De l'indexation des mots clés à l'indexation des concepts (2ème partie)

Retour au sommaire de la lettre

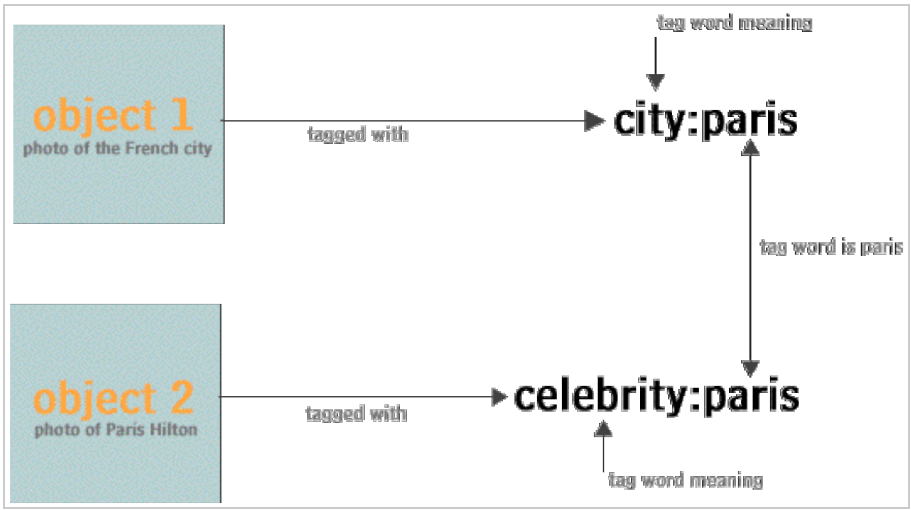
Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Depuis que les moteurs de recherche existent, le référencement existe. Mais les moteurs ont grandement évolué depuis près de 20 ans. D'un système "simple" d'indexation de mots clés isolés, ils sont passés à la détection de syntagmes, de synonymes puis aujourd'hui aux entités nommées et aux "index de concepts"... Nous continuons ce mois-ci notre exploration de ces technologies avec la seconde partie de notre article. Une évolution que les référenceurs devront prendre en compte à l'avenir pour s'adapter aux progrès toujours plus importants que font les moteurs de recherche pour mieux analyser le Web et rendre leurs résultats toujours plus pertinents...

Identifier les entités nommées : une étape de plus dans l'indexation

La "reconnaissance des entités nommées", que nous avons souvent évoquée dans la lettre d'Abondance, est une technique permettant non seulement d'identifier au sein des pages web (et des requêtes) des noms de personne, de lieux, des marques, des numéros de téléphone, des raisons sociales de société etc., mais aussi d'identifier que les termes "Michael Jackson" font allusion, selon les pages, soit à un plombier du Bronx qui porte ce nom, soit à la Popstar.

Pour pouvoir utiliser cette information dans un moteur de recherche, il convient d'ajouter une étape supplémentaire de "balisage" (tagging) dans le processus d'indexation. Nous avons vu que dans le processus traditionnel, beaucoup d'étapes avaient pour objectif de supprimer des signaux, du code et des caractères inutiles. Ce processus de "toilettage" et de "normalisation" ou de "standardisation" est un processus qui détruit de l'information. Là, au contraire, il s'agit d'ajouter de l'information à la page : on part d'un texte non structuré, et on ajoute de l'information structurée.



Exemple de taggage (balisage sémantique) sur le terme Paris : dans le premier cas, le terme est étiqueté comme correspondant à une référence à la ville de Paris, dans le deuxième cas à la célébrité Paris Hilton.

On peut ensuite générer un index spécialisé en stockant pour chaque entité nommée les pages qui les contiennent, et compléter l'index inversé avec ces informations : ainsi, pour renvoyer les pages qui parlent de Michael Jackson, on pourra chercher dans les pages associées à l'entité nommée "Michael Jackson" comme dans les pages qui contiennent "Michael Jackson".

Inutile de souligner combien le premier jeu de pages risque d'être plus pertinent que le second. En fait la difficulté ici est d'identifier les entités nommées dans la requête, or ce n'est pas toujours possible faute de contexte.

Si la requête "beat it michael jackson" est sans équivoque, la requête sur "Michael Jackson" tout court est ambigüe : l'algorithme du moteur doit être capable de retourner des résultats associés aux différentes entités nommées possibles, et mêmes aux pages qui contiennent les termes "Michael Jackson" sans qu'il ait été possible d'identifier clairement l'entité nommée évoquée dans la page. L'algorithme doit donc d'une part travailler à partir de probabilités de *matching* avec une entité nommée dans la requête, et classer les pages retournées également en fonction de la probabilité que la recherche de l'internaute porte sur une entité nommée plus qu'une autre.

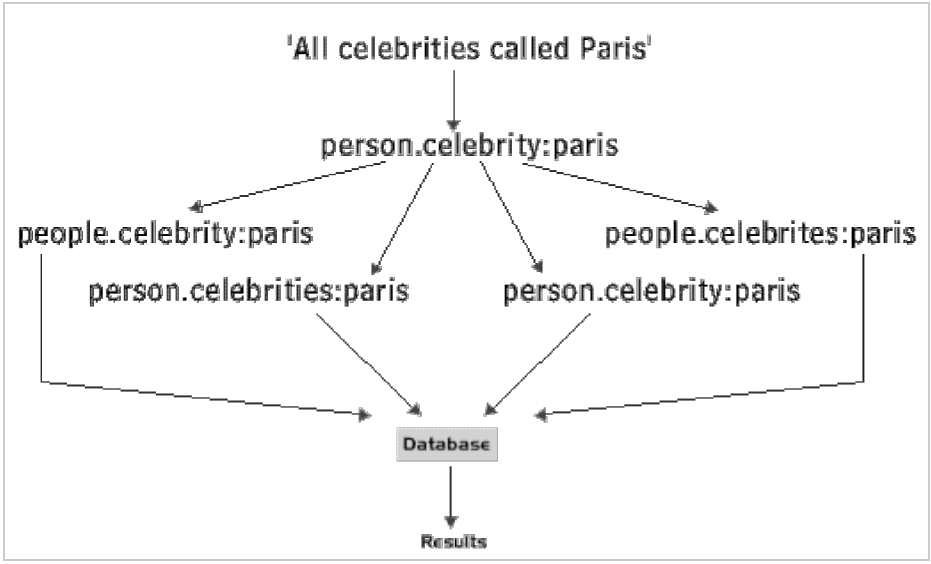


Illustration d'une requête « sémantique » sur la requête « All celebrities called Paris » : la requête, pourtant peu ambigüe terme correspond à plusieurs catégories d'entités nommées, sans que l'on puisse déterminer la plus pertinente. Le moteur cherchera à présenter des résultats composites construits à partir des items correspondants aux différents cas possibles.

Les entités nommées et les synonymes : une évolution à prendre en compte pour le SEO

La technique de reconnaissance des entités nommées permet d'associer la même étiquette à différentes graphies. Cela permet de détecter des synonymies plus ou moins parfaites entre groupe de mots, qui ne figurent pas dans les dictionnaires de synonymes usuels. Par exemple : "L'Elysée" et "La présidence de la République", "Le Président des USA" et "le Président des Etats Unis" et "Barack Obama" et "le président Obama".

Le recours de plus en plus évident aux techniques de REN (*Reconnaissance des Entités Nommées*) dans un moteur comme Google conduit certaines stratégies SEO à devenir obsolètes. Par exemple : une page sur Barack Obama sera *de facto* optimisée pour toutes les variantes des dénominations attachées à l'entité nommée "Barack Obama". Il est donc inutile en théorie :
- de faire figurer toutes ces variantes dans la page ;
- ou de créer une page optimisée sur chaque variante.

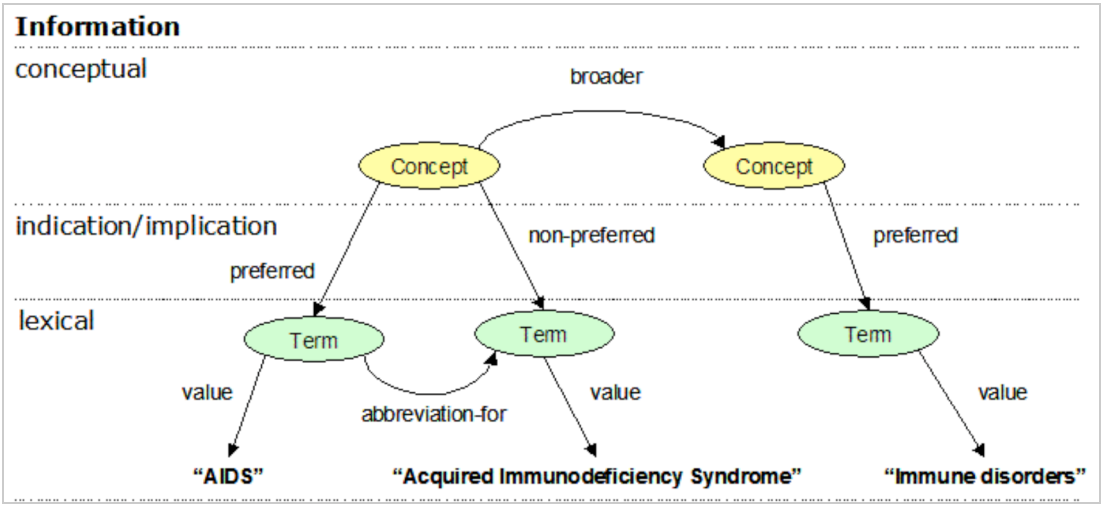
En pratique aujourd'hui, le fonctionnement du moteur est mixte, et majoritairement syntaxique (il cherche donc à renvoyer des pages jugées similaires à la requête en analysant les termes de la requête et les termes placés dans les pages). Le recours aux entités nommées concerne certains types de requêtes et est plus ou moins visible dans les pages de résultats. Il est intéressant néanmoins d'apprendre à reconnaître les cas d'utilisation de la REN pour adapter son approche en conséquence.

Le caractère mixte du moteur correspond à la situation actuelle, mais est-ce que ce sera encore vrai demain ? Google évolue progressivement vers un moteur sémantique, et sur le moteur Google US, apparaissent déjà des informations qui montrent qu'il fait mieux que reconnaître et stocker des infos sur les "entités nommées" : aujourd'hui, il stocke et exploite aussi des informations sur les relations entre les entités nommées, ou plutôt sur des "concepts".

La notion de "concept"

L'approche de Google consiste à essayer de reconnaître quelque chose de plus abstrait et plus général qu'une entité nommée : c'est ce que les spécialistes du traitement automatisé du langage appellent un "concept". Les "entités nommées" ne couvrent que les objets réels, les concepts (dans ce contexte) peuvent inclure des personnes, des lieux mais aussi des choses abstraites comme une théorie ("La Relativité"), une idéologie ("Le Communisme"), un terme abstrait concernant une famille d'objets réels ("la Montagne"), un concept ("le réchauffement climatique"), une série d'évènements historiques ou non ("révolution") etc.

L'extension de la notion "d'entité nommée" à quelque chose de plus générique est logique : la manière de catégoriser les entités est assez arbitraire et c'est le sujet de nombreuses polémiques. Même les contours de ce que l'on appelle une "entité nommée" sont assez flous dans la pratique. Pour un moteur de recherche généraliste, limiter son approche sémantique aux seuls objets dits "réels" ne peut en aucun cas être satisfaisant, son objectif est forcément beaucoup plus large.



Exemples de relations sémantiques dans un graphe conceptuel.
Le concept « SIDA » s'élargit au concept plus large des « désordres immunitaires ». La valeur « AIDS » est une valeur possible pour le terme associé à ce concept de SIDA.

Comment bâtit-on un "index de concepts" ?

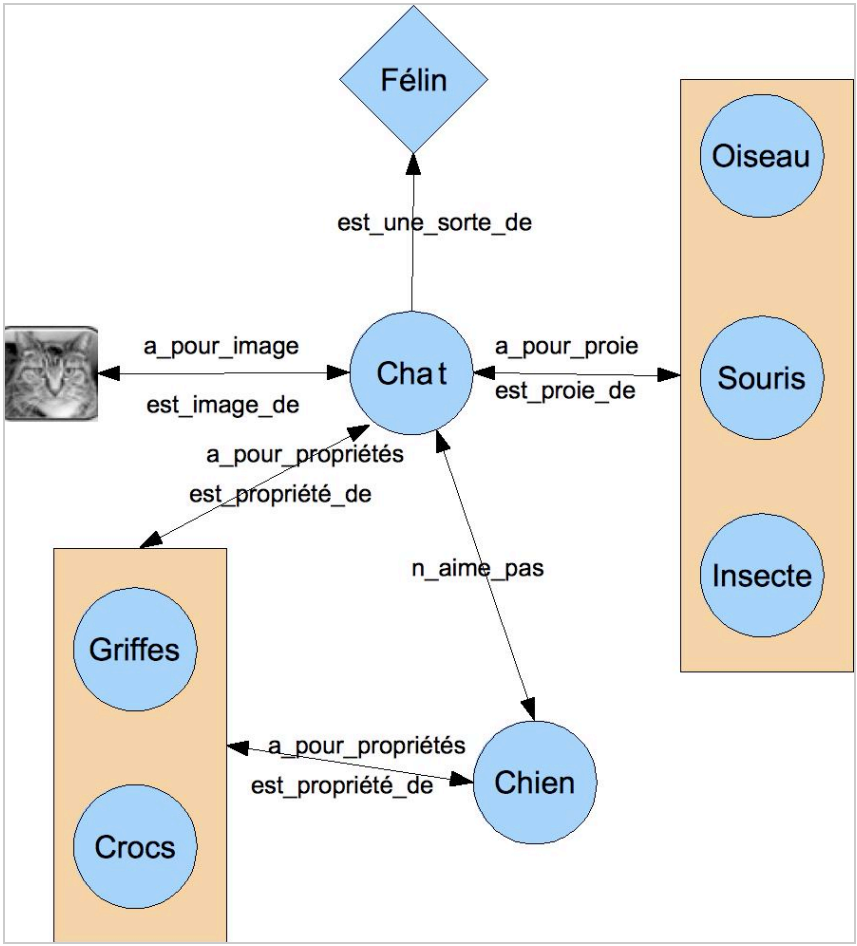
Google dans son Knowledge Graph, qui est un avatar des "graphes de concept", cherche à identifier :

- les **concepts** (ou "**connaissances**") proprement dits. L'identification des concepts est habituellement un travail de linguiste, mais on peut supposer que Google, qui a besoin de développer des solutions viables à grande échelle et universelles (fonctionnant dans de nombreuses langues et pays), serait confronté à un problème de scalabilité avec une approche humaine. Il est donc probable que Google a mis au point une approche automatique par "induction" pour détecter des candidats au statut de concept, ce qui n'exclut pas le recours à des linguistes.

- une fois le concept identifié, une approche de type "**fouille de données**" (« **data mining** ») permet de détecter les "attributs" (l'âge, le poids, l'altitude...) qui sont habituellement

associés avec ce concept, ainsi que la valeur de ces attributs pour l'item identifié (25 ans, 50 kilos, 4807 m).

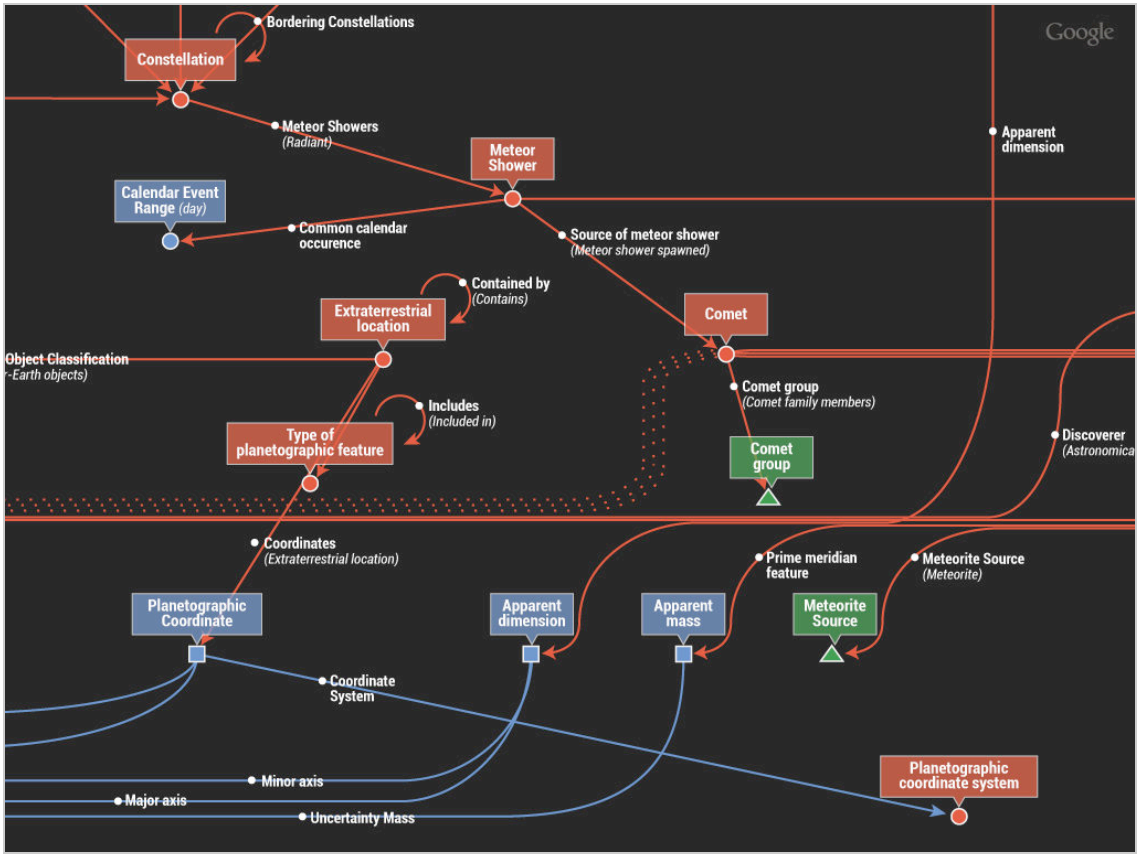
- en parallèle, le moteur cherche à identifier les relations entre les concepts. Il s'agit non seulement d'identifier le fait que le concept A est relié à B, mais aussi d'identifier la nature de la relation. Par exemple "Danton", "Révolution Française" et "la Marseillaise" sont trois entités différentes mais que l'on peut relier. Les relations pourraient être : *"Danton" est un acteur de la "Révolution Française". "La Marseillaise" est un chant de la "Révolution Française"*. Notons que ces relations sémantiques ne sont pas transitives : si Danton est relié à la Révolution et la Révolution avec la Marseillaise, cela ne signifie pas forcément qu'il faut relier Danton à la Marseillaise.



Exemple de « graphe sémantique » autour du concept de CHAT. Dans cette illustration, le « graphe » est stocké sous la forme d'une ontologie reliant des termes.

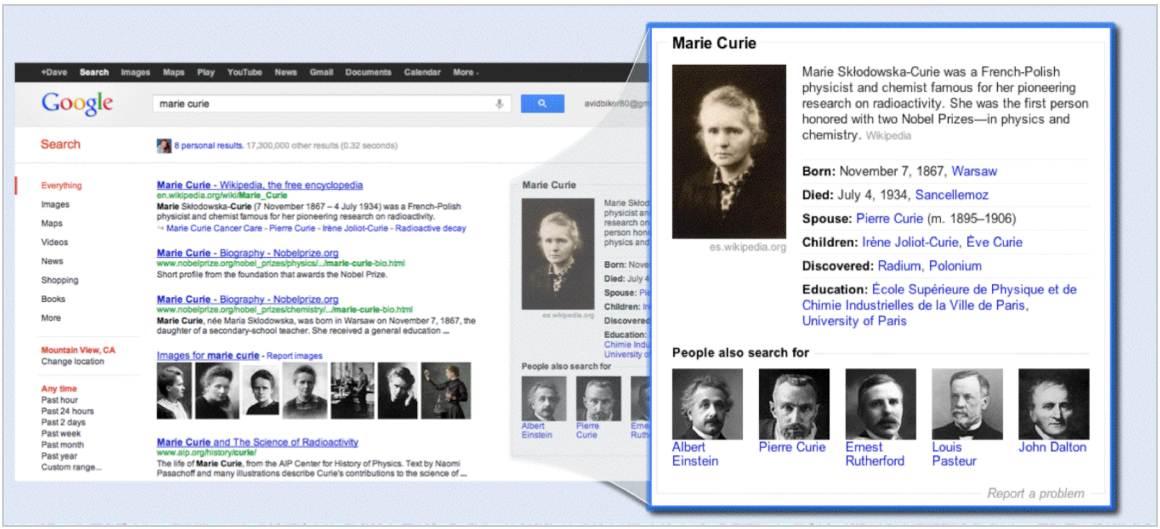
Faire du SEO sur un index de concepts

Dans l'implémentation actuelle, les pages contiennent à la fois les liens bleus traditionnels, et les informations extraites du "knowledge graph" de Google. Nous ne détaillerons pas ici pourquoi cette situation risque de durer longtemps, car ce point a déjà été abordé dans nos deux articles précédents dans la lettre "Recherche & Référencement" d'Abondance.



Le « graphe des connaissances » de Google. Le système de stockage de l'information (le « data model ») permet en réalité de ne pas avoir à traiter (en théorie) différemment les concepts et les attributs.

Il est évidemment inutile, à moins d'être partenaire de Google pour obtenir un lien vers son contenu, d'optimiser son site pour être LA source crawlée et fouillée par Google pour alimenter son « knowledge graph ». Dans l'implémentation actuelle, Google ne renvoie pas vers ses sources, il fournit les données, point à la ligne. Il n'y a rien à optimiser pour figurer sur la partie droite de la page : on ne peut pas y être présent, c'est tout.



Le "Knowledge Graph" de Google.

Par contre, il peut être particulièrement dommageable que son contenu ne soit pas correctement "associable" à une entité nommée ou à un concept. Dans une évolution où les liens bleus sont affichés de plus en plus souvent en fonction de critères sémantiques, il devient indispensable de vérifier que l'association « terme » avec l'objet référencé soit débarrassée de toute ambiguïté.

Comment s'assurer que son contenu est bien référencé dans l'index sémantique?

L'ironie de la situation, c'est qu'avec un index sémantique, on en revient à une problématique de « référencement » et non plus de « positionnement ». Ce que l'on veut, c'est que le terme « On » présent dans une page soit bien identifié à la ville du Vietnam, et non avec le pronom « on » ! L'objectif étant qu'une page d'un site de tourisme soit bien référencée parmi les pages parlant de « On ».

Comment faire ? Tout simplement en aidant les moteurs à identifier correctement les entités nommés et les concepts, en particulier :

- en évitant absolument les **erreurs de graphie** (d'orthographe) sur le site.

- en faisant attention au code qui entoure le terme à référencer : il doit faciliter l'isolement de ce terme (pour éviter des rapprochements ambigus avec d'autres termes sur la page) ou au contraire, permettre de créer un contexte facilitant la désambiguïsation.
Ex :
Ville : On
Ce bout de code html fournit à la fois une structure qui isole le terme du reste du texte, et apporte un contexte (« Ville: ») qui maximise les chances que l'entité nommée « On » soit correctement étiquetée.

- On peut aussi utiliser directement des formats de type « web sémantique » pour coder les pages ou mieux, les formats prévus par Schema.org qui vont aider les moteurs à identifier la nature des informations présentes sur les pages.

Une évolution à surveiller de près dans les mois qui viennent

Le basculement de Google vers un fonctionnement sémantique a commencé timidement il y a plusieurs années, sans que cela pèse beaucoup dans le comportement général du moteur. Le mouvement s'accélère depuis quelques semaines avec le déploiement du knowledge graph sur Google US .

Pour le moment, le nombre de requêtes impactées reste faible et ces changements ne sont pas encore visibles en France. Mais il est clair que l'implémentation de cette fonctionnalité en France n'est qu'une question de semaines. Et que ce comportement va être étendu à un nombre beaucoup plus élevé de requêtes.

Déjà, le comportement des internautes a changé : Amit Singhal, le patron du service Search Quality chez Google, en charge de l'algorithme, a révélé dans un article paru sur la version online du Wall Street Journal que le nombre de requêtes effectués par les utilisateurs avait fortement augmenté depuis ces changements (http://blogs.wsj.com/digits/2012/05/29/google-search-engine-surging-following-revamp/).

Pour le moment, les liens bleus situés à gauche des box d'informations extraites du Knowledge Graph sont encore faits de manière traditionnelle (mixte : peu de sémantique/ majoritairement syntaxique). Mais pour combien de temps ?

Les mois qui viennent s'annoncent passionnants, et cette évolution est clairement à surveiller...

Bibliographie

Le site de schema.org
http://schema.org/

Ingénierie des connaissances : représentation ontologique
http://liris.cnrs.fr/alain.mille/enseignements/IGC_M2_2008/session4/cadre_droit.htm

How does Google collect and rank results?
Un article pédagogique en anglais signé Matt Cutts
http://si110.cms.si.umich.edu/sites/si110.cms.si.umich.edu/files/InfoFinding/Google%20Index%20&%20Rank%20How-to.pdf

L'indexation des syntagmes : la fin du raisonnement par mots clés ?
Philippe YONNET
Lettre recherche et référencement d'abondance : 16 juin 2010

Brevets sur l'indexation des syntagmes :

Concept Disambiguation via Search Engine Search Results
http://appft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220120130972%22.PGNR.&OS=DN/20120130972&RS=DN/20120130972
Invented by David Ahn, Michael Paul Bieniosek, Andrei Peter Makhanov, Franco Salvetti, and Giovanni Lorenzo Thione
Assigned to Microsoft
US Patent Application 20120130972. Published May 24, 2012. Filed: November 23, 2010

Multiple index based information retrieval system
United States Patent Application 20060106792
Attribué à : Patterson; Anna L. (Google) May 18, 2006
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.html&r=1&f=G&l=50&d=PG01&p=1&S1=20060106792&OS=20060106792&RS=20060106792

Phrase identification in an information retrieval system
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.html&r=1&p=1&f=G&l=50&d=PG01&S1=20060018551.PGNR.&OS=dn/20060018551&RS=DN/20060018551
Inventors: Patterson; Anna Lynn; (San Jose, CA)
United States Patent Application 20060018551
January 26, 2006

Index server architecture using tiered and sharded phrase posting lists
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=7,693,813.PN.&OS=pn/7,693,813&RS=PN/7,693,813
Invented by Pei Cao, Nadav Eiron, Soham Mazumdar, Anna Patterson, Russell Power, and Yonatan Zunger
Assigned to Google
US Patent 7,693,813. Granted April 6, 2010. Filed March 30, 2007

***Philippe YONNET** , Directeur SEO & Merchandising, Twenga.*