# Vers un moteur de recherche conceptuel grâce à la sémantique ?

Retour au sommaire de la lettre

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

*On entend souvent parler de moteurs de recherche sémantique, sans savoir souvent ce que l'on met sous cette appellation. Qu'est-ce qu'un moteur de recherche sémantique ? Existe-t-il des moteurs sémantiques qui ne le soient pas vraiment ? Qu'entend-on par "moteur conceptuel" et comment se fait-il qu'aucun moteur réellement sémantique n'ait réussi à percer jusqu'à aujourd'hui sur le Web ? Pourtant, Google a réussi à intégrer dans ses algorithmes des "morceaux" de tels outils, notamment avec les technologies d'Applied Semantics et d'Orion. Alors, un moteur sémantique et conceptuel peut-il émerger dans un proche avenir ? Voici quelques éléments de réponse...*

Historiquement, les personnes qui ont développé des moteurs de recherche ont exploré deux voies différentes : les **moteurs de recherche syntaxiques** et les **moteurs de recherche sémantiques**.

Les **moteurs de recherche syntaxiques** considèrent les textes comme des suites de caractères sans signification. Les documents retournés par ces moteurs sont classés en utilisant principalement des calculs de similarité entre chaînes de caractères. Google est un illustre représentant de cette classe de moteurs, ainsi que Bing, Yahoo ou Ask...

Dans les **moteurs de recherche sémantiques**, le contenu est considéré comme une suite de termes associés à des concepts, et le classement des documents prend en compte les relations sémantiques entre ces concepts.

En théorie, les moteurs de recherche sémantiques ont une supériorité certaine sur les moteurs de recherche syntaxiques. Les résultats renvoyés par ces moteurs contiennent moins de bruit : les résultats non-pertinents sont peu nombreux. Leur précision est meilleure : c'est-à-dire le rapport du nombre de documents pertinents trouvés au nombre total de documents renvoyés. Mais ceci est vrai en théorie seulement, car en pratique, la supériorité des moteurs syntaxiques est écrasante, et aucun moteur de recherche sémantique n'a réussi à sortir du lot.

Dans la suite de cet article, nous allons essayer de mieux comprendre ce qu'est réellement un moteur sémantique (après avoir fait le tour de tout ce que l'on appelle, parfois à tort, les *moteurs sémantiques*). Nous essaierons également de comprendre pourquoi les moteurs sémantiques purs ne parviennent pas à concurrencer les moteurs syntaxiques. Enfin, nous nous intéresserons aux fonctionnalités sémantiques que des moteurs comme Google ont introduites récemment.

## Les faux amis : les moteurs "dits sémantiques"

Pour éviter toute confusion, on appellera par la suite dans cet article "moteur conceptuel" les moteurs sémantiques qui s'appuient réellement sur les relations sémantiques entre concepts pour fonctionner. Les "*concept based search engines*", comme on les appelle en anglais, sont effectivement fondamentalement différents de bien des moteurs qualifiés aujourd'hui de sémantiques.

Il existe en effet trois catégories de moteurs "dits sémantiques", mais non parfaitement conceptuels.

### *Les moteurs sémantiques travaillant à partir de métadonnées*

La première catégorie rassemble les moteurs fonctionnant sur des pages web issues du "web sémantique". Rappelons ici que le web sémantique regroupe des pages web écrites à partir d'une famille de langages normalisés par le consortium W3C, à partir d'une initiative de Tim Berners Lee. Ces pages utilisent un système de métadonnées pour rendre leur contenu plus intelligible par des logiciels. Ces métadonnées permettent d'étiqueter le contenu textuel, pour

par exemple indiquer que tel nom correspond à un écrivain, tel chiffre à un code postal, tel paragraphe à un résumé etc.

Le terme "sémantique" utilisé ici est un peu abusif, car il laisse entendre que ces techniques permettent de s'intéresser au sens du contenu, alors qu'il ne permet que de catégoriser des contenus, et de "structurer" des informations habituellement non structurées sur des pages web normales. Néanmoins, le web sémantique permet réellement de travailler sur des concepts et des relations sémantiques entre concepts, mais ces relations sont trop limitées en général pour réussir à créer des moteurs réellement et pleinement "conceptuels".

Le web sémantique représente encore aujourd'hui un sous ensemble très réduit comparé au reste du web. La plupart des moteurs généralistes qui s'appuient sur des métadonnées pour fonctionner cherchent donc à "structurer" eux-mêmes automatiquement les pages qui ne sont pas issues du web sémantique. Certains utilisent leur propre système de balises propriétaires, d'autres utilisent vraiment des langages W3C comme OWL et RDF.

L'un des exemples caractéristiques de ce type de moteur est Hakia, qui s'appuie sur un balisage avancé des entités nommées pour filtrer les documents.

*Page de résultats de Hakia sur la requête "Joe Biden"*

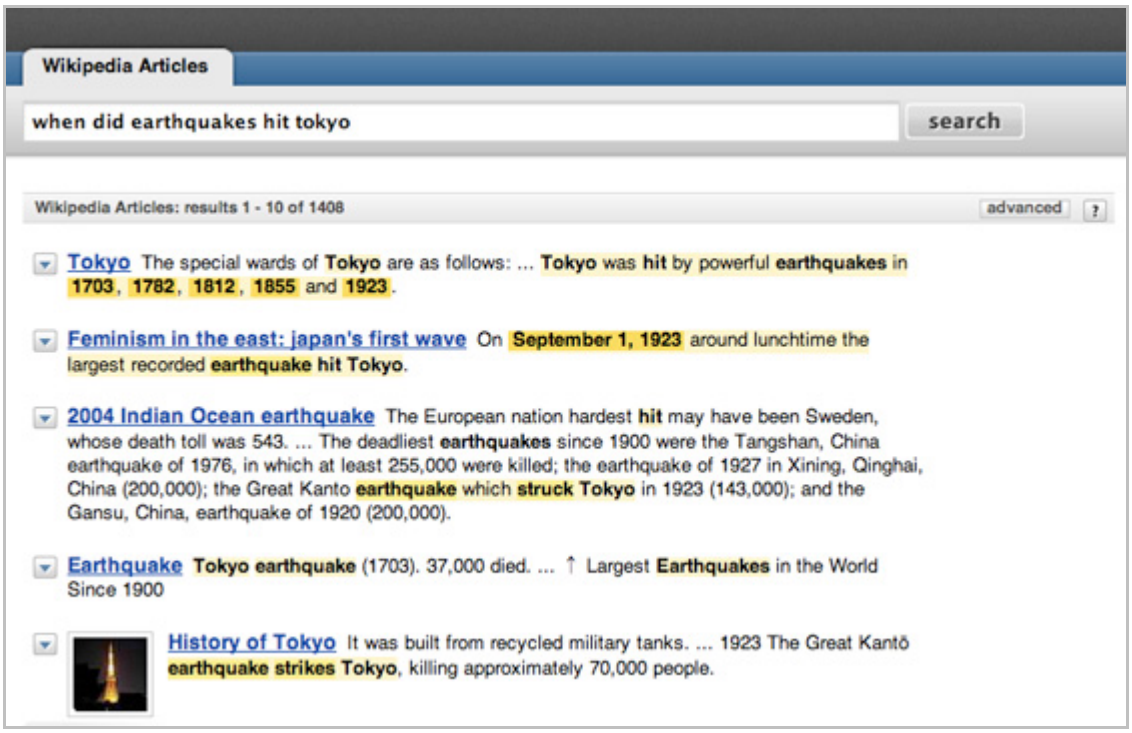
### *La recherche en langage naturel*

Certains moteurs qualifiés de "sémantiques" cherchent avant tout à essayer de comprendre le sens de la requête pour filtrer les résultats et pour ne fournir, si possible, que des résultats pertinents. En effet, le problème avec les moteurs 100% syntaxiques, c'est que les requêtes tapées par les internautes sont de simples suite de mots clés, parfois totalement ambiguës. Par exemple, si quelqu'un tape "java" dans un moteur de recherche, que cherche-t-il ? Des infos sur le langage Java, sur la danse, sur une île d'Indonésie ? En fait il manque des informations de contexte pour parvenir à désambiguïser cette requête, selon le terme consacré.

Dans la vraie vie, un humain aurait probablement posé oralement une question du genre "Où puis-je apprendre à danser la java ?", ce qui lève clairement toute ambiguïté.

Certains moteurs dits sémantiques s'appuient sur donc sur des requêtes exprimées en langage naturel pour fonctionner.

Un bon exemple de moteur de ce type est Powerset, dont la technologie a été rachetée par Microsoft pour l'intégrer dans Bing. Powerset utilise à la fois la recherche en langage naturel et utilise ensuite les métadonnées ajoutées dans son index pour présenter des résultats correspondant uniquement au "concept" ou à l'entité nommée identifiée dans la requête. Là encore, la manipulation de concepts est très limitée dans Powerset, et le moteur ne fonctionne correctement que sur des thématiques précises, à partir d'un petit nombre de ressources structurées (comme Wikipedia par exemple).

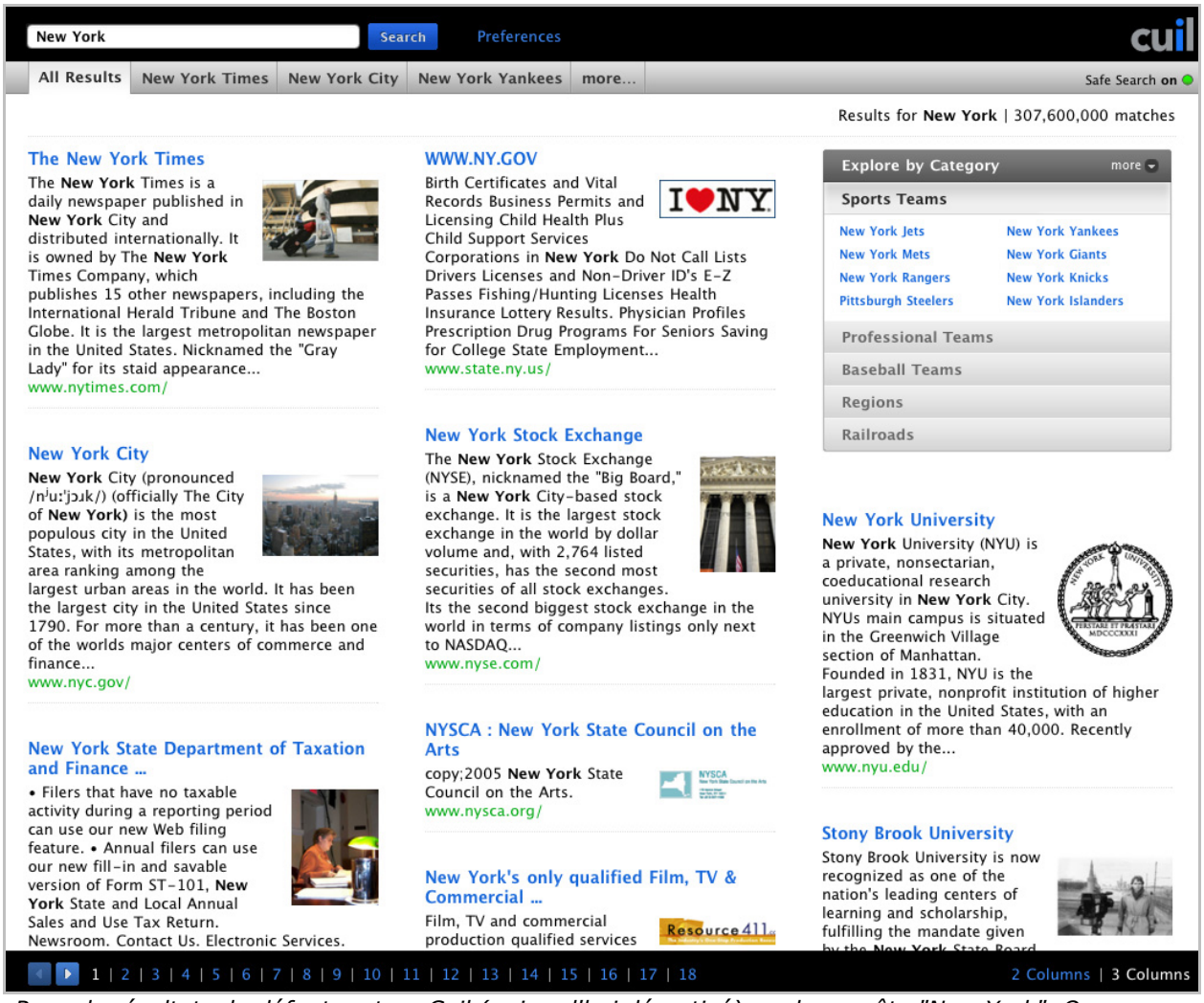


*Exemple de résultat fourni par Powerset sur la base des informations fournies par Wikipedia*

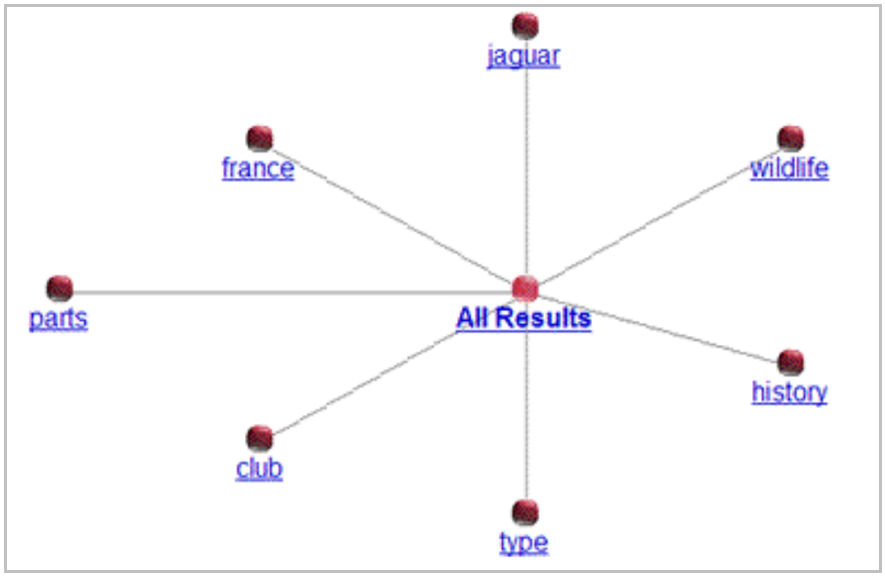
***Les moteurs à expansion de requête sémantique***

Toujours pour résoudre les problèmes engendrés par la frappe de mots clés ambigus, d'autres moteurs dits sémantiques choisissent une approche différente : proposer toutes les réponses correspondant aux différents sens du terme, mais triés et regroupés en fonction d'un concept commun.

L'utilisateur se voit communiquer, après la requête, une liste de liens vers ces différents groupes de réponses ("clusters" en anglais).




*Page de résultats du défunt moteur Cuil (aujourd'hui désactivé) sur la requête "New York". On note dans l'encadré grisé des suggestions de requêtes très complètes et précises qui permettent d'affiner efficacement la recherche.*




*Navigation visuelle entre concepts associés autour de la requête Jaguar proposée sur le moteur australien Mooter en 2004. La technologie associée a été abandonnée, le moteur est revenu à une logique syntaxique depuis.*

## *Les "vrais" moteurs de recherche sémantiques : la recherche sur les concepts*

A la différence des moteurs de recherche précédents, un vrai moteur de recherche conceptuel adopte une approche différente. Dans un premier temps, un traitement sémantique de la requête est nécessaire. L'idée ici est d'associer une phrase, ou une suite de mots clés avec un "concept" prédéfini dans le système.

Tout se passe comme si le système comprenait le sens de la requête : en réalité, il faut garder à l'esprit qu'il s'agit d'une illusion, le logiciel ne fait qu'imiter très imparfaitement le fonctionnement du cerveau humain, et ne sait que calculer des correspondances et des similarités entre ce qui est tapé par un utilisateur et des entrées dans sa base de données.

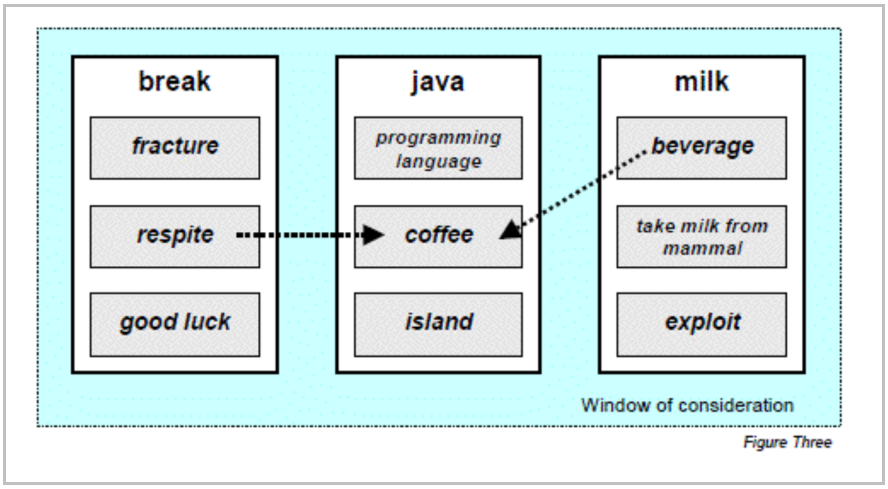
Le problème principal est ici d'avoir assez de données sur le contexte pour parvenir à déterminer sans risque d'erreur à quels concepts se rattache la phrase entrée par l'internaute. Par exemple la requête "spécificités de la syntaxe du Java" appelle clairement des documents parlant de la syntaxe du langage Java, à l'exclusion de toute autre chose. Mais pour identifier le concept "syntaxe du langage Java", il faut réaliser plusieurs opérations successives.

La première passe consiste à s'appuyer sur la grammaire pour limiter les acceptions possibles des termes (on procède à une analyse syntaxique). Le problème, c'est que la même orthographe peut correspondre à des formes verbales (un verbe conjugué), un adverbe, un adjectif ou un nom. Les noms propres peuvent également avoir la même graphie que des verbes conjugués (Martin et martin-pêcheur)...

Dans l'exemple ci-dessus, la grammaire est d'un grand secours : on identifie facilement un groupe nominal ("du Java") et l'utilisation de "du" indique clairement que le seul sens possible est le langage informatique (la danse aurait impliqué "de la" et l'île de "de").

L'analyse syntaxique permet souvent d'éliminer des possibilités, mais une seconde passe est souvent nécessaire. L'objectif de ce deuxième traitement est d'éliminer les ambiguïtés qui peuvent subsister. La solution pour y parvenir est de s'appuyer sur le contexte pour y repérer des termes qui collent ou ne collent pas avec tel "sens" du mot.

Par exemple, dans l'exemple illustré ci-dessous, le terme "java" est clairement identifié comme correspondant au terme d'argot américain désignant un café, parce qu'il est accompagné des termes "respite" et "beverage" ( "pause" et "boisson").




*Tiré de la documentation d'Applied Semantics : les différents sens des termes sont choisis en fonction du contexte. Ici Java=coffee est clairement identifié grâce à la présence des termes "respite" et "beverage" à proximité*

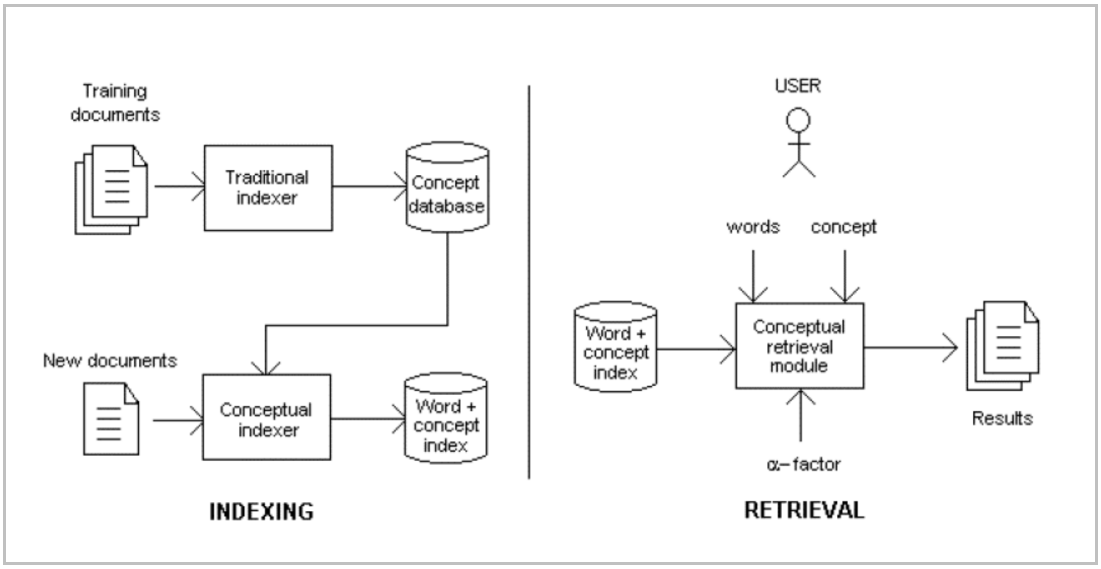
Une fois le concept identifié, le moteur de recherche conceptuel va essayer de renvoyer des documents associés aux mêmes concepts. Cela signifie que, lors du processus de création de la base de données (l'index) du moteur conceptuel, chaque document a lui même été analysé, découpé en éléments simples, puis a fait l'objet d'une analyse syntaxique et de multiples traitements de désambiguïsation.

A l'issue de ces traitements, des metadonnées (balises) sont ajoutées pour associer les termes à des catégories précises (par exemple Victor Hugo->écrivain, Victor Hugo->personne, etc.) et des portions de documents à des thématiques précises (alinéa 2 ->syntaxe Java).

Pour parvenir à de tels résultats, il faut faire appel à des bases de données spéciales baptisées **ontologies**. Les ontologies servent à stocker les termes et les relations qui les unissent sous la forme de triplets {terme1,relation avec,terme2} : par exemple {branche,est un élément,arbre} ou {chaud, est le contraire de, froid} ou encore {perroquet,est une espèce de,oiseau}. L'ontologie du moteur conceptuel contient également beaucoup de relations sémantiques dites "latérales", c'est à dire une simple relation de co-occurence entre des termes sans que l'on puisse identifier une relation simple entre eux (comme l'association café<->pause).

Un moteur conceptuel est donc la combinaison d'un moteur utilisant les métadonnées et la reconnaissance d'entités nommées, et celle d'un moteur utilisant la reconnaissance du langage naturel, et s'appuyant sur un index bâti autour d'une ontologie.

Il reste néanmoins à faire mieux que sélectionner une collection de documents associés aux mêmes concepts que ceux contenus dans la requête. Comme sur un moteur syntaxique, il faut parvenir à classer les résultats, pour présenter en premier les documents censés être les plus pertinents. Différents algorithmes ont été inventés, certains font appels à un calcul de similarité comme sur un moteur classique (syntaxique), d'autres à un calcul de probabilité de pertinence.




## *Un exemple de tentative de moteur conceptuel : Oingo*

Historiquement, on recense peu de tentatives de lancement de moteurs de recherche commerciaux et généralistes complètement conceptuels. Oingo constitue l'un des rares moteurs de ce type : ce qui est intéressant, c'est que c'est un exemple ancien.

A la fin des années 90, Adam Weissman et Gilad Elbaz, deux spécialistes de linguistique informatique et du traitement automatisé des langues décident de lancer un moteur de recherche travaillant sur le sens ou les concepts. Avec l'aide de quelques membres de leur famille, ils créent la société Applied Semantics (http://www.appliedsemantics.com/), et commencent par recruter une armée de linguistes, de grammairiens, de lexicologues et de philologues pour les aider à construire les bases de données nécessaires à leur entreprise (les fameuses ontologies). Au bout de quelques mois, ils lancent leur moteur : Oingo, et déposent fin 1999 un brevet décrivant leur algorithme.

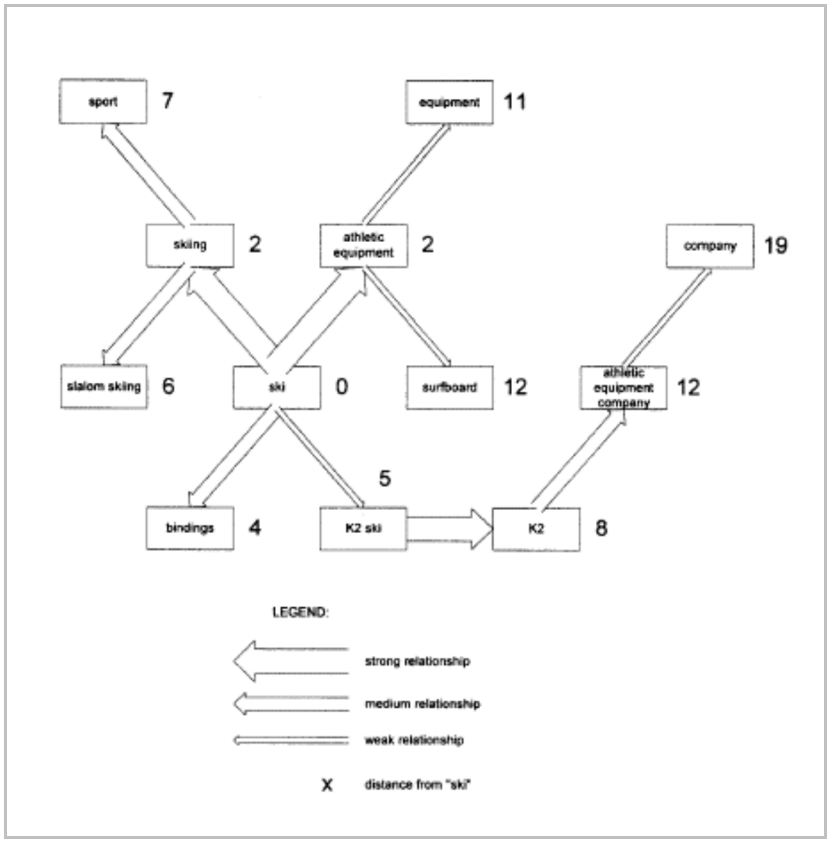
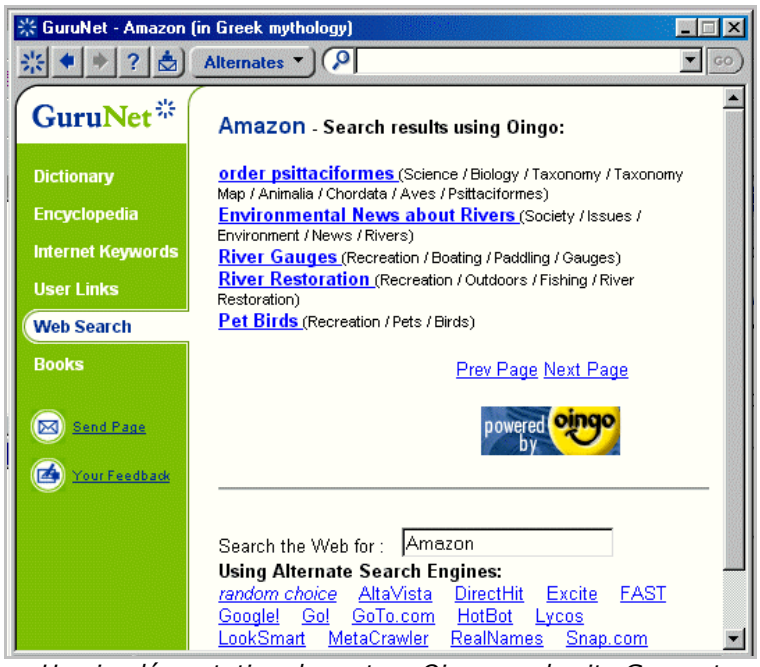



*Illustration issue du brevet sur le moteur Oingo : l'ontologie stocke non seulement les relations entre les termes, mais aussi un score mesurant la force de la relation unissant les concepts*

Faute de moyens pour faire connaître leur moteur auprès du grand public, Applied Semantics cherche des partenaires désireux de réutiliser leur technologie, qu'ils baptisent CIRCA (*Conceptual Information Retrieval and Communication Architecture*). Le seul partenariat important qu'ils parviendront à négocier sera l'accord signé avec USA Today pour réutiliser la technologie CIRCA dans le moteur de recherche interne du site d'actualités américain.




*Une implémentation du moteur Oingo sur le site Gurunet*

Constatant le peu d'engouement pour leur moteur de recherche, surtout comparé avec le succès éclatant de Google à la même époque, les fondateurs d'Applied Semantics décident avec beaucoup d'à propos de réorienter leur stratégie vers des sources de monétisation plus évidentes. La technologie CIRCA sera donc détournée pour créer une méthode d'affichage de publicités contextuelles (Adsense), et directement réexploitée pour créer un service de pages de parking sponsorisées pour domaines (Domainsense).

En 2003, la technologie CIRCA attise les convoitises de Yahoo! et de Google. C'est finalement le géant de Mountain View qui achètera la société Applied Semantics, et l'ontologie CIRCA permettra de créer la technologie cachée derrière le programme de publicité contextuelle Adsense !

## *Pourquoi tous les moteurs de recherche sémantiques finissent-ils par disparaître ?*

On peut remarquer que pour la plupart des exemples cités dans cet article, il s'agit soit de moteurs disparus (comme Mooter), soit de technologies rachetées par des moteurs syntaxiques (comme Powerset ou CIRCA).

En fait, trois obstacles empêchent ces moteurs conceptuels de vraiment surclasser les moteurs syntaxiques.

Tout d'abord, pour obtenir des résultats de qualité (un fort rappel avec une excellente précision), il faut que les concepts associés à la recherche de l'internaute soient parfaitement identifiés. Beaucoup d'informations de contexte sont nécessaires (comme une question détaillée et correctement formulée en langage naturel).

Or taper une telle question représente un travail compliqué, et un internaute préfère aujourd'hui taper trois mots clés qu'une longue phrase en langage naturel. Dans l'attente de la généralisation des interfaces en langage parlé comme Google Voice sur les smartphones, les moteurs conceptuels sont condamnés à utiliser des requêtes ambiguës et donc de fournir des résultats dont la précision est dégradée. L'exploitation de l'historique des recherches et du comportement de l'utilisateur pourrait permettre une meilleure désambiguïsation, mais on ne peut pas aller assez loin dans ce domaine sans dépasser les limites habituellement fixées en matière de respect des données privées.

Le deuxième obstacle est que ces moteurs ne s'appuient pas sur des algorithmes universels : ils reposent entièrement sur des processus d'analyse syntaxique qui sont différents d'une langue à l'autre. Car même si travailler sur des concepts permet de s'affranchir de la barrière linguistique et permet une recherche parfaitement multilingue, dans le même temps, associer un passage de texte à un concept nécessite des traitements spécifiques à la langue du texte.

De plus, pour bien fonctionner, ces moteurs nécessitent des bases de connaissances énormes. Et dans la pratique, ces bases ne sont pas constituées entièrement automatiquement. Les ontologies nécessitent, pour être opérationnelles, l'intervention d'experts en lexicologie. Le résultat final, c'est que ces ontologies sont très imparfaites, contiennent des erreurs, et surtout sont très lacunaires. Or, lorsque l'information sur les concepts ou les relations entre termes vient à être absente, ces moteurs ne sont plus pertinents du tout.

Le dernier obstacle réside dans les performances médiocres de ces moteurs en termes de temps de réponse. Les algorithmes de type probabilistes sont difficiles à optimiser de ce point de vue, et le retard accumulé sur les moteurs syntaxiques en matière de scalabilité est difficile à rattraper.

Le résultat, c'est que ces technologies servent soit sur des moteurs spécialisés, pour lesquels on peut créer des ontologies très performantes sur un espace sémantique limité (par exemple dans le domaine médical, pour lequel des moteurs sémantiques impressionnants ont été réalisés), soit en tant que technologie d'appoint pour des moteurs de recherche syntaxiques.

En règle générale, les technologies sémantiques manipulant des concepts servent à réaliser des "expansions de requêtes" ou des "raffinements de requêtes", c'est à dire des suggestions

de requêtes montrées à l'utilisateur pour l'aider à reformuler sa demande pour obtenir selon les cas plus de résultats, des résultats plus précis, ou effectuer une demande voisine.

Pour l'anecdote, on peut noter qu'un brevet récemment officialisé déposé par les fondateurs d'Applied Semantics décrit justement comment utiliser une technologie proche de CIRCA pour générer des suggestions de requêtes.

Mais Google a plus récemment intégré une autre technologie conceptuelle : la technologie Orion.

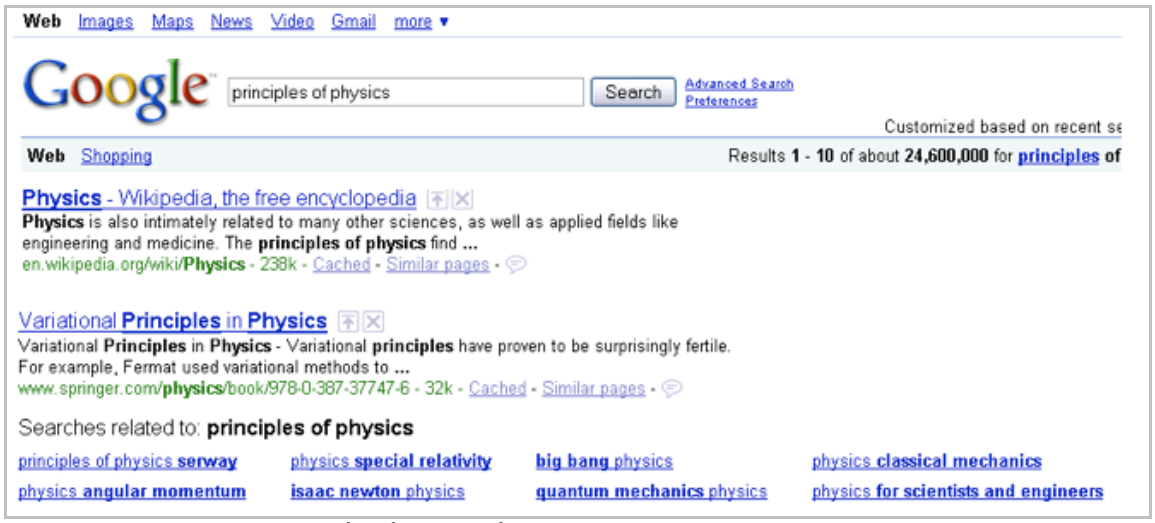
## Google Et la technologie Orion

En 2006, un jeune chercheur né en Israël, Ori Allon, annonce le lancement prochain d'un moteur de recherche basé sur une technologie sémantique "révolutionnaire" qu'il baptise "Orion".

Ce jeune diplômé de l'Université de Galles du Sud n'a en réalité jamais sorti son moteur de recherche : sa technologie a été immédiatement achetée par Google, ainsi que son silence : tous les documents décrivant sa technologie ont immédiatement disparu du net.

Trois ans plus tard, fin mars 2009, Google a implémenté cette technologie dans son moteur : elle sert à définir les recherches associées qui apparaissent en bas de page lorsque l'on tape une requête.

Voilà comment Ori Allon présentait le fonctionnement de sa technologie :
"*prenez une requête comme la Révolution Américaine [...]. Orion retournera des résultats avec des extraits de textes contenant cette phrase. Mais il donnera également des résultats pour l'Histoire Américaine, George Washington, la Guerre Révolutionnaire Américaine, la Déclaration d'Indépendance, la Boston Tea Party etc... On obtient beaucoup plus d'informations pertinentes pour chaque recherche.*"




*"Related searches" associées à la requête principles of physics. Les termes en gras sont les termes/concepts associés trouvés par le système : big bang, angular momentun, Serway, Isaac Newton, special relativity etc... Le terme "physics" est conservé pour empêcher une trop grande dérive de la requête.*

La technologie Orion sert ici de source d'expansion de requêtes, à partir d'une recherche sur les concepts associés a une requête. Il s'agit clairement d'une implémentation réussie d'une recherche de type sémantique/conceptuelle dans un moteur de recherche syntaxique.

## Vers un moteur de recherche conceptuel ? Non !

Régulièrement, des start-ups lancent des "Google killers" en s'appuyant sur des technologies sémantiques souvent qualifiées de "révolutionnaires". Force est de constater que ces

technologies sont en fait connues depuis longtemps : Oingo est presque contemporain des débuts de Google. La plupart de ces start-ups finissent par disparaître. Les technologies les plus intéressantes finissent par être rachetés par les géants de la recherche syntaxique (Applied Semantics racheté par Google, Orion également, Powerset par Bing).

En fait, pour les raisons évoquées plus haut, il est difficile aujourd'hui d'envisager un moteur de recherche conceptuel grand public et universel. Ces technologies sont plus adaptées pour créer d'impressionnants moteurs de recherche verticaux (spécialisés sur un domaine particulier : biologie, médecine, droit ...). Elles peuvent également jouer le rôle de technologie d'appoint efficaces pour des moteurs syntaxiques.

Néanmoins, les progrès effectués sur le balisage automatique et le traitement automatisé des langues laissent penser que d'autres applications pourraient voir le jour. L'approche sémantique peut se révéler plus fructueuse pour faciliter la recherche et la navigation sur un smartphone par exemple...

Demain ne verra donc pas l'avènement des moteurs de recherche sématiques contextuels. Mais après-demain ? Qui sait ...

## *Pour approfondir le sujet...*

***Brevets***

*Meaning-based information organization and retrieval*
http://patft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=6453315.PN.&OS=PN/6453315&RS=PN/6453315
Inventé par Adam J. Weissman et Gilad Israel Elbaz
United States Patent 6,453,315, attribué à Applied Semantics le 17 Septembre 2002, enregistrée le 1er novembre 1999

*Methods and systems for detecting and extracting information*
http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,689,536.PN.&OS=pn/7,689,536&RS=PN/7,689,536
Inventé par Adam J. Weissman et Gilad Israel Elbaz
US Patent 7,689,536, attribuée à Google le 30 mars 2010, enregistrée le 18 mars 2003

***Liens sur la technologie Circa***

*CIRCA Technology Overview*
http://lsdis.cs.uga.edu/SemWebCourse_files/WP/Applied-Semantics-CIRCA_wp.pdf

*CIRCA Technology: Applying Meaning to Information Management*
http://static.twoday.net/blackcat/files/applying%20meaning%20to%20information%20management.pdf

*Ontology Usage and Applications*
http://lsdis.cs.uga.edu/SemWebCourse_files/WP/Applied-semantics-ontology_usage.pdf

***Liens sur la technologie Orion***

*Two new improvements to Google results pages*
http://googleblog.blogspot.com/2009/03/two-new-improvements-to-google-results.html

***Autres liens sur le sujet***

*Discovering missing background knowledge in ontology matching*
http://eprints.biblio.unitn.it/archive/00000953/01/005.pdf
Fausto Giunchiglia and Pavel Shvaiko and Mikalai Yatskevich - February 2006
Université de Trente

*Incorporating Conceptual Matching in Search*
http://www.searchlores.org/library/incorporating.pdf
Juan M. Madrid, Susan Gauch
EECS Department University of Kansas Lawrence

*KeyConcept: A Conceptual Search Engine*
http://citeseer.uark.edu/publications/ITTC2003.pdf
Susan Gauch, Juan M. Madrid, Subhash Induri, Devanand Ravindran, Sriram Chadalavada
EECS Department, University of Kansas, Lawrence

*AgroExplorer: a Meaning Based Multilingual Search Engine*
http://www.cse.iitb.ac.in/~pb/papers/agro-icdl-feb04.pdf
Dr. Pushpak Bhattacharyaa, Sarvjeet Singhb, Tushar Chandrac , Upmanyu Misrac, Ushhan D. Gundeviac

*Concept Search: Semantics Enabled Syntactic Search*
http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-334/paper-10.pdf
Fausto Giunchiglia, Uladzimir Kharkevich, and Ilya Zaihrayeu
Department of Information Engineering and Computer Science University of Trento, Italy
SemSearch 2008, CEUR Workshop Proceedings

*A Real-time Integration of Concept-based Search and Summarization on Chinese Websites*
http://acl.ldc.upenn.edu/W/W00/W00-1319.pdf
Joe F Zhou and Weiquan Liu
Intel China Research Center

*Web Intelligence: Conceptual Search Engine and Navigation*
http://www-bisc.cs.berkeley.edu/cs199-299/Fall2003/WebIntelMasoud.pdf
Masoud Nikravesh, BISC Program, Computer Sciences Division, EECS Department
University of California, Berkeley

***Philippe Yonnet,*** ***Global SEO Strategist, WEB DMUK (Londres) – Easyroommate / Vivastreet***

**Réagissez à cet article sur le blog des abonnés d'Abondance :**
**http://blog-abonnes.abondance.com/2011/01/vers-un-moteur-de-recherche-contextuel.html**