Le Web sémantique (3ème partie) : les applications du web sémantique : qui ? quoi ? pour quoi faire ?

Retour au sommaire de la lettre

Domaine :	**Recherche**	Référencement
Niveau :	**Pour tous**	Avancé

Nous continuons avec cet article notre exploration, du Web Sémantique. Après la découverte du concept et de ses définitions, puis de ses fomats dans les articles précedents, nous allons essayer de comprendre à quoi servent concrètement aujourd'hui ces notions. Une attention toute particulière a en effet été portée ces dernières années sur les outils sémantiques de recherche, due essentiellement au développement des outils issus du web sémantique. Le "search" est considéré comme une technologie fonctionnelle, limitée à une approche syntaxique. Il manquait cependant encore aux moteurs de recherche l'interprétation sémantique des requêtes ou du contenu. Des technologies comme Search Monkey ou Powerset ont ainsi mis à profit, depuis quelques temps, les technologies du "Natural Language Processing" ou du "Semantic Web". Les résultats sont certes concluants mais restent encore limités à une partie infime du Web...

Le Web sémantique, de plus en plus appelé "*Web of data*" (Web de données) est passé de l'utopie à la réalité en étant inséré dans des outils de recherche d'informations tels que les moteurs web. Ces derniers semblent miser sur la recherche de l'information sous toutes ses formes. Quelles sont les raisons de ce quasi engouement, y-a-t-il une réelle demande de l'internaute pour une telle technologie ? Nous allons tenter de faire dans cet article un état des lieux des outils incorporant des aspects du Web sémantique.

Le web sémantique dans les moteurs de recherche: pour quoi faire ?

La sémantique est techniquement un véritable défi. Satya Nadella, vice-président de la recherche chez Microsoft, résume la situation (http://blogs.msdn.com/livesearch/archive/2008/07/01/powerset-joins-live-search.aspx) : "*Les moteurs de recherche sont incapables aujourd'hui de comprendre que les mots "arbuste" et "arbre" renvoient à un même concept. On ne comprend pas pourquoi la requête sur le mot "cancer" renvoie tantôt à l'horoscope, tantôt à la maladie, et comment une recherche ou une page web font référence à l'un ou à l'autre*". Selon Satya Nadella toujours, "*un tiers des recherches n'obtiennent pas de réponse adéquate lors de la première recherche et au premier clic*". Une part très importante qu'il faut donc faire baisser...

Des moteurs de recherche sémantiques

Peter Mika, qui travaille sur la technologie Search Monkey de Yahoo!, pointe les défaillances des moteurs de recherche actuels. La liste suivante montre quelques exemples de ces limitations :
- Il est quasiment impossible pour un moteur de retourner en résultat le second sens d'un terme. Il ne considère que le sens dominant. Par exemple si l'on tente de chercher "bush", seules les réponses concernant le président apparaissent dans les premiers résultats, occultant la marque de "bière" (http://www.beerbelgium.com/Dubuisson.htm).
- Lorsqu'aucun mot-clé clair et pertinent n'existe, il n'est pas possible de décrire l'objet pour améliorer les résultats de recherche.
- La technologie de recherche actuelle n'est pas capable de satisfaire des requêtes complexes qui demandent l'intégration d'une information de type temporelle, analytique, prédictive, etc.

Un exemple parlant est celui de la recherche d'opinion de plus en plus présente en ce qui concerne les produits ou les services. Parallèlement, un certain nombre de résultats encourageants ont été récoltés dans la recherche d'opinion avec l'analyse des sentiments, c'est souvent des cas où l'utilisateur souhaite retrouver des affirmations positives ou négatives

au sujet d'un produit ou d'un service. Les résultats sont agrégés et organisés. L'intégration de l'information n'est pas possible sans une représentation structurée du contenu.

Yahoo! et Search Monkey

Au printemps dernier, on nous annonçait l'arrivée de *Search Monkey*. La firme Yahoo! a mis à disposition des développeurs une plate-forme pour mettre au point des applications qui agrémentent, selon la volonté des utilisateurs, les résultats de sa recherche (http://developer.yahoo.com/searchmonkey/). Pour ce faire, il rend disponible les données de son index, soit les données structurées avec des microformats, RDFa et déjà indexées, par défaut, par Yahoo, soit directement l'ensemble des données de la page Web à partir desquelles le développeur peut créer une extraction personnelle.

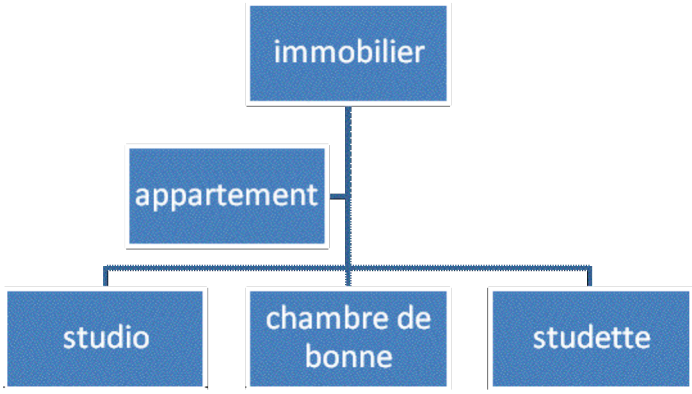
Très concrètement, que fait Search Monkey ?

- Il permet d'ajouter des informations aux résultats de Yahoo Search.
- Il permet de personnaliser la présentation des résultats.
- Il permet d'ajouter une dimension sociale aux sites dans la mesure où il s'agit d'une technologie collaborative.
- Il permet d'ajouter de l'information automatiquement en faisant des liens entre l'information des pages d'un site et le vocabulaire structuré présent dans les bases de Yahoo!.

En effet, Yahoo! structure l'information en la représentant sous la forme d'ontologies qui exploitent les données structurées en RDF. L'information d'un site, ainsi structurée en RDF est ajoutée aux ontologies existantes. A partir de là, toute information du site peut être reliée à d'autres informations externes au site grâce aux relations qui existent au sein même de l'ontologie.

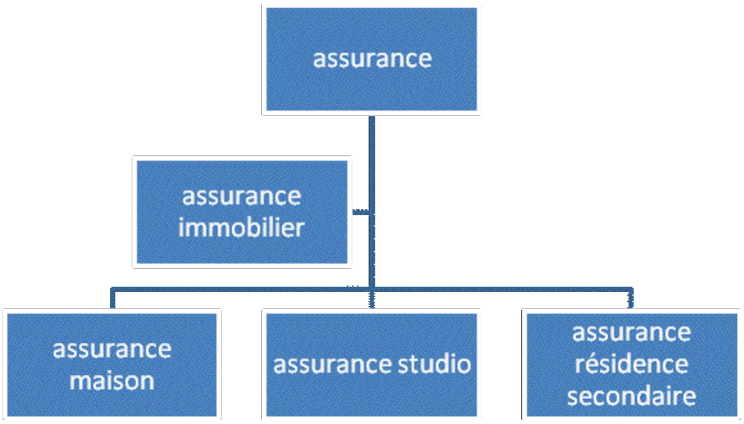
Prenons un exemple simple : Un site d'immobilier a structuré son information en RDF, cette information va être représentée sous la forme d'une ontologie. Imaginons que nous soyons sur une page de location de studios. Mon information est représentée de telle manière que le moteur "sait" qu'un studio est un "petit appartement" et qu'"un petit appartement" fait partie du domaine de l'"immobilier".

On représentera cette connaissance selon le schéma suivant :



Ontologie du secteur "immobilier"

De la même manière, il existe au sein de la plate-forme *Search Monkey*, une ontologie qui représente les informations concernant le secteur des assurances.
Cette ontologie peut être représentée pour partie selon le schéma suivant :

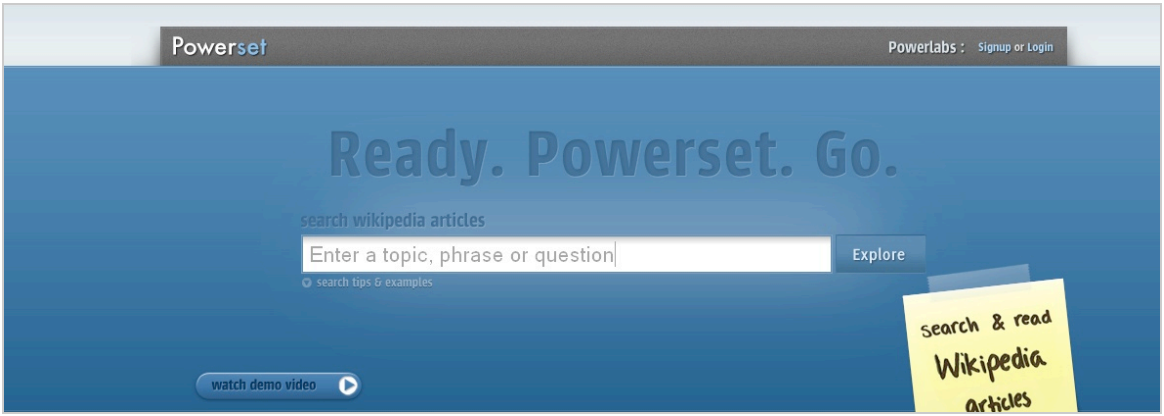


Ontologie du secteur "assurances"

Avec cette connaissance, le moteur reliera automatiquement "assurance immobilier" à "immobilier". Il en déduira également que "studio" est relié à "assurance immobilier" et par-delà à "assurance studio". Concrètement, la page "studio" du site immobilier sera reliée automatiquement à la page "assurance studio" d'un autre site concernant les assurances.

Powerset

A la même époque que Search Monkey, une jeune société installée dans la Silicon Valley sort un moteur de recherche sémantique : Powerset (http://www.powerset.com/). L'objectif est de révolutionner les pratiques de recherche des internautes en dépassant la simple recherche par mots-clés. Powerset indexe le web différemment en essayant d'analyser le sens des pages qui pourraient correspondre aux recherches. A la différence des autres moteurs, il ne s'arrête donc pas uniquement aux mots-clés.



Powerset mise sur les requêtes en langage naturel. Qu'est-ce que ça signifie ? Que l'on peut poser une question sous forme écrite comme on la poserait à l'oral dans sa langue maternelle. De fait, *exit* les mots-clés, le moteur se doit de "comprendre" la question et trouver l'information la plus juste qui puisse y répondre.

Powerset propose de tester son moteur sur les pages de Wikipédia. Pourquoi Wikipédia ? Parce que pour qu'un moteur de recherche sémantique fonctionne, il lui faut de l'information. Et pour qu'il y ait de l 'information, il faut qu'il y ait du texte, du contenu avec des mots... Beaucoup de mots. Or, quoi de plus approprié qu'une encyclopédie pour trouver une réponse à une question ?

Comment fonctionne Powerset ?

Powerset, à la différence de Search Monkey, fonctionne sur deux technologies : celle du "Web sémantique" avec la structuration de l'information (RDF) et sa représentation sous la forme d'ontologies et celle du traitement automatique du langage (ou *Natural Language Processing*). La première technologie consiste en la structuration des données Wikipédia en RDF puis sa représentation sous la forme d'une immense ontologie. La seconde consiste en l'analyse linguistique des articles afin d'identifier les marqueurs linguistiques qui peuvent réagir à l'expression d'une réponse en rapport avec une question.

Prenons un exemple simple : Si on prend la requête formulée en langage naturel :

Qu'est-ce que Powerset ?

L'analyse des marqueurs linguistiques préalable permettra au moteur d'identifier les deux points du titre de cet article :

Powerset : un moteur de recherche en langage naturel

Les deux points expriment linguistiquement qu'il s'agit d'une annonce thématique. C'est-à-dire qu'en aval des deux points, très souvent on détaille ce qu'il y a en amont ou bien on donne une explication. Le moteur ressortira donc en réponse l'extrait de l'article avec le titre surligné comme réponse à la question.

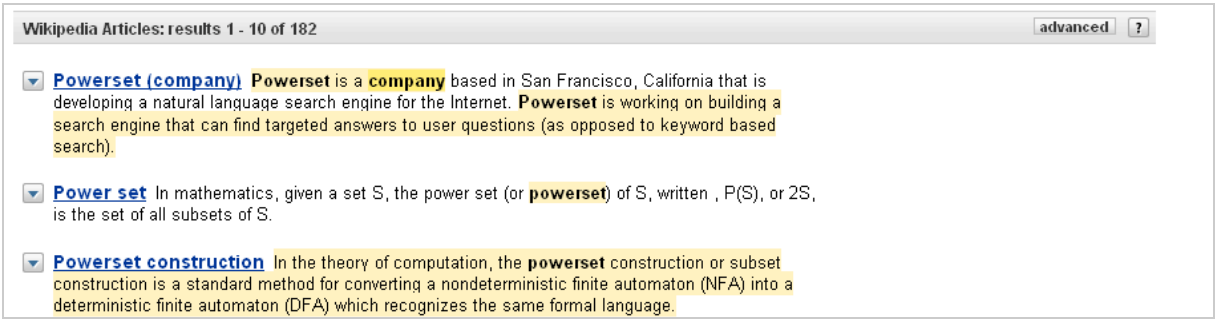
Cas concret : ***Le test "Question"***

A la question *What is Powerset?* Le moteur affiche en réponse la partie surlignée de l'article Wikipédia qui lui semble la plus pertinente.

*Powerset is a **company.***

En outre, il va surligner une autre partie qui vient compléter la première réponse :

Powerset is working on building a search engine that can find targeted answers to user questions (as opposed to keyword based search).



Au point de vue pertinence, le test est concluant. Le moteur retrouve la réponse linguistiquement exacte à la question.
La question est formulée en mot interrogatif (*what*) +verbe (*is*) et la réponse retrouvée reprend la même structure que la question.

***What is** Powerset ? → **Powerset is** ...*

Passons à présent au ***test "mot-clé"*** :

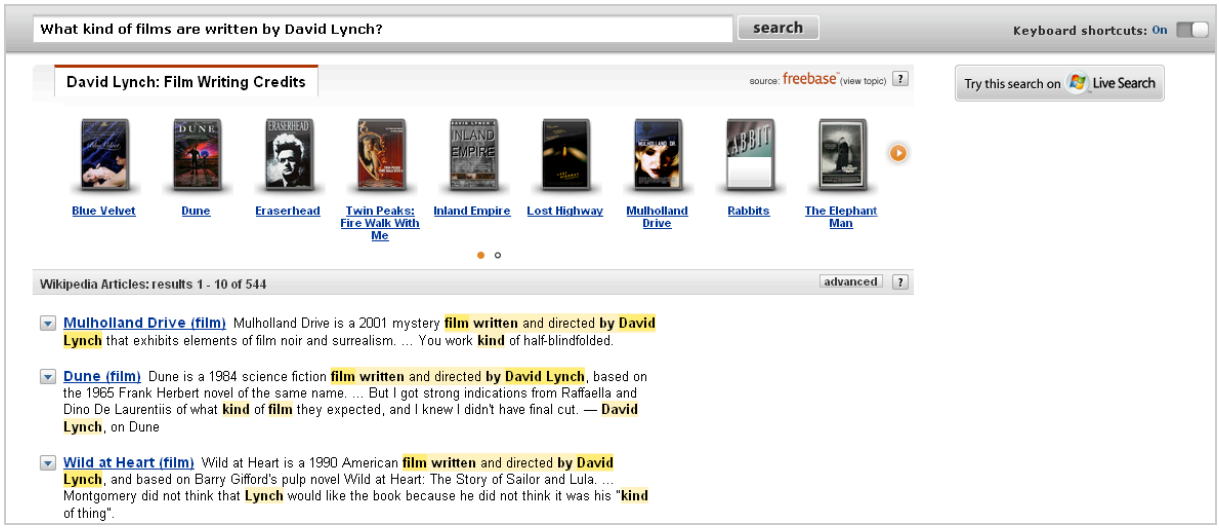
Au mot-clé "*powerset*", le moteur nous renvoie la même page de résultats que précédemment. A première vue, il n'y a pas de différence flagrante. Cependant, la réponse surlignée n'est pas exactement la même. Précédemment, "*company*" était surligné en jaune comme réponse première et la plus pertinente. A présent, le mot n'est plus surligné. Le paragraphe entier est surligné de la même couleur. Il n'y a donc pas de distinction de la partie la plus pertinente par rapport à la question.

Powerset is a company based in San Francisco, California that is developing a natural language search engine for the Internet. Powerset is working on building a search engine that can find targeted answers to user questions (as opposed to keyword based search).

Le moteur renvoie donc une réponse plus pertinente pour une requête formulée sous la forme de question que sous la forme de mot-clé. En effet, la question est fermée, elle oriente la réponse, alors que le mot-clé est ouvert et ne prédit pas la réponse.

Comment sont affichés les résultats ?

Comme le montre la capture d'écran ci-dessous, Powerset affiche des résultats qui présentent des caractéristiques intéressantes :
- Chaque lien de résultat offre un condensé des informations qu'on peut trouver sur chaque résultat.
- Selon la question, les réponses sont agrémentées d'informations trouvées sur Wikipedia, en rapport direct avec le sujet principal.
- Quand cela est possible, les résultats sont accompagnés de médias complémentaires.



Résultats de recherche à la question: What kind of films are written by David Lynch?

Au-delà de Wikipédia...

La question qui se pose est la suivante : au delà de Wikipédia, le moteur est-il toujours aussi pertinent ? En d'autres mots, quand on lance le moteur sur de l'information peu ou pas structurée, parfois pauvre en contenu comme peut l'être le Web pris de façon globale, y-a-t-il assez de matière pour que le moteur réagisse favorablement ?

Powerset n'ayant pas été lancé en dehors de la base de Wikipédia (complétée par celle de Freebase), aucune donnée ne nous est encore fournie de ce côté-là. Néanmoins, on peut penser que sans une matière (texte) fortement structurée tant qualitativement que quantitativement, le défi du moteur sera difficile à relever. Son récent rachat par Microsoft pourrait cependant permettre de visualiser d'autres applications d'ici peu.

Etat des lieux des moteurs de recherche sémantiques

Qu'en est-il, fin 2008, de cette mise sur le marché de moteurs de recherche sémantiques ? Y-a-t-il eu un bouleversement des pratiques de recherche sur le Web ? Des manières de structurer l'information au sein des sites web ?

De toute évidence, ces nouvelles technologies n'ont quasiment pas eu d'impact sur le quotidien de l'internaute. Alors que les usagers de la Toile ont été formatés à la recherche sur le web par mots-clés, il est difficile d'envisager un renversement instantané des pratiques. Mais il est vrai que la propagation du Web sémantique se fera par touches successives et surtout sur le long terme puisque – et c'est là la clé du web de demain - l'Internet deviendra de plus en plus collaboratif. Les technologies – et qui plus est le Web sémantique - fonctionneront donc à la mesure de la collaboration de tout un chacun.

La technologie coûte de l'argent, des ressources et du temps. Des moteurs comme Powerset n'ont pas encore indexé tout le web. Cela nécessiterait une grosse levée de fonds. Son rachat par Microsoft permettra peut-être de palier ce besoin... Il reste à voir ce que donneront ces moteurs une fois le Web indexé dans de plus fortes proportions...

Myriam MALORON
veille@brioude-internet.fr
Brioude Internet Référencement

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2009/01/le-web-smantique-3me-partie-les.html