Plus d'infos sur le moteur de recherche Ask.com

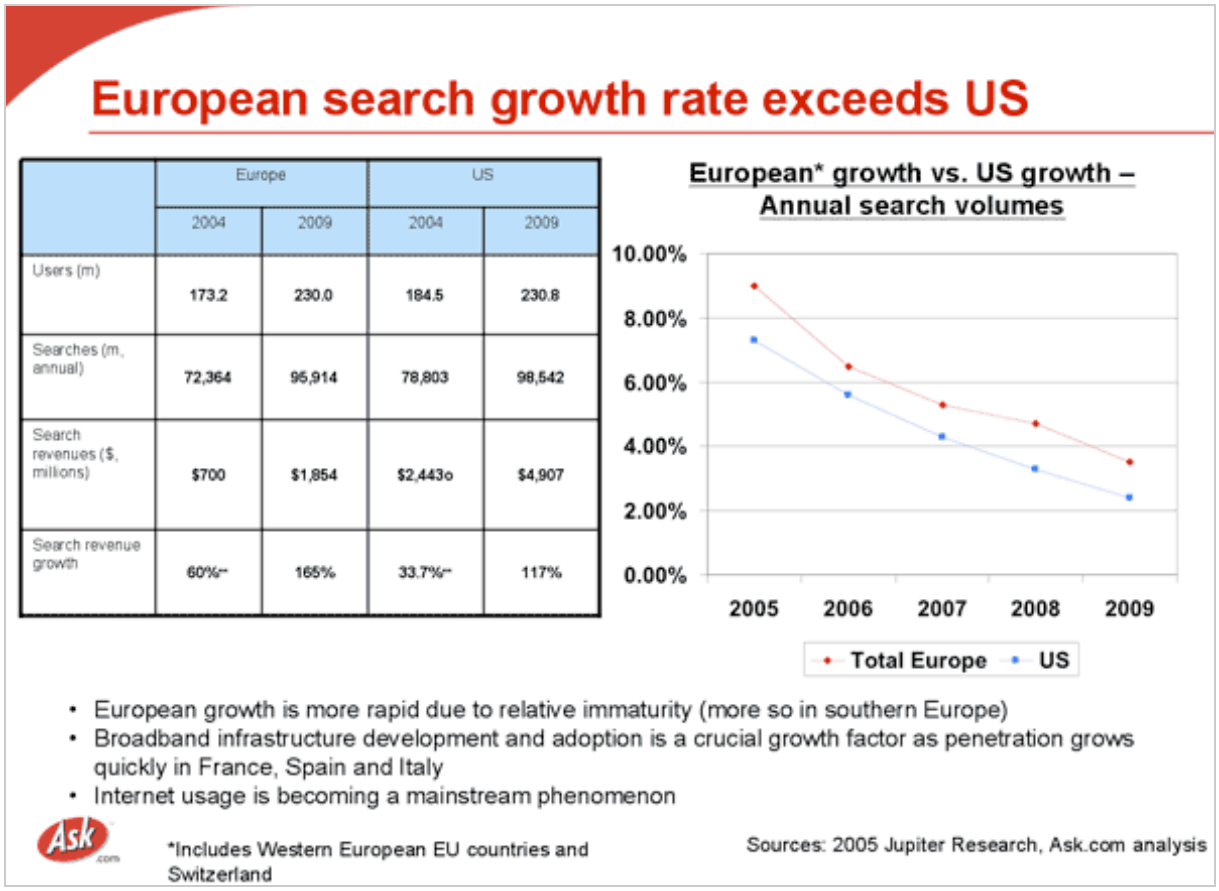
Retour au sommaire de la lettre

Ask.com est l'un des principaux moteurs de recherche aux Etats-Unis et en Grande-Bretagne; Il a ouvert un site en version bêta en France il y a quelques semaines de cela. A l'occasion d'un voyage au centre européen de R&D du moteur à Pise en mai dernier, nous avons pu glaner un certain nombre d'informations sur le mode de fonctionnement de Ask.com, moteur dont la vision se démarque de celle de Google, notamment au niveau de la notion de "popularité". Explications...

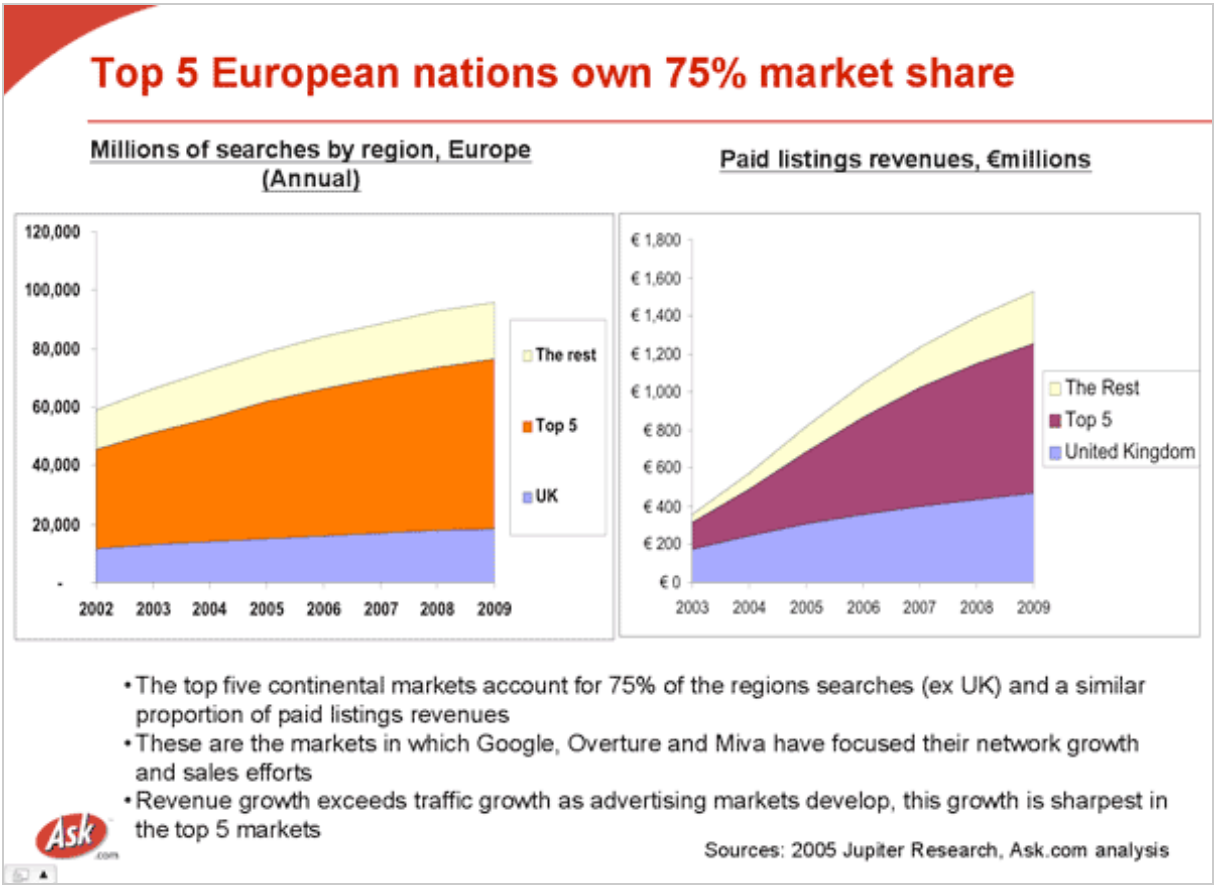
Au mois de mai 2006, nous avons eu l'occasion d'aller à Pise (Italie) visiter le centre de Recherche et Développement du moteur de recherche Ask.com et de rencontrer notamment Antonio Gulli, Directeur *Advanced Projects* de Ask et responsable R&D du moteur de recherche en Europe. C'était une bonne occasion de mieux connaître ce nouvel outil qui a débarqué en France il y a peu et de faire un point sur son algorithme de pertinence qui se démarque nettement de celui de Google, notamment en ce qui concerne la notion d'*ExpertRank* pour le calcul de la popularité des pages...

Le marché européen des moteurs de recherche

Dans un premier temps, lors de notre périple à Pise, Miguel Acosta, *VP Business Development* Europe, nous a présenté de façon plus large le marché des moteurs de recherche en Europe et aux Etats-Unis. Parmi les slides présentés, deux nous ont plus particulièrement intéressés :



Selon ces données :
- Le nombre de requêtes effectuées sur les moteurs de recherche est proche aux Etats-Unis et en Europe - entre 72 et 78 milliards par an - et leur taux de croissance est encore très important d'ici 2009.
- Le marché des revenus du SEM (*Search Engine Marketing*) - notamment les liens sponsorisés - est encore aujourd'hui beaucoup plus fort aux Etats-Unis.
- La croissance de l'Europe est plus forte actuellement que celle des Etats-Unis.



Les 5 pays les plus importants en Europe (hors Grande-Bretagne) sont la France, l'Allemagne, l'Italie, l'Espagne et les Pays-Bas qui représentent 75% des parts de marché du continent.

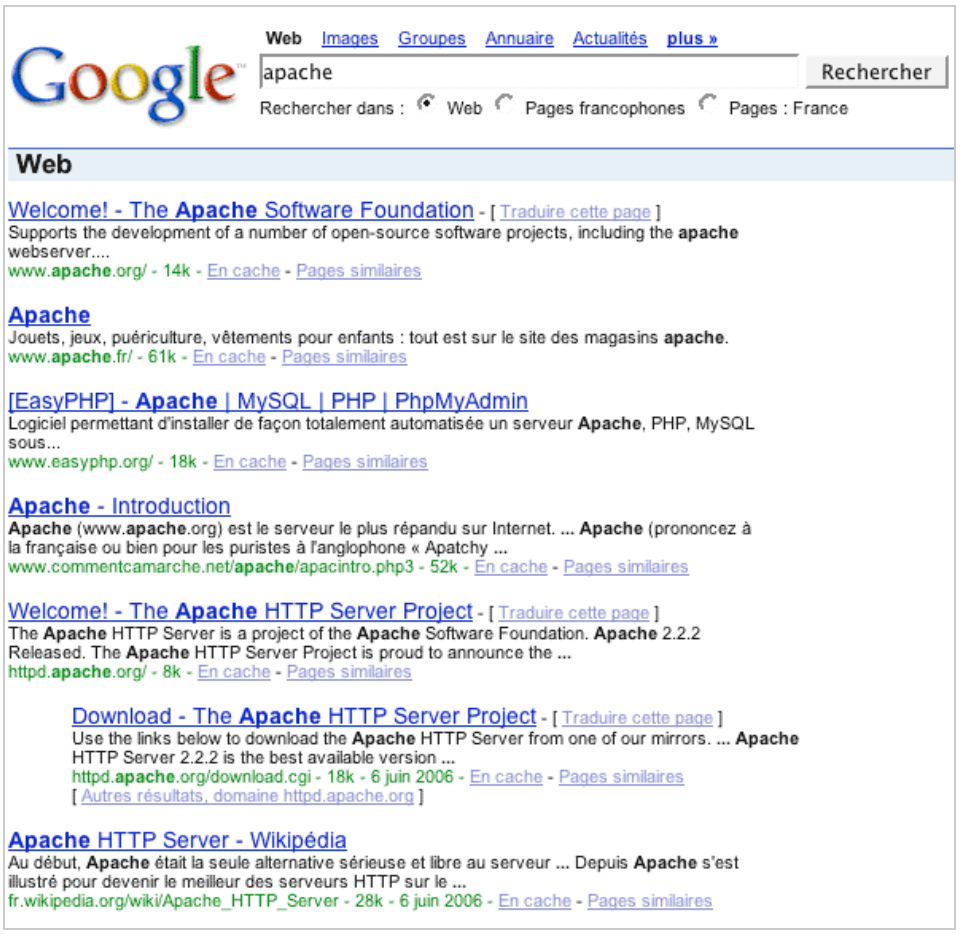
Notons enfin qu'aux Etats-Unis et en Grande-Bretagne, Ask.com se positionne dans le "Top 5" des outils de recherche utilisés par les internautes locaux, avec Google, Yahoo!, MSN et AOL (Sources : http://actu.abondance.com/2006-22/trafic-us.php, http://actu.abondance.com/2006-13/moteurs-uk.php, http://actu.abondance.com/2005-09/barometre-uk.php). Arrivera-t-il au même résultats en France ? C'est ce que l'avenir nous dira...

L'algorithme de pertinence de Ask.com

Nous avons pu, lors de notre présence à Pise, discuter avec Antonio Gulli (*voir photo ci-contre*), créateur en 1998 d'Arianna, premier moteur de recherche italien, et qui est aujourd'hui chargé de coordonner les travaux de recherche du moteur en Europe. Une bonne occasion de tenter de comprendre comment fonctionne Ask.com...

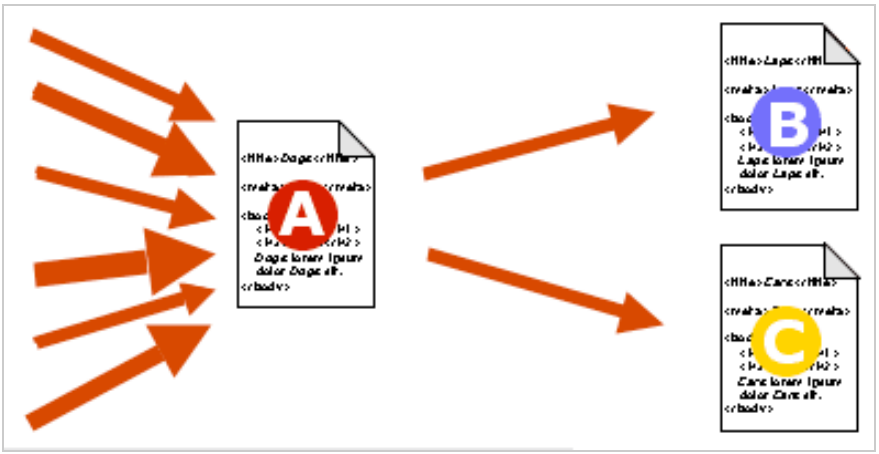
Ce moteur de recherche est basé sur la technologie de Teoma, rachetée en 2001 par Ask.com. Selon Antonio Gulli, cet algorithme est avant basé sur la notion de diversité...

En effet, de nombreux termes, dans toutes les langues, ont plusieurs sens. Par exemple, une requête sur le mot clé "Apache" désigne les serveurs web bien connus, mais également une tribu d'indiens, un avion, un hélicoptère, etc. Or, selon Antonio, si vous tapez ce mot clé sur Google, la requête apache vous donnera dans son immense majorité des liens vers les systèmes de serveurs web :



Il en sera de même pour des mots clés comme python, qui présentera de façon large des informations sur le langage de programmation du même nom en laissant un peu de côté le serpent, etc.

Pourquoi ? Parce que Google tient compte de la popularité des pages avec son système de PageRank **de façon générale**. En d'autres termes, pour calculer la popularité d'une page web, Google recherche dans son index global, toutes les pages qui ont mis en place un lien vers elle et va analyser ces liens.



Dans le schéma ci-dessus, A va "voter pour" (i.e. "faire un lien vers") B et C. Plus la popularité de A sera forte, plus ce vote augmentera la popularité de B et de C. Ce système itératif permet de prendre en compte deux aspects importants de l'interconnexion des pages web : la **quantité** des liens et leur **qualité**.

Nota : pour mieux comprendre le fonctionnement du PageRank, nous vous recommandons de (re)lire, entre autres, les articles suivants :
- *Un point sur les brevets : le PageRank de Google* (http://abonnes.abondance.com/archives/2003-04/pagerank-brevets.pdf)
- *Comment est calculé l'Indice de Popularité sur les moteurs de recherche* (http://abonnes.abondance.com/archives/2002-02/index.html)
- *Comment optimiser une page web pour les moteurs de recherche (4ème partie)* (http://abonnes.abondance.com/archives/2004-02/opti-liens.pdf)

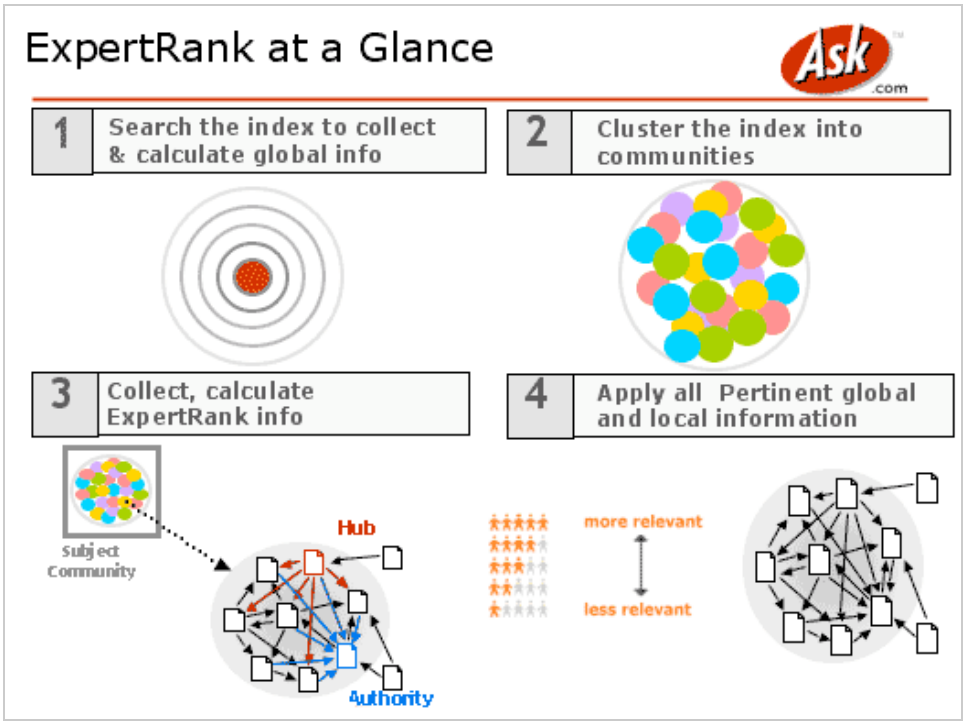
Mais la "faiblesse" du PageRank, selon Antonio Gulli, est que ce calcul s'effectue sur tout l'index du moteur. Donc, si la "communauté" des programmeurs en langage Python est plus grande, plus forte sur le Web que celle qui apprécie les serpents du même nom, la première va "phagocyter" la seconde qui apparaîtra très peu dans les résultats de recherche... Bref, la parole est plutôt, avec le PageRank, aux communautés les plus "visibles", les plus "présentes" sur le web.

L'idée de la technologie Teoma, qui prévaut aujourd'hui sur Ask.com, est de définir au préalable, lors du calcul de la popularité, quelles sont les communautés sur le Web dans différents domaines, puis d'appliquer le calcul de la popularité uniquement sur les pages présentes à l'intérieur de chaque communauté. Cela donne le schéma suivant, par exemple pour le mot clé "apache" :

1. Identification préalable des différentes communautés représentées par le terme "Apache" : serveurs web / tribu indienne / avion / hélicoptère / etc.
2. Prise en compte et identification des pages web de l'index global à l'intérieur de chacune de ces communautés.
3. Calcul d'un "ExpertRank", notion de popularité d'une page à l'intérieur de sa communauté uniquement.

On peut donc définir la notion de popularté, nommée "ExpertRank" par Ask.com, comme étant un "PageRank communautaire"... C'est ce qui fait la grande différence entre les algorithmes de pertinence de Ask d'un côté et ceux de Google, Yahoo! et MSN de l'autre...

Le schéma de calcul du système ExpertRank peut être décrit au travers de ce schéma, fourni par Ask.com :



Les algorithmes de Kleinberg

En fait, ce système est loin d'être un inconnu pour les experts des moteurs de recherche, puisqu'il est basé sur les algorithmes de Kleinberg, très connus dans le micrososme du "*search*". Ces algorithmes ont notamment été intégrés dans des projets comme HITS ou CLEVER et WEBFOUNTAIN d'IBM.

Jon Kleinberg (*photo ci-contre*) est professeur à l'université de Cornell. Il est l'auteur de nombreux travaux sur l'analyse de l'interconnexion des pages sur le Web, notamment lors d'un passage d'un an au laboratoire de recherche Almaden d'IBM en tant que "visiteur scientifique". Il est l'auteur de l'algorithme HITS (*Hyperlinked Induced Topic Search*) qui est inspiré par des systèmes permettant de "peser" l'intérêt d'une publication scientifique en fonction du nombre de citations que contiennent les autres publications... Vous voyez où nous voulons en venir... :-)

Kleinberg a notamment défini deux types de sites "importants" dans l'évaluation de sliens :

- Les "sites de référence" ou "**Authorities**". Ils contiennent de l'information importante sur un domaine et sont considérés comme des "incontournables" sur une thématiques donnée. Cherchez bien sur les domaines qui vous intéressent, vous en trouverez très rapidement...

- Les "sites de redirection" ou "**Hubs**". Ces sites redirigent en fait l'internaute vers les "sites de référence". Un exemple-type est celui d'un annuaire spécialisé qui va proposer une liste de "sites incontournables" dans un domaine précis...

Les "Hubs" sont donc des sites qui dirigent l'internaute vers les "Authorities" qui, eux, détiennent une info intéressante et le réel contenu. Le "Hub" est donc caractérisé par ses liens sortants vers les "Authorities" qui, eux, sont logiquement caractérisés par leurs liens entrants depuis les "Hubs" et leurs liens internes vers l'information du domaine en question. Le tout à l'intérieur d'une communauté donnée...

Le principe utilisé par l'algorithme HITS est le suivant : "*un document a un poids "authority" élevé s'il est pointé par de nombreux documents ayant un poids "hub" élevé et, vice versa, un document a un poids "hub" élevé s'il pointe vers beaucoup de documents ayant un poids "authority" élevé.*"

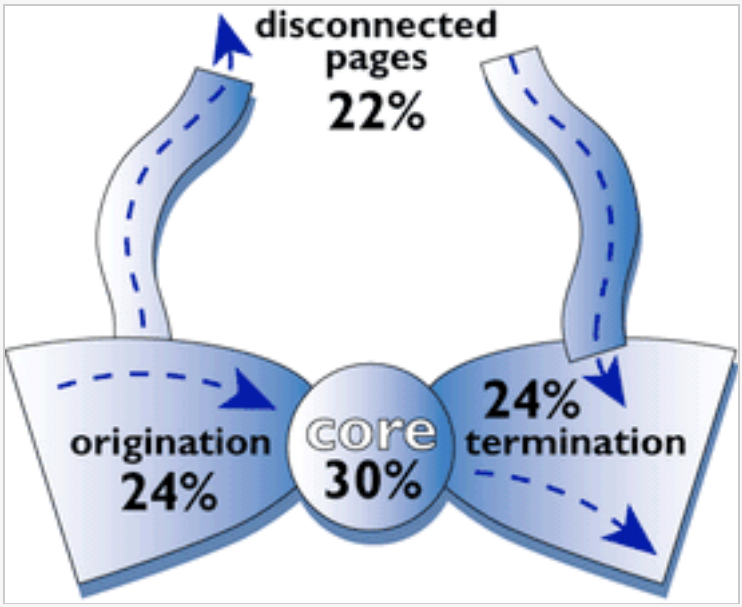
Lorsque l'analyse des liens et de l'interconnexion des pages permet de mettre en avant et de "découvrir" des "Hubs" et des "Authorities" cohérents, on peut se dire que l'on a découvert une "communauté". C'est sur cette notion qu'est basée l'algorithme de pertinence de Ask.com, adaptée de HITS sous la houlette de Apostolos Gerasoulis, professeur d'informatique à l'Université de Rutgers (New Jersey), en 1999 sous le nom de Teoma (qui signifie "expert" en langue gaëlique), technologie de recherche rachetée ensuite par Ask en 2001...

Bien entendu, tous les calculs sont effectués de façon automatique, aucune intervention humaine n'étant nécessaire pour définir quels sites sont des "hubs" ou des "authorities"... De la même façon, comme pour le PageRank, tous les calculs sont effectués de façon automatique au moment de l'indexation des informations. L'identification des communautés n'est donc pas effectuée au moment de la saisie de la requête, comme pour des sytèmes de *clustering* proposés par des outils comme Vivisimo (http://www.vivisimo.com/), Clusty (http://www.clusty.com/) ou Exalead (http://www.exalead.com/)...

La théorie du noeud papillon

Les algorithmes de Kleinberg ont également donné, en l'an 2000, naissance à la théorie du "noeud papillon" (http://domino.research.ibm.com/comm/pr.nsf/pages/news.20000511_bowtie.html). A cette époque, Altavista, Compaq et IBM avaient réalisé une étude conjointe sur le "web déconnecté". Les scientifiques de ces différents centres de recherche avaient achevé la représentation graphique d'une carte topographique complète du Web

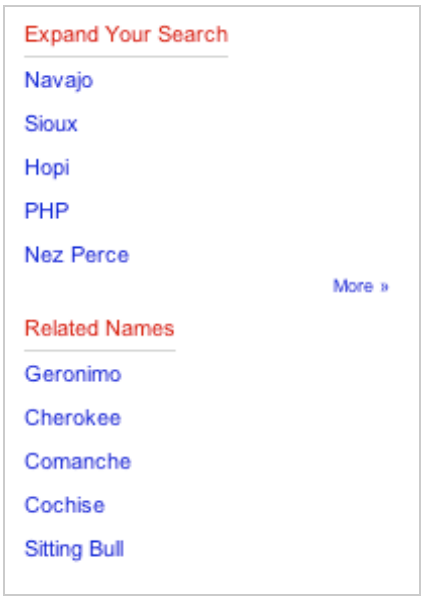
mondial et découvert l'existence de divisions entre différentes zones d'Internet, pouvant rendre la navigation sur le Web difficile, voire impraticable.
Les recherches qui avaient été effectuées auparavant, basées sur de simples échantillonnages du Web, avaient permis de conclure à un haut degré de connectivité entre les sites.



La recherche effectuée par IBM, Compaq et Altavista sur l'analyse de plus de 200 millions de pages Web, prouvait (contrairement à ce que l'on croyait) que le Web mondial était fondamentalement divisé en quatre grandes zones, chacune comprenant approximativement le même nombre de pages. On pouvait constater de même qu'un nombre impressionnant de sites Web était inaccessibles par le biais des liens hypertextes. Or, ces liens sont ce qu'un internaute utilise le plus au cours de ses navigations sur le reseau.
La théorie du "noeud papillon" permet d'appréhender la dynamique comportementale du Web et son organisation complexe. C'est au fur et à mesure des recherches que la représentation du Web s'est profilée en forme de noeud papillon.
Le " **noeud** " est constitué du "noyau ultra connecté". Les internautes peuvent aisément naviguer entre ces sites, via les liens hypertextes. Ce noyau compact constitue le coeur du réseau Internet. La partie gauche du "noeud papillon" contient les pages "**de création**". Elles permettent l'accès au coeur du Web (le noyau hyper-connecté) mais l'inverse n'est pas possible (le "noyau dur" n'a pas de liens vers elles).
La partie droite du "noeud papillon" est le contraire de l'aile droite. Les pages **de destination** sont accessibles depuis le noyau ultra connecté, mais aucun retour vers le noyau n'est possible ; c'est par exemple le cas des sites institutionnel d'entreprise qui reçoivent beaucoup de liens mais qui n'en offrent que très rarement. Des "culs de sac du Web" en quelque sorte.
La quatrième et dernière zone contient des pages "**déconnectées**". Elles sont accessibles mais ne donnent pas accès au noyau ultra connecté et ne sont "pointées" par aucune page du web.

Ainsi, sur le moteur de recherche Ask.com, les différente communautés sont proposés notamment sur la droite de la page de résultats, dans la rubrique "*Narrow your search*". Exemple pour la requête "apache" :

Ensuite, à l'intérieur de chacune des communautés identifiées dans un premier temps, des mots clés de recherche plus précis sont proposés dans deux rubriques : "*Expand your Search*" et "*Related Names*" :



Il est à noter que ces zones d'information ne sont pas encore disponibles sur le site français d'Ask.com (http://fr.ask.com/) mais qu'elles devraient l'être lors de son lancement officiel en fin d'année...

Conclusion

La notion d'*ExpertRank* et la prise en compte de la popularité des pages en termes de communautés sont certainement les particularités les plus intéressantes du moteur de recherche Ask.com. C'est également ce qui fait sa grande différence avec Google. On peut d'ailleurs s'étonner que sa communication institutionnelle soulève peu cette différenciation à une époque où tous les moteurs semblent être des clones les uns des autres. Deux outils semblent vraiment "différents" dans leur approche de la pertinence aujourd'hui : Ask.com et Exalead (sachant que l'approche "folksonomique" tentée par Yahoo! et d'autres reste encore à valider selon nous). Pourquoi ne pas faire valoir de façon plus forte cette différence pour Ask.com ?... Cela reste une énigme pour nous...

Webographie :

- **Comment référencer son site (ou blog) sur Ask.com ?**
http://www.blog-moteurs.com/ask/2006/06/comment_rfrence.html

- **L'algorithme HITS et le projet CLEVER**
http://www.webmaster-hub.com/publication/article61.html
http://www.webmaster-hub.com/publication/article82.html

- **La page perso de Jon Kleinberg**
http://www.cs.cornell.edu/home/kleinber/

- **Authoritative Sources in a Hyperlinked Environment** (John Kleinberg)
http://www.cs.cornell.edu/home/kleinber/auth.pdf

- **The Web as a graph**
http://cs.brown.edu/research/webagent/pods-2000.pdf

- **Optimisation et modélisation du graphe du Web**
http://www.enst.fr/_data/files/docs/id_511_1127992030_271.pdf

- **Analyse d'hyperliens en vue d'une meilleure description des profils**
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_110.pdf

- **Accessibility of information on the Web** (Steve Lawrence, C. Lee Giles)
http://nicomedia.math.upatras.gr/courses/mnets/mat/Lawrence&Giles_AccessibilityOfInformationOnTheWeb.pdf

Merci également à l'équipe de Ask.com France pour leur aide.