A quoi ressemblera le "Google Next Generation" ?

Vous l'avez peut-être remarqué, le moteur de recherche Google est depuis quelque temps sur la sellette. D'une part, il connaît une résurgence du spam dans ses pages de résultats (dernièrement, sur le forum Abondance consacré au référencement, un visiteur proposait de taper la requête "Alcatel Hopital Espagnol" pour s'apercevoir que les liens proposés n'avaient pas grand chose à voir avec la demande initiale et faisaient la part belle aux sites X... De nombreux observateurs du Web disent également avoir constaté une baisse globale de la pertinence du moteur de recherche, pertinence qui avait fait le succès de l'outil.

Enfin, Google voit se profiler à l'horizon deux gros concurrents : Microsoft qui développe pour bientôt (2004 ? 2005?) son moteur basé sur une technologie "maison" et Yahoo! qui est en train d'effectuer l'amalgame entre Overture, AltaVista, Inktomi et AllTheWeb, tous rachetés dernièrement.

Bref, il semble urgent pour Google de réagir. Et, connaissant les loustics (pardon ;-)), il y a de fortes chances qu'ils le fassent dans les mois qui viennent pour proposer un nouvel outil de recherche entièrement remis à neuf avec de nouvelles spécificités. Quand ? Peut-être pas tout de suite. Car Google pourrait caler sa stratégie sur celle de ses concurrents. En effet, il semble peu probable que Yahoo!, de son côté, propose dans des délais très brefs un nouvel outil de recherche "révolutionnaire". Son prochain moteur web devrait être, si notre intuition est bonne, fortement basé sur la technologie Fast / AllTheWeb. Si cette dernière a largement prouvé son efficacité dans les mois qui viennent de s'écouler, elle ne constitue pas pour autant une rupture technologique forte par rapport à ce que propose aujourd'hui un moteur comme Google. Yahoo! devant aujourd'hui résoudre les soucis - pas si évidents - d'intégration de plusieurs technologies très différentes à la base, il n'est pas sûr que son souci principal soit de révolutionner le monde du "search" mais plutôt d'obtenir un moteur au moins aussi pertinent que le Google actuel, ce qui serait déjà une bonne chose et une première étape pour lui. Il n'est pas dit qu'il ne lance pas un outil de recherche réellement innovant à long terme, mais il ne semble pas que ce projet soit promis à une courte échéance.

Il n'en n'est pas de même de Microsoft. Partant de zéro (ou quasiment), la firme de Redmond peut réeellement créer un outil révolutionnaire, basé sur de nouvelles idées et naviguant sur des voies différentes de celles empruntées par Google. Bref, surprendre avec un nouvel outil réellement original pour prendre d'importantes parts de marché au leader actuel. Il faut bien avouer que si Microsoft sortait un moteur de recherche "classique", du style"Google-like", tout le monde serait bien déçu et le projet de la firme de Redmond perdrait beaucoup de son intérêt (même si le fait d'intégrer le moteur à Windows et Explorer reste un atout de poids pour le projet de Microsoft).

Pas de nouveau Google avant 2005 ?

Or, il n'est pas sûr que Microsoft lance son nouveau moteur en 2004. On entend parfois parler, dans les couloirs ;-), de 2005. Dans ce cas, Google ferait certainement une erreur stratégique en lançant un nouvel outil trop tôt, laissant ainsi à Microsoft le temps de se ressaisir et de le contrer plus facilement. L'idéal pour Google serait certainement de lancer son nouveau moteur deux à trois mois, par exemple, avant celui de Microsoft. Résultat : un effet d'annonce très fort pour Google et un délai trop court pour Microsoft pour se retourner et partir sur d'autres pistes de réflexion... Mais Google peut-il se permettre d'attendre 2005 pour lancer un nouvel outil, au vu de la baisse de pertinence actuelle de ses résultats et de la concurrence proche de Yahoo! ? Difficile équation à résoudre pour le leader actuel.. Notons cependant que tout cela n'est que supposition, l'avenir nous dira certainement de quoi il en retourne.

Cependant, il nous semble important, à la lueur des "rumeurs" qui circulent et des tendances actuelles de nombreux moteurs de recherche et acteurs du monde du "search", de voir ce que peut proposer Google dans les prochains mois. A quoi ressemblera le "Google Next Generation" ? Nous avons tenté, dans les pages suivantes, un exercice prospectif (indiquons bien qu'il s'agit d'un pur exercice de type "cas d'école", sans qu'aucune information précise ne nous ait été fournie par Google lui-même) pour tenter de répondre à cette question. Voici donc quelques pistes de réflexion pour essayer de bâtir le "moteur de recherche" du futur...

Sur la base des quelques pistes évoquées ci-dessous, à vous de vous faire votre propre idée de ce à quoi resssemblera Google dans quelques mois...

Géolocalisation : une fonction très "tendance"

Tout d'abord, la fonctionnalité certainement la plus "à la mode" est celle de géolocalisation de l'internaute. Le but est de savoir où se trouve l'utilisateur du moteur, géographiquement parlant (pays, région, ville, etc.), afin de lui fournir une information la plus proche possible de ses attentes. Bien entendu, cette fonction est également intéressante dans le domaine de la publicité. Si un internaute tape le terme "menuisier", l'idéal sera de lui proposer, notamment, des liens sponsorisés d'artisans et de sociétés qui sont dans son pays, et, mieux, dans sa région, voire sa ville... Au vu du "boom" actuel des liens sponsorisés, nul doute que les fonctions de géolocalisation intéressent tout le monde...

Historiquement, de nombreuses expérimentations et plusieurs outils de recherche ont déjà proposé ce type d'outils. Mirago (http://www.mirago.fr/) propose un moteur de recherche géographique depuis plusieurs mois (voir lettre R&R de mars 2002). DeepIndex avait ensuite lancé le projet GIIPPS (http://www.giipps.com/), proposant un certain nombre de réflexions à ce sujet. Overture a dernièrement ouvert ses laboratoires de recherche avec un projet de localisation (http://localdemo.overture.com/), suivi par le Google Labs (http://labs.google.com/location/), nous permettant de mieux visualiser, de façon concrête, ce que ce type d'application peut proposer comme amélioration dans le cadre de la recherche d'information sur le Web. Overture vient d'ailleurs d'implanter une fonction "géolocalisation" en test sur AltaVista, pour certains internautes uniquement (10% des connectés voient s'afficher des résultats "géolocalisés" dans la page de résultats du moteur).

Le moteur Gigablast (http://www.gigablast.com/) annonçait également, il y a peu de temps, les "geo-sensitive metatags". Il s'agit de balises Meta spécifiquement étudiées pour la géolocalisation, indiquant plusieurs données géographiques. Exemple :

```
<meta name="zipcode" content="87112,87113,87114">
<meta name="city" content="albuquerque, abq, rio rancho">
<meta name="state" content="new mexico">
<meta name="country" content="usa, united states of america">
<meta name="author"   content="john doe">
<meta name="language" content="english">
```

On le voit, la géolocalisation est une technique très "tendance". Pour arriver à localiser l'internaute (pour lui proposer des résultats qui lui conviennent parfaitement) ou un site web (pour le proposer à l'internaute local s'il propose des caractéristiques locales), de nombreuses possibilités sont offertes sur le Web :
- Analyse des adresses IP (une adresse IP peut être localisée géographiquement dans de nombreux cas).
- Analyse de la langue du navigateur de l'internaute.
- Analyse des pages web pour y trouver une adresse postale.
- Déclaration spontanée des utilisateurs (Deepindex : http://www.geo.deepindex.com/, Google : http://services.google.com/georeport/).
- Utilisation et analyse de services connexes (mail, formulaires, création de sites perso, concours, etc.), nécessitant une identification de l'internaute.
- Traitement humain par une équipe éditoriale (par exemple, Yahoo! ajoute à chaque site indexé dans son annuaire ses coordonnées postales).
- Etc.

Il y a donc de très fortes chances que la géolocalisation soit au centre des réflexions de Google à l'heure actuelle. D'ailleurs, les AdWords sur le site www.google.com ne sont visibles que pour les internautes américains. Il en est de même de la plupart des liens sponsorisés sur les portails américains. Une première application... avant bien d'autres ?

Personnalisation : Google y travaille...

La personnalisation est l'un des grands axes de recherche actuels de Google, renforcé par l'acquisition de la société Kaltix (http://www.kaltix.com/) il y a peu.
Difficile d'en savoir plus pour l'instant de ce côté, mais le but est globalement de fournir une réponse plus pertinente à l'utilisateur en tentant de mieux le connaître. Pour simplifier, s'il tape le mot clé "moteur" et qu'il a saisi dans les jours précédents des termes comme "Renault" ou "Ferrari", on peut penser qu'il s'intéresse plutôt aux moteurs de voiture. Si son historique est peuplé de "Google" et d'"Altavista", ce sont plutôt les moteurs de recherche qui seront privilégiés.

De nombreux sites, notamment dans le commerce électronique (Amazon) s'y sont essayé dans les années qui viennent de s'écouler, avec plus ou moins de succès. Les moteurs de recherche (Inktomi, AltaVista) n'ont pas été en reste avec des projets qui n'ont pas toujours été couronnés de succès. La complexité avec la personnalisation vient du fait que cette fonctionnalité ne doit en rien être considérée comme une contrainte par l'utilisateur. Bref, elle doit être le plus transparente possible, point que les moteurs de recherche ont eu du mal à intégrer jusqu'à maintenant. Le problème de la personnalisation est également un problème de stockage. Si le moteur doit garder sur ses serveurs bon nombre d'informations pour chacun de ses visiteurs, le coût de stockage peut vite se montrer prohibitif. Autre possibilité à sa disposition : passer par des cookies, mais là aussi, la taille des fichiers cookies est limitée. Pas si facile. Il semblerait cependant que certaines sociétés (notamment françaises) réfléchissent à d'autre voies que les cookies pour stocker des infos de personnalisation sur le disque dur de l'utilisateur. Mais chut, plus d'infos très bientôt..; ;-)
En tout cas, là aussi, il s'agit clairement d'une idée "dans l'air du temps"...

L'acquisition de Kaltix amènera-t-elle Google à proposer un véritable outil de personnalisation des recherches sur le web ? Pour l'instant, on sait peu de choses, Kaltix n'a que très peu communiqué sur ses projets, hormis le fait qu'il est issu du "Stanford Webbase Project" (http://www-diglib.stanford.edu/~testbed/doc2/WebBase/) et que ses créateurs ont déjà travaillé, à l'époque, sur l'algorithme PageRank de Google. Difficile de savoir, en l'état actuel des choses, où le méneront ses réflexions. Mais il est clair, comme l'a dit dernièrement Eric Schmidt, "chief executive officer" de Google, sur le site Cnet (http://news.com.com/2100-1024-5088153.html), que la personnalisation sera au coeur de la prochaine version de Google. Si c'est lui qui le dit...

Contextualisation

La contextualisation des recherches semble également être au menu des laboratoires de Google, qui a déposé un brevet dans ce sens, intitulé : "Ranking search results by reranking the results based on local inter-connectivity" (http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=/netahtml/search-bool.html&r=3&f=G&l=50&co1=AND&d=ptxt&s1=google.ASNM.&OS=AN/google&RS=AN/google).
Le but est de chercher dans l'index du moteur, non pas de façon globale, mais uniquement sur un "corpus" de documents traitant de la thématique recherchée. Par exemple, si vous tapez "ferrari", l'ambition des systèmes de contextualisation est de lancer la requête non pas sur les 3 milliards de documents de l'index mais uniquement sur l'ensemble des pages de cet index ayant trait aux automobiles... La pertinence de la requête doit logiquement être bien plus forte dans ce cas (on évitera ainsi, par exemple, les pages qui parlent de... Lolo Ferrari).

Les corpus thématiques peuvent s'évaluer en analysant les liens établis entre les sites du Web. C'est ce que fait, par exemple, Teoma (http://www.teoma.com/) qui, pour une requête (exemple : "insurance") propose dans sa rubrique "Resources" une liste de "sites Hub" incontournables du domaine. C'est également le domaine de Linkkit (http://www.linkkit.com/), technologie française qui a fait l'objet d'un article sur cette lettre en avril 2003.

Bien sûr, la contextualisation peut tout à fait être couplée aux techniques de personnalisation pour effectuer des requêtes uniquement dans les thèmes habituellement recherchés par l'internaute. La conjonction des deux idées peut également mener vers des systèmes de veille, apportant l'information de façon régulière à l'internaute, par exemple par mail, sans que ce dernier fasse une requête, l'outil ayant "compris" les thèmes d'intérêt de l'internaute. Quand le "push" complète le "pull", donc... D'autre part, il semblerait que la contextualisation des requêtes soit une bonne solution pour lutter contre le spam des index des moteurs.

Statistiques / Sémantique : déjà en cours...

Googe n'a jamais caché que l'une de ses quêtes était de "mieux comprendre le Web". Pour cela, des voies prenant en compte des algorithmes sémantiques et statistiques peuvent être considérées comme pertinentes. En France, des sociétés comme Sinequa (http://www.sinequa.com/) et Exalead (http://www.exalead.com/), l'une des technologies de "web search" les plus pertinentes sur le Web francophone, ont montré avec succès qu'il s'agissait là de reflexions qui pouvaient mener à d'excellents résultats. De plus, Exalead démontre également que ces technologies peuvent également grandement aider à lutter à la lutte contre le spam...

Google a déjà commencé à intégrer des fonctions sémantiques dans ses outils. La fonction de recherche "tilde" (recherche de synonymes) a été mise en place il y a quelques semaines de cela (http://actu.abondance.com/2003-32/google-synonymes.html). Pour son service de liens sponsorisés AdWords, des termes "connexes" sont aujourd'hui proposés (http://actu.abondance.com/2003-42/adwords.html). De plus, les AdSense (https://www.google.com/adsense/) font appel à la fois à des technologies developpées en interne et à des algorithmes émanant de la société Applied Semantics (http://www.appliedsemantics.com/), rachetée il y a peu par Google (http://actu.abondance.com/2003-17/applied-semantics.html). Google avait également racheté la société Outride (http://www.google.com/press/pressrel/outride.html), émanation du Xerox Palo Alto Research Center, en 2001. Cette société était également spécialisée dans l'approche sémantique du traitement des données et dans le Datamining.

Un faisceau de faits et de présomptions qui semble indiquer que les aspects sémantiques et statistiques sont bien des projets dans les "cartons" de Google à moyen terme...

Lutte contre le spam

Si vous utilisez souvent Google, vous avez sûrement remarqué que les pages de résultats de Google sont énormément spammées, certainement beaucoup plus qu'avant. Certains sites trustent les premières positions avec des pages n'ayant aucun rapport avec la requête initiale (voir l'article "Premières dérives pour les liens sponsorisés ?" dans notre lettre du mois dernier).

Google n'a aujourd'hui pas d'autre solution que de lutter fortement contre ce spam qui pollue ses résultats. Deux solutions donc :

- Soit les nouvelles mesures (géolocalisation, personnalisation, contextualisation, analyse statistique et sémantique ou un mix de certaines d'entre elles) suffisent pour combattre ce spam et éjecter des résultats les pages des spammeurs actuels, de façon "naturelle".

- Soit Google met en place une politique plus restrictive vis à vis du spam, ce qu'il ne semble pas avoir fait jusqu'à maintenant, certainement parce que le système du PageRank le mettait à l'abri de ce type de désagréments, ce qui n'est plus le cas aujourd'hui, certains webmasters peu scrupuleux ayant trouvé le moyen de "dérouter" l'index du moteur.

Comme le PageRank a permis à Google de rendre, en son temps, son index peu perméable au spam, on peut parier que de nouvelles fonctionnalités pourraient renverser la situation actuelle et "épurer" à nouveau les pages de résultats de Google. L'expérience d'Exalead et de Linkkit, notamment, montrent que les nouvelles voies de réflexions évoquées dans cet article peuvent suffire à lutter contre ce fléau.

Bref, bien que tout ce nous avons écrit dans cet article relève de la plus pure extrapolation et d'une vision qui n'appartient qu'à nous, on peut raisonnablement penser que les principales voies de réflexion actuelles de Google sont évoquées dans les pages précédentes. Seront-elles toutes prises en compte ? Quand la nouvelle génération du moteur verra-telle le jour ? Pour l'instant, ces questions restent malheureusement sans réponse...