainsi, une page qui est mise à jour quotidiennement (par exemple, un site d'actualité) sera visitée chaque jour ou tous les deux jours par le robot tandis qu'une page rarement modifiée sera « crawlée » toutes les quatre semaines en moyenne. De plus, la mise à jour du document dans l'index du moteur est *quasi* immédiate. Ainsi, une page souvent mise à jour sera le plus souvent disponible à la recherche sur le moteur un ou deux jours plus tard. Ces pages récemment crawlées sont par exemple identifiables sur Google car la date de crawling est affichée. Exemple ici sur une recherche effectuée le 28 mars 2007 :

Figure 2-4
Affichage par Google de la date d'indexation de la page

Le résultat proposé à la figure 2-4 montre bien que la page proposée a été crawlée (sauvegardée par les spiders) deux jours auparavant et qu'elle a été immédiatement traitée et disponible dans les résultats de recherche.

Le « Minty Fresh Indexing »

À la mi-2007, Google a accéléré son processus de prise en compte de documents, certaines pages se retrouvant dans l'index du moteur quelques minutes seulement après leur création/modification. Ce phénomène est appelé *Minty Fresh Indexing* par le moteur de recherche. Matt Cutts, dont nous avons déjà parlé au chapitre précédent, explique ce concept sur son blog : *http://www.mattcutts.com/blog/minty-fresh-indexing/.*

On pourra noter que la technique de suivi des liens hypertextes par les spiders peut poser plusieurs problèmes pour :

- l'indexation des pages qui ne sont liées à aucune autre et ne peuvent donc pas être repérées par les crawlers qui n'ont aucun lien à « se mettre sous la dent » (si tant est que les robots aient des dents...). Il en est ainsi des sites qui viennent d'être créés et qui n'ont pas encore de *backlinks* (liens entrants) qui pointent vers eux ;
- l'indexation des pages dynamiques de périodiques ou de bases de données (ces pages étant moins facilement prises en compte, nous y reviendrons au chapitre 5) ;
- les pages pointées par des documents proposant des liens qui ne sont pas pris en compte par les moteurs de recherche, comme beaucoup de ceux écrits en langage JavaScript. Là aussi, nous y reviendrons (chapitre 5).

Le passage des spiders sur les sites peut être vérifié par les webmasters en analysant les fichiers « logs » sur les serveurs (ces fichiers indiquent l'historique des connexions qui ont eu lieu sur le site, y compris celles des spiders). La plupart des outils statistiques comprennent dans leurs graphiques ou données une partie « visites des robots ». Attention cependant, ces outils doivent le plus souvent être spécifiquement configurés pour prendre en compte tous les robots émanant de moteurs français. Les outils statistiques, notamment d'origine américaine, ne prennent pas toujours en compte ces spiders « régionaux »...

Pour tracer les robots...

Plusieurs applications en ligne permettent également d'analyser les visites des robots sur des pages données (voir notamment les solutions gratuites *http://www.robotstats.com/* et *http://www.spywords.com/).* Des « marqueurs » doivent être intégrés par les webmasters dans les pages et les services surveillent si l'un des visiteurs est le robot d'un moteur de recherche.

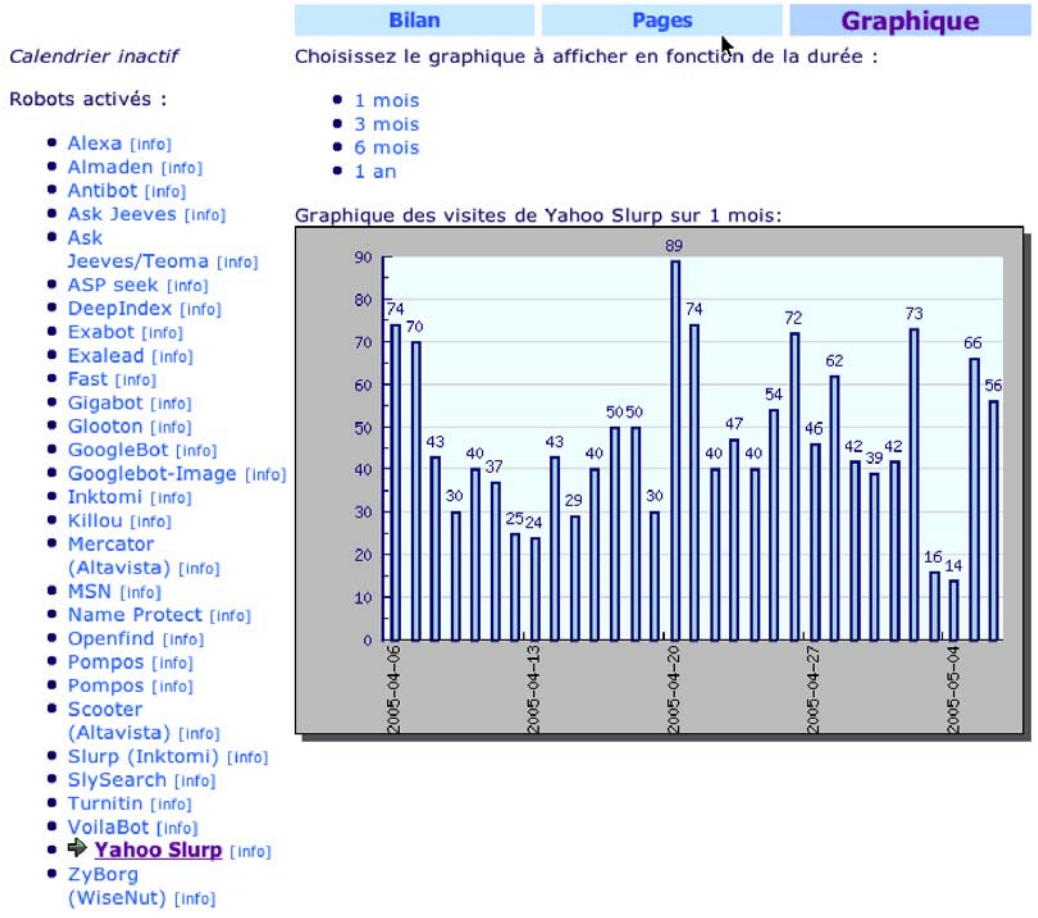

Figure 2-5

Exemple de statistiques fournies par un utilitaire de statistiques en ligne (ici Robotstat, utilitaire français : http://www.robotstats.com/)

Le moteur d'indexation

Une fois les pages du Web crawlées, le spider envoie au moteur d'indexation les informations collectées. Historiquement, plusieurs systèmes d'indexation des données ont été utilisés :

- Indexation des seules balises meta (*meta tags*) insérées par les webmasters dans le code source des pages HTML. Ces balises qui comprennent, entre autres, le résumé et les mots-clés attribués par l'auteur à la page. Très peu de moteurs fonctionnent encore ainsi aujourd'hui.

- Indexation des titres (informations qui sont de moins en moins utilisées car les titres des documents ne reflètent pas toujours le contenu de la page) ou du début du code des documents. Là aussi, ce mode de fonctionnement est devenu très rare.
- Indexation en texte intégral (c'est de loin le cas le plus fréquent). Tous les mots d'une page, et plus globalement son code HTML, sont alors indexés.

Le plus souvent donc, les systèmes d'indexation se chargent d'identifier en « plein texte » (voir encadré ci-dessous) l'ensemble des mots des textes contenus dans les pages ainsi que leur position. Certains moteurs peuvent cependant limiter leur capacité d'indexation. Ainsi, pendant de longues années, Google s'est limité aux 101 premiers kilo-octets des pages (ce qui représentait cependant une taille assez conséquente). Cette limite ne semble aujourd'hui plus d'actualité. Microsoft Live Search, pour sa part, semblait se limiter aux 150 premiers kilo-octets des pages au moment où ces lignes étaient écrites. D'autres moteurs peuvent effectuer une sélection en fonction des formats de document (Excel, Powerpoint, PDF…).

Enfin, sachez que, comme pour les logiciels documentaires et les bases de données, une liste de mots « vides » (par exemple, « le », « la », « les », « et »…), appelés *stop words* en anglais, est le plus souvent automatiquement exclue (pour économiser de l'espace de stockage) ou ces mots sont systématiquement éliminés à l'occasion d'une requête (pour améliorer la rapidité des recherches).

Le traitement des stop words par les moteurs de recherche

On a souvent tendance à dire que la plupart des moteurs de recherche ignorent les *stop words* tels que, en anglais : « the », « a », « of », etc., ou en français : « le », « la », « un », « de », « et », etc. Ceci est exact, comme le montre l'explication de Google dans son aide en ligne :

Google ignore les chaînes de caractères dont le poids sémantique est trop faible (également désignés par « mots vides » ou « bruit ») : « le », « la », « les », « du », « avec », « vous », etc., mais aussi des mots spécialisés tels que « http » et « .com » ainsi que les lettres/chiffres d'un seul caractère, qui jouent rarement un rôle intéressant dans les recherches et risquent de ralentir notablement le processus.

On pourrait donc logiquement s'attendre à ce qu'une requête sur les expressions « moteur de recherche » et « moteur recherche » donnent les mêmes résultats. Eh bien non... S'il y a un certain recouvrement entre les deux pages de résultats, elles ne sont pas identiques, loin de là. Alors, pourquoi cette différence ?

Cela semble venir du fait que Google tient compte de la proximité des mots entre eux dans son algorithme de pertinence. Par exemple, sur la requête « moteur de recherche », Google ne tient pas compte du « de » mais il se souvient tout de même qu'il existe un mot entre les deux termes. Alors que sur la requête « moteur recherche », les pages qui contiennent ces deux mots l'un à côté de l'autre seront mieux positionnées, toutes choses égales par ailleurs, que celles qui contiennent l'expression « moteur de recherche »...

Pour être plus clair, raisonnons sur un exemple : sur l'expression « franklin roosevelt » (*http://www.google.fr/search?q=franklin+roosevelt*), la majorité des pages identifiées comme répondant à la requête contiennent le nom ainsi orthographié : « Franklin Roosevelt ». Insérons maintenant n'importe quel *stop word* entre les deux termes et lançons la requête « franklin le roosevelt » (*http://www.google.fr/search?q=franklin+le+roosevelt*). Résultat : la plupart des pages contiennent le nom différemment orthographié, sous la forme « Franklin *quelque-chose* Roosevelt ». Google s'est donc souvenu que la requête était sur trois termes, même si le deuxième n'a pas été pris en compte. Et ça change tout au niveau des résultats...

(suite)

Vous voulez une autre démonstration ? Tapez la requête « franklin * roosevelt » (*http://www.google.fr/search?q=franklin+*+roosevelt*) et vous obtiendrez quasiment la même réponse que pour « franklin le roosevelt ». Rappelons que l'astérisque, sur Google, permet de remplacer un mot, quel qu'il soit, dans une requête. Là encore, le moteur s'est souvenu que la requête s'effectuait sur trois termes, le premier et le dernier seulement étant pris en compte...

Comment faire, alors, pour que Google prenne en compte le *stop word* s'il vous semble important pour votre recherche ? Il existe deux façons de le faire : soit avec les guillemets, soit avec le signe +.

- Les guillemets vont vous permettre d'effectuer la requête « moteur de recherche » (*http://www.google.com/search?q=%22moteur+de+recherche%22*), les trois mots dans cet ordre et les uns à côté des autres. Dans ce cas, Google prend bien en compte le mot vide dans son algorithme.
- Le signe + vous permet, à l'aide de la requête « *moteur +de recherche* » (*http://www.google.com/search?q=moteur+%2Bde+recherche*), d'inclure de façon obligatoire le mot vide dans la recherche. En revanche, par rapport à l'exemple ci-dessus (avec les guillemets), les mots ne seront pas obligatoirement dans cet ordre et les uns à côté des autres dans les résultats. L'utilisation des guillemets permet donc d'obtenir des résultats plus précis...

L'index inversé

Au fur et à mesure de l'indexation et de l'analyse du contenu des pages web, un index des mots rencontrés est automatiquement enrichi. Cet index est constitué :

- d'un index principal ou maître, contenant l'ensemble du *corpus* de données capturé par le spider (URL et/ou document…) ;
- de fichiers inverses ou index inversés, créés autour de l'index principal et contenant tous les termes d'accès (mots-clés) associés aux URL exactes des documents contenant ces termes sur le Web.

L'objectif des fichiers inverses est simple. Il s'agit d'espaces où sont répertoriés les différents termes rencontrés, chaque terme étant associé à toutes les pages où il figure. La recherche des documents dans lesquels ils sont présents s'en trouve ainsi fortement accélérée.

Pour comprendre le fonctionnement d'un index inversé, prenons, par exemple, une page A (disponible à l'adresse *http://www.sanglots.com/*) comprenant la phrase « Les sanglots longs des violons de l'automne » et une page B (*http://www.violons.com/*) contenant les mots « Les violons virtuoses : les premiers violons du Philharmonique de Radio France ».

Figure 2-6

Deux pages prêtes à être indexées par un moteur de recherche