Un nouveau moteur de recherche chez Free France

Si vous êtes un webmaster consciencieux, vous avez peut-être vu passer dernièrement dans vos logs (la mémoire des connexions de votre serveur) un robot surnommé "Pompos". Une rapide enquête sur le Web amène à cette page :
http://pompos.iliad.fr/
qui présente "Pompos" comme le robot d'un nouveau moteur de recherche francophone. La domiciliation de l'adresse chez Iliad (propriétaire de Free) est étonnante. Free plancherait-il sur un nouvel outil de recherche ?

C'est ce que nous avons demandé aux concepteurs et chefs de projet Fabien Menemenlis (à droite sur la photo) et Philippe Develter (à gauche donc), effectivement salariés d'Iliad. Ils ne sont pas des débutants dans le domaine des bases de données et de la recherche puisque Fabien Menemenlis a développé le site http://www.societe.com/ et Philippe Develter le service Minitel 36 17 ANNU et le site http://www.annu.com/. Fabien est plutôt spécialisé sur la partie "moteur" et Philippe sur la partie "crawl".

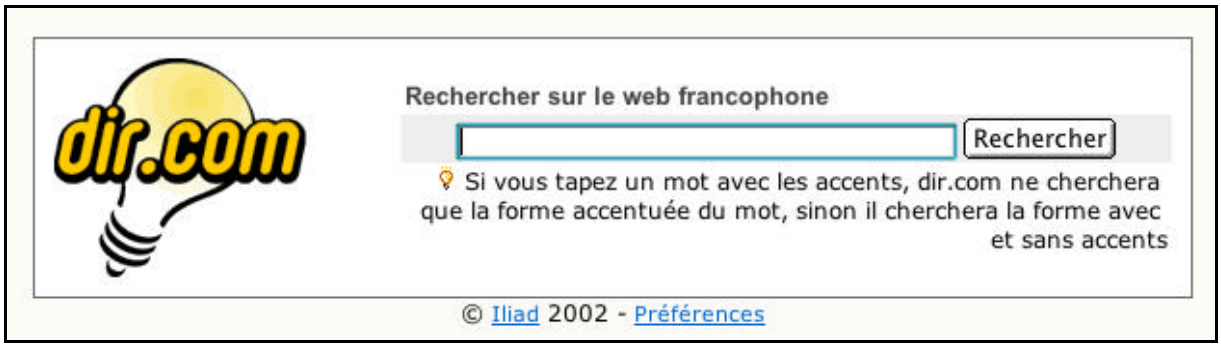
Pompos fait en fait partie d'un projet mené en interne chez Iliad pour développer un moteur de recherche dans une optique strictement francophone au départ. Le site ouvrira à l'url http://www.dir.com/ qui pointe actuellement sur le site annu.com.
Date de lancement prévue, voire espérée ;-) : fin janvier 2002.

Formats indexés

Ce moteur crawle actuellement tous les liens html (pages dynamiques comprises), les documents PDF, PostScript, Word .doc et .rtf, Powerpoint .ppt, Flash .swf et Excel .xls. D'autres formats seront envisagés par la suite.
Dans un premier temps, l'index du moteur sera remis à jour et reconstruit toutes les semaines.
Notons, donc, que Dir.com sera, dès son lancement, le deuxième moteur de recherche, après Fast, à indexer le format Flash de Macromedia.
L'index comportera dans sa version de lancement 30 millions de documents francophones.

Page d'accueil

Voici à quoi devrait ressembler la page d'accueil du moteur :



Un design très simple, "à la Google", et un outil qui ne fait QUE de la recherche d'information de type "moteur de recherche". Slogan actuel : "Quand on le cherche, il trouve !" :-))

Syntaxe d'interrogation

La syntaxe d'interrogation ressemble à ce qui est proposé sur un certain nombre de moteurs à l'heure actuelle :

- ET : signe +. Taper :
+moteur +recherche
pour la requête "moteur ET recherche"
Mais la requête :
moteur recherche
est équivalente (l'espace est considéré par défaut comme un ET)

- OU : pas de possibilités (les développeurs de l'outil estiment que cette syntaxe est trop peu utilisée par les internautes pour qu'elle soit implémentée)

- SAUF : point d'exclamation. Taper :
moteur !recherche
pour le requête "moteur SAUF recherche"

- Recherche sur un site web donné : site:. Taper :
+free +site:www.abondance.com
pour rechercher les pages du site Abondance contenant le mot "free".
ou :
site:www.abondance.com
pour avoir toutes les pages du site qui sont indexées par le moteur.

- Recherche sur les liens pointant vers un site web donné : link:. Taper :
link:www.abondance.com
pour rechercher les pages qui contiennent un lien vers le site Abondance.

- Les requêtes sont limitées dans un premier temps à 10 termes distincts à la suite.

Pages de résultat

Le "look" des pages de résultat est le suivant :

Sont affichées les informations suivantes :

- Titre (complet, pas de limitations de taille).
- Résumé en gris : contenu de la balise Meta Description si elle existe (250 caractères max).
- Résumé en noir : "snippet" : extrait du texte de la page contenant le mot demandé (250 caractères max).
- Url de la page. Bizarrement, celle-ci est affichée avec le numéro de port ("80"). Cela sera corrigé dans la version finale.
- Lien vers la fonction [Cache], similaire à celle de Google : Dir.com sera donc le deuxième moteur, après Google et à notre connaissance, à proposer ce type de fonctionnalité.
- Le clustering (un lien par site web) étant activé, un lien vers toutes les pages du site en question est également proposé.

Indexation

Nous avons vu, en début d'articles, que Dir.com indexera de nombreux formats. Pas de réel obstacle technique à l'indexation de son site par le moteur, puisque ce dernier prendra en compte les pages dynamiques (ASP, CFM, PHP, etc.) ainsi que le Flash, etc.

Dans une page, voici les champs qui seront indexés ou non :

- Titre (balise <Title>) : Oui
- Balise Meta "Description" : Oui
- Balise Meta "Keywords" : Non
- Texte visible : Oui, tout le contenu de la page est indexé, quelle que soit sa longueur.

Prise en compte du fichier Robots.txt : Oui
Pris en compte de la balise Meta "Robots" : Oui

Critères de pertinence

Certains champs ou critères de pertinence sont, bien entendu, plus importants pour le moteur Dir.com :

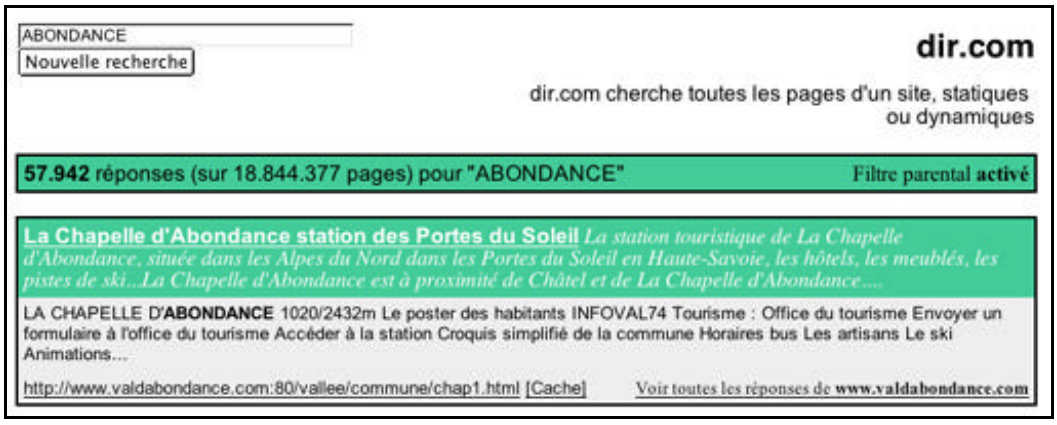
- Champ <Title>
- Indice de popularté, calculé à 2 niveaux (quantitatif et qualitatif), à la manière de Google.
- Début du texte visible. Un mot en haut de page est plus important que le même terme en bas de page.
- La mise en exergue de certains mots du texte (taille de la police, gras, texte des liens, etc.)

Le moteur donne d'abord les pages où les mots recherchés sont côte-à-côte dans la page pour donner ensuite celles où les mots sont un peu plus éloignés pour aller jusqu'aux plus éloignés. Imaginons une recherche sur "LECTEUR DVD" . Les premières réponses comporteront d'abord la chaine "LECTEUR DVD" puis peut-être "LECTEUR POUR LIRE UN DVD" puis enfin "UN LECTEUR NOUS A SIGNALÉ QU'IL NE POUVAIT LIRE UN DVD", etc. Selon les deux chefs de projets, peu de moteurs le font car l'indexation de toutes les positions de mots est coûteuse en ressources (Google le fait cependant).

Préférences

Des préférences d'affichage des résultats seront proposées sur la version finale du site. Pour l'instant, seuls deux choix sont proposés :

Le filtre parental permet d'exclure certaines pages "extrêmes" et le site propose 3 présentations, 3 "looks" différents : "Standard", "Vert" ou "Bleu", ce qui change la couleur des pages de résultats. Voici le look "vert" :



Thématisation

L'une des principales caractéristiques, et l'une des plus intéressantes, du site Dir.com, sera la catégorisation automatique des pages web (encore en développement au moment de nos tests) : chaque page crawlée passera dans une "moulinette" qui octroiera, lorsque cela sera possible, un thème donné au contenu identifié. Cela permettra de fournir des résultats beaucoup plus fins et

précis à l'internaute, qui pourra ainsi orienter sa recherche sur telle ou telle thématique parmi celles proposées. Certains moteurs (Voila, Inktomi) s'étaient déjà lancés dans ce type de fontionnalités. Nul doute que nous suivrons de façon assidue les travaux des concepteurs de Dir.com dans ce domaine et surtout les fonctions et possibilités qu'ils vont en tirer pour en faire profiter les internautes...

Conclusion

Pour l'instant, le moteur "Dir.com" est considéré par Iliad comme un projet de R&D. Le moteur ne tourne aujourd'hui "que" sur 5 machines, ce qui serait passablement insuffisant pour un moteur de recherche en phase de production avancée. Mais il ne s'agit encore que d'une phase de test....

Quelques innovations ou originalités intéressante sont à mettre à son crédit :

- Indexation de nombreux formats de documents dont le Flash.
- Réel effort d'indexation des pages dynamiques.
- Fonction [Cache].
- Thématisation automatique des pages web.

Côté pertinence, les premiers essais que nous avons pu faire sont encourageants. Sur les mots classiques que nous utilisons pour tester un outil de recherche, les sites incontournables du domaine étaient toujours présents dans les 10 premiers liens. Pour un site de test, deux mois avant son lancement, bravo ! L'essai est prometteur, attendons de voir maintenant comment il sera transformé !

Il n'est cependant pas dit que ce moteur remplace l'outil de recherche (actuellement Google) sur le portail Free.fr. L'idée est plutôt de lancer l'outil dans un premier temps, sans réel modèle économique, puis de voir ce que cela donne. S'il est performant, il n'est pas dit que, d'une part il n'aille pas titiller les "gros" moteurs mondiaux avec un index anglophone et que, d'autre part, il s'installe sur la "home" de Free... Cela, seul l'avenir le dira...

Ceci dit, nous aurons certainement le plaisir, fin janvier prochain (si tout va bien ;-)), le plaisir de tester "en direct live" un nouvel arrivant sur le marché des moteurs de recherche francophones. Bienvenue à eux !!

Nota : toutes les informations ci-dessus ont été relues et validées par les concepteurs du moteur Dir.com. Merci à eux pour le temps qu'ils ont bien voulu nous consacrer !