Google veut devenir un moteur syntaxique et sémantique : quel impact pour le référencement ? (2ème partie)

Retour au sommaire de la lettre

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

En février 2012, plusieurs sites web ont repris les mots d'Amit Singhal, grand prêtre de l'algorithme de pertinence de Google, expliquant que le moteur de recherche allait fortement sémantiser ses résultats dès cette année et répondre directement, dès l'affichage de la SERP, aux questions posées par les internautes et comprises par Google. Mais comment Google a-t-il mis en place ces algorihmes sémantiques et quelle sera leur portée en termes de SEO ? Cet article, en deux parties, explique le concept d'un moteur de recherche sémantique et analyse la façon dont Google pourrait bien fonctionner dans les mois qui viennent. Une petite révolution semble bien se préparer dans le domaine du SEO...

Le mois dernier, nous avons initié cet article sur les projets à court terme de Google en tant que moteur de recherche sémantique. Nous avons passé en revue les objectifs d'un tel moteur, ses limites, les premières expériences disponibles en ligne avant d'explorer le fonctionnement de ces outils.

Nous continuons, donc, au travers de cette seconde partie, à explorer la façon dont Google pourrait intégrer la sémantique dans son moteur d'ici quelques semaines...

Le tournant : le rachat de Metaweb

Le rachat de la société Metaweb par Google en juillet 2010 (http://www.abondance.com/actualites/20100719-10404-google-acquiert-metaweb-et-freebase-base-de-donnees-dentites-nommees.html) a clairement accéléré le rythme avec lequel Google a pu préparer son évolution (même partielle) vers un moteur sémantique.

Metaweb est la société éditrice du site Freebase (http://www.freebase.org/), une initiative qui permet à des contributeurs d'enrichir une base de données publiques sur des entités diverses et variées. Lors du rachat de sa maison mère par Google, Freebase contenait déjà des informations sur 12 millions d'entités, soit quatre fois plus que la version anglophone de Wikipedia. Metaweb intéressait également Google pour son savoir-faire dans le domaine de l'indexation des faits, des concepts, et en particulier la découverte des entités nommées. Depuis lors, Google a réussi à augmenter la taille de son graphe de connaissances, pour atteindre un chiffre de 200 millions d'entités nommées.

La fiche de freebase à propos de l'entité nommée « Tour Eiffel » (alias Eiffel Tower en anglais).

Que faire avec des centaines de millions d'entités ?

Plus on sait reconnaître d'entités, mieux on peut répondre aux questions des internautes. Selon une étude de Bing (http://research.microsoft.com/pubs/120889/fp0700-yin.pdf), 20 à 30% des requêtes tapées sur le moteur de Microsoft seraient des entités nommées, et plus de 70% des requêtes contiennent au moins une ou plusieurs entités nommées !

Graphes d'entités contre Graphe social ?

Il existe clairement un enjeu stratégique derrière cette évolution de Google. Jusqu'ici, la matière sur laquelle Google travaillait était une masse de documents (des pages web), mais sans connexion directe avec le monde réel : une page qui parle d'une célébrité n'est pas cette célébrité.

Avec un graphe de connaissances et de concept, on passe dans une autre dimension : on ne stocke plus de l'information sur des documents, mais sur ce dont parle les documents : des lieux, des faits, des gens, des dates, des quantités.

L'analogie avec le graphe social de Facebook est évidemment très forte, sauf que les graphes de connaissances étendent les possibilités liées au stockage, et l'analyse des relations entre individus et leurs habitudes, leurs préférences, à un grand nombre de nouvelles entités.
Il est devenu évident que l'un des futurs terrains de combat pour les majors de l'Internet sera l'information sur les individus, les lieux, les entreprises etc.

L'intérêt de cette approche avec les nouvelles interfaces vocales

L'évolution vers un moteur sémantique présente également un intérêt certain avec l'avènement des interfaces vocales grand public, dont Siri d'Apple constitue l'archétype le plus connu. Les réponses données par Siri exigent de pouvoir identifier des éléments précis correspondant à la question formulée par l'internaute. On pourrait s'interroger sur la qualité actuelle de ces systèmes, et sur le rythme d'adoption des interfaces vocales par les internautes, mais il semble désormais acquis que cette nouvelle manière d'interagir avec son téléphone, sa tablette ou son ordinateur finira par modifier sérieusement la manière dont les internautes cherchent l'information sur internet.

Si l'interface vocale permet de libérer les mains de l'internaute, et lui permet aussi de formuler ses demandes en langage (presque) naturel, la « lecture » des informations en retour est plus lente que pour une interface visuelle. Bien sûr, on peut toujours regarder les résultats sur son écran, mais ces interfaces deviennent vraiment confortables quand toute l'interaction peut se gérer à la voix et à l'oreille. S'agissant d'une recherche d'information, cela sous-entend que l'on puisse évidemment « lire » directement la réponse, et non lire une liste de pages et ensuite plusieurs pages qui contiennent éventuellement la réponse.

L'évolution vers un moteur sémantique se révèle donc particulièrement utile, voire indispensable pour développer des applications vocales efficaces.

Google formule-t-il un réel avertissement ou recherche-t-il sur un effet d'annonce ?

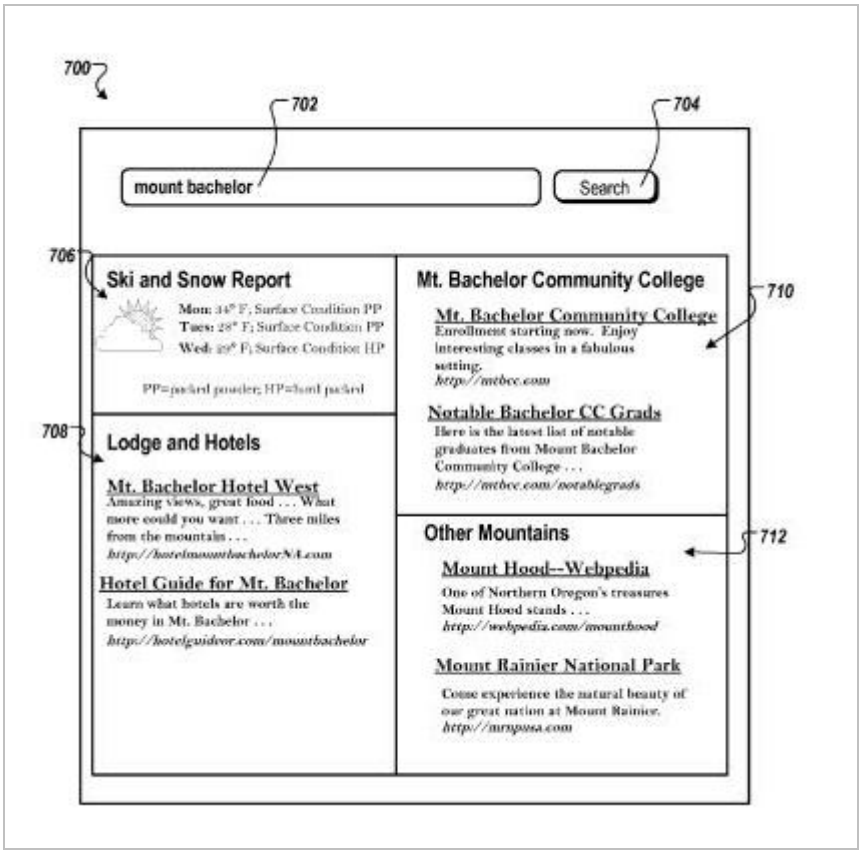
Amit Singhal, dans un message diffusé sur Google+ (https://plus.google.com/115744399689614835150/posts/3vLRVL7C4QS) à la suite de l'article du Wall Street Journal, semblait surpris par les commentaires présentant ce changement comme révolutionnaire. Il s'est clairement attaché à minimiser la portée de ces changements. Il faut dire que si les chiffres d'entités collectées semblent impressionnants, en réalité on est loin du compte. Google annonce avoir collecté des faits et des données sur plusieurs centaines de millions d'entités (200 millions). Pour pouvoir couvrir l'ensemble des questions, la quantité d'entités à couvrir se compte en réalité au minimum en... milliards.

Il semble en fait que l'effet d'annonce n'était pas exactement le but recherché par Amit Singhal, mais qu'il s'agit bel et bien d'une entreprise de communication. Quel que soit l'impact réel de ce changement, il est apparu probablement suffisamment disruptif pour qu'une « préparation » des esprits longtemps à l'avance soit nécessaire. L'introduction brusque de cette fonctionnalité, dans une période où les critiques à l'encontre de Google se font de plus en plus fréquentes et sévères, pouvait s'avérer risquée et surtout, mal comprise.

Un impact limité dans un premier temps ?

Même si le pourcentage de requêtes impactées (10% minimum, et jusqu'à 20%) en ferait peut-être la mise à jour la plus visible depuis la mise en place de la recherche universelle il y a cinq ans de cela, cela reste malgré tout limité (mais sans doute plus impactant que Panda). D'autant plus que l'apparition de « réponses » liées à des entités nommées dans Google n'empêche pas l'affichage d'autres résultats sous forme de « liens bleus ».

En réalité, pour l'instant, rien n'a filtré sur la manière dont ces nouveaux résultats seront affichés. On en est réduit à des conjectures et à l'examen des brevets pour essayer de deviner comment de tels résultats pourraient être affichés. Un brevet attire particulièrement l'attention : ***Identifying Query Aspects*** (voir les references completes dans la bibliographie). On y trouve de nombreuses allusions à Freebase, et la description d'une page de résultats composite présentant des résultats associés à une entité nommée et aux entités « associées ».



Exemple de page de résultat, tirée d'un brevet de Google, présentant des élements (informations, pages web) associées à une entité nommée. On constate des points communs flagrants avec les solutions d'affichage retenues en leur temps par les moteurs Hakia et Cuill.

Le référencement dans un Google sémantique

En découvrant les intentions de Google, beaucoup de webmasters ont réagi très négativement en apprenant que Google allait utiliser les informations extraites des sites web pour fournir directement des réponses. Rappelons que c'est partiellement vrai, car Google envisage aussi

d'utiliser des sources de données issues de fichiers de partenaires et d'organisations gouvernementales (par exemple la base du fameux « World Fact Book » de la CIA). Mais cette perspective a ému suffisamment les webmasters pour que beaucoup se demande s'il fallait laisser Google accéder à ses pages web !

La question n'est pas ridicule pour des sites disposant de données exclusives. Qui plus est, certaines applications des moteurs sémantiques posent des problèmes juridiques nouveaux qui risquent une fois de plus de chambouler la législation sur les droits d'auteurs et les droits voisins avant même que le législateur ait eu le temps de s'en mêler et avant même que les conséquences réelles aient pu être évaluées.

Pour les autres sites, les plus nombreux, bloquer l'accès de ses pages à Googlebot ou gêner son travail relèvera plus de l'acte militant que d'un comportement rationnel.

Il est évident qu'on ne référence pas un site pour un moteur sémantique comme pour un moteur de recherche traditionnel (les moteurs dits « syntaxiques »). La rupture ici est fondamentale. On passe d'un système où l'enjeu était de « rapprocher » une page d'une requête, afin que les mesures de similarité et d'importance produisent une apparition en tête des classements. Dans le contexte d'un moteur sémantique, on ne parle plus de « proximité ou de similarité» mais de « correspondance exacte ».

L'un des poids clés, si l'on a un contenu intéressant sur Napoléon, sera donc de faciliter l'identification exacte de l'entité nommée Napoléon, notamment :

- En utilisant un code source permettant d'identifier facilement la nature des entités nommées (par exemple un des référentiels de Schema.org, la norme proposée par les principaux moteurs de recherche, ou d'autres formats issus du web sémantique). A défaut, la présence d'un code bien structuré facilitera le travail des robots d'indexation.

- En évitant les problèmes de graphies qui empêchent l'identification d'une entité (en sachant qu'il est possible, mais dans une certaine mesure seulement, de reconnaître que *Brittney Spears* est une variante du nom d'une personne unique appelée *Britney Spears*, et qu'il ne s'agit pas d'une entité différente).

- En faisant attention au contexte d'apparition des entités nommées : c'est ce contexte qui autorise une identification correcte :

  * Les informations présentes dans le contexte entourant l'apparition d'un nom de lieu, de personne, doivent être suffisamment riches pour «désambiguïser » efficacement des graphies communes à plusieurs entités.

  * La syntaxe entourant une entité nommée doit être correcte : l'analyse syntaxique poussée de la phrase permet souvent de comprendre la nature d'une entité nommée. En cas de fautes d'orthographe, d'accord, de conjugaison, le résultat est souvent inattendu.

Mais deux remarques s'imposent...

La première est que la reconnaissance des entités nommées est déjà à l'œuvre dans plusieurs fonctionnalités clés de Google, et que les conseils donnés plus hauts sont donc déjà valables !

La seconde est que l'optimisation pour la partie sémantique du moteur va s'ajouter au SEO, comme le SMO s'est déjà ajouté au référencement traditionnel, sans se substituer aux règles d'optimisation habituelles.

Comment se préparer au changement ?

Dans quelques semaines ou quelques mois, le lancement effectif des fonctionnalités annoncées par Amit Singhal mettra fin aux spéculations. On saura enfin à quel point le moteur de recherche actuel sera transformé - ou pas... Et on pourra alors commencer à étudier les solutions retenues par Google, et en tirer des conclusions complémentaires, si nécessaire, sur les optimisations envisageables, et les changements de stratégie à opérer pour s'adapter.

En attendant, les référenceurs vont devoir ajouter une nouvelle corde à leur arc pour s'adapter à cette nouvelle donne : les connaissances en linguistique, et en traitement automatisé du langage. C'était déjà utile quand Google était un pur moteur syntaxique. Mais cela deviendra indispensable pour gérer la mutation progressive de Google en moteur sémantique ...

Bibliographie

Liens vers les articles évoqués plus haut (ce mois-ci ou le mois dernier) :

http://mashable.com/2012/02/13/google-knowledge-graph-change-search/

http://online.wsj.com/article/SB10001424052702304459804577281842851136290.html

Exemples de moteurs sémantiques « en ligne »

Wolfram Alpha : http://www.wolframalpha.com
Hakia : http://www.hakia.com
Cluuz : http://www.cluuz.com

Articles scientifiques divers sur le sujet des moteurs de recherche d'entités nommées

Utilisation de bases de connaissance (ici Wikipedia) pour la résoudre les cas de correspondances ambigües dans le processus de reconnaissance d'entités nommées (article publié par des chercheurs de Google :
Using Encyclopedic Knowledge for Named Entity Disambiguation
http://research.google.com/pubs/archive/70.pdf

Méthode de découverte d'entités nommées par induction:
A Context Pattern Induction Method for Named Entity Extraction
Partha Pratim Talukdar - CIS Department, University of Pennsylvania,Thorsten Brants - Google, Inc, Mark Liberman Fernando Pereira - CIS Department
http://research.google.com/pubs/archive/34509.pdf

Les arbres d'entités conceptuelles:
Knowledge acquisition
Tom Mitchell, Google inc.
http://research.google.com/pubs/archive/74.pdf

Sans surprise, les sujets évoqués dans cet article ont fait l'objet d'une activité de recherché très active chez Google, Bing mais aussi de nombreux laboratoires de recherche, on trouvera donc d'autres articles passionnants sur ces pages :
http://research.google.com/pubs/DataMining.html

Brevets

Identifying Query Aspects
Invented by Fei Wu, Jayant Madhavan, and Alon Halevy
Assigned to Google
US Patent Application 20100198837
Published August 5, 2010
Filed: July 30, 2009
http://www.google.com/patents/US20100198837

Editing a network of interconnected concepts
Inventors:Weissman; Adam J. (Santa Monica, CA), Elbaz; Gilad Israel (Santa Monica, CA)
Assignee:Google Inc. (Mountain View, CA)
Appl. No.:10/748,399
Filed: December 30, 2003, Published November 2011
http://www.google.com/patents/US20040236737

Query rewriting with entity detection
Inventors: Zhou; Hong (Sunnyvale, CA), Bharat; Krishna (San Jose, CA), Schmitt; Michael (Newfalm, DE), Curtiss; Michael (Sunnyvale, CA), Mayer; Marissa (Palo Alto, CA)
Assignee: Google Inc. (Mountain View, CA)
Appl. No.: 10/813,572
Filed: March 31, 2004
http://www.google.com/patents/US20050222977

Entity-Based Search Results and Clusters on Maps
Invented by Joy Ghanekar, Jerry Cheng, Edward Stanley Ott, IV, and Marc Eliot Davis
Assigned to Yahoo!
US Patent Application 20100211566
Published August 19, 2010
Filed: June 23, 2009
http://www.google.com/patents/US20100211566

Creation and Enrichment of Search Based Taxonomy for Finding Information from Semistructured Data
Invented by Sudharsan Vasudevan, Rohan Monga, Hemanth Sambrani, and N S Sekar
US Patent Application 20090282010
Assigned to Yahoo!
Published November 12, 2009
Filed June 18, 2008
http://www.google.com/patents/US20090282010

Disambiguation of Named Entities
Inventors: Razvan Constantin Bunescu and Alexandru Marius Pasca
US Patent Application 20070233656
Published October 4, 2007
Filed: June 29, 2006
http://www.google.com/patents/US20070233656

Implicit name searching
Invented by Rajat Mukherjee, Irfan Presswala, and Kalpana Ravinarayanan
Assigned to Yahoo
US Patent Application 20080228720
Published September 18, 2008
Filed March 14, 2007
http://www.google.com/patents/US7917489

Autres brevets divers de Google sur les sujets évoqués ici :

Detecting Entities and New Words (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20100180199.PGNR.&OS=DN/20100180199RS=DN/20100180199)
Query rewriting with entity detection (http://www.google.com/patents/US20050222977)
Agent Rank (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20110213770.PGNR.&OS=DN/20110213770RS=DN/20110213770)
Query rewriting with entity detection (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20110270863.PGNR.&OS=DN/20110270863RS=DN/20110270863)
Delegating authority to evaluate content (http://www.google.com/patents/US8069175)
Associating an Entity with a Category (http://www.google.com/patents/US20100070339)
Query-independent entity importance in books (http://www.google.com/patents/US20100281034)

Autres brevets de Microsoft

Named Entity Recognition in Query (http://www.google.com/patents/US20110231347)

Web-scale entity relationship extraction (http://www.google.com/patents/US20110251984)
Webpage entity extraction through joint understanding of page structures and sentences (http://www.google.com/patents/US20110078554)
Providing comparison experiences in response to search queries (http://www.google.com/patents/US20110153528)
Applying model of a persona to search results (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20110252014.PGNR.&OS=DN/20110252014RS=DN/20110252014)
Scalable Incremental Semantic Entity and Relatedness Extraction from Unstructured Text (http://www.google.com/patents/US20110264997)
Identifying location names within document text (http://www.google.com/patents/US20110264664)
Comparisons of entities of a particular type (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=/netahtml/PTO/search-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20110270864.PGNR.&OS=DN/20110270864RS=DN/20110270864)
Entity category determination (http://www.google.com/patents/US20110314018)
Named Entity transliteration using corporate corpora (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20100106484.PGNR.&OS=DN/20100106484RS=DN/20100106484)
Entity-specific tuned searching (http://www.google.com/patents/US7739270)
Entity-specific search model (http://www.google.com/patents/US20080005076)
Identifying synonyms of entities using web search (http://www.google.com/patents/US20100293179)
Identifying synonyms of entities using a document collection (http://www.google.com/patents/about/12_478_120_IDENTIFYING_SYNONYMS_OF_ENTIT.html?id=mbDlAAAAEBAJ)

Brevets de Yahoo !

System and method for adding identity to web rank (http://www.google.com/patents/US7840548)
Scalable semi-structured named entity detection (http://www.google.com/patents/US8073877)
Extracting entities from a web (http://www.google.com/patents/US20090182759)
Mining knowledge sources for improved entity extraction (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20110022598.PGNR.&OS=DN/20110022598RS=DN/20110022598)
Mining knowledge sources with auto learning for improved entity extraction (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20110022550.PGNR.&OS=DN/20110022550RS=DN/20110022550)
Implicit Name Searching (http://www.google.com/patents/US7917489)
Large scale entity-specific resource classification (http://www.google.com/patents/US20110264651)

***Philippe YONNET** , Directeur SEO international, Twenga.*

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://www.abondance.com/abonnes/blogpro/