Sémantique et moteur de recherche : l'exemple du moteur d'Orange

Retour au sommaire de la lettre

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

La notion de sémantique est aujourd'hui sur toutes les lèvres lorsqu'on parle de moteur de recherche. Google a fait plusieurs annonces à ce sujet (voir nos articles par ailleurs dans cette lettre à ce sujet). Mais l'équipe qui conçoit le moteur de recherche d'Orange travaille également depuis de nombreux mois sur cette thématique, notamment en se basant sur les données issues de Wikipedia. Ces équipes nous décrivent ici les algorithmes conçus pour analyser et "comprendre" les questions posées par les internautes en langage naturel et les mécanismes mis en place pour fournir une réponse qui soit la plus pertinente possible...

On parle de plus en plus de sémantique lorsqu'est abordé le sujet des moteurs de recherche sur le Web. C'est peut-être oublier que celle-ci est présente depuis longtemps dans les algorithmes utilisés en ligne pour mesurer la pertinence d'une requête.

Ainsi, dès Février 2010, le site d'Orange "Le Moteur" (http://www.lemoteur.fr/) a introduit en tête de sa page de résultats une réponse sémantique lorsque la requête de l'internaute est posée sous la forme d'une question. Cette première version de réponse sémantique dans le moteur Orange est décrite dans l'article du 12 Février 2010 disponible à l'adresse http://www.abondance.com/actualites/20100212-10243-le-moteur-dorange-integre-la-semantique-aux-recherches-geographiques.html.

L'objectif avec cet outil, que nous appellerons par la suite "moteur de question/réponse", était d'introduire une rupture d'usage à la classique réponse sous la forme d'une liste de documents, en apportant directement à l'internaute une réponse à sa question. Aujourd'hui, la rupture d'usage entrevue il y a deux ans trouve naturellement sa place car depuis début 2010 deux faits majeurs sont venus conforter cet usage :

- l'arrivée à maturité des techniques de reconnaissance vocale, banalisées aujourd'hui dans le grand public avec la sortie de Siri pour l'Iphone 4S,

- l'allongement de la taille des requêtes tapées sur les différents moteurs de recherche (voir l'article du 26 Janvier 2012 : http://www.abondance.com/actualites/20120126-11246-plus-de-4-mots-cles-en-moyenne-dans-les-requetes-moteurs.html). Le cercle vertueux qui semble s'amorcer fait que plus les requêtes vont en s'allongeant, plus les requêtes contiennent du contexte, et plus les réponses des moteurs peuvent être pertinentes s'ils savent tenir compte de ce contexte. Ceci incitera les internautes à continuer à allonger leurs requêtes, etc.

Présentation du moteur de question/réponse Orange

L'approche disruptive de réponse directe aux questions posées dans les requêtes donne au moteur Orange de nouvelles perspectives. Ce nouveau moteur sémantique s'additionne au moteur classique afin de composer la page de réponse finale du moteur du portail Orange. Par exemple à une requête comme "*nom des habitants de Beaulieu dans l'Ardèche*", le moteur

Orange affiche en première réponse "*Belliloquois*" même si la requête comporte des erreurs comme dans la figure ci-dessous. Nous reviendrons sur le traitement linguistique de la requête plus loin.

L'utilité d'une telle fonction est évidente. En trouvant directement la réponse à sa question, l'utilisateur gagne du temps et est encouragé à poser d'autres questions. En effet, dans d'autres moteurs de recherche qui ne possèdent pas cette fonction, l'utilisateur est censé rechercher sa réponse dans les extraits cités des documents proposés dans la page des réponses ou d'ouvrir un par un les documents pour éventuellement trouver la réponse, mais sans aucune garantie. Ce parcours est souvent fastidieux et inconfortable pour l'utilisateur. Cette pénibilité est encore plus grande lorsque le terminal est petit comme un téléphone mobile.

Un tel moteur de question/réponse a été déployée depuis 2010 dans le moteur de recherche du portail Orange (http://www.lemoteur.fr, http://www.voila.fr et http://www.orange.fr) pour répondre à des questions relatives à des lieux, des personnes, des épreuves sportives (une aide à l'utilisation de cette fonction est accessible sur les pages de l'assistance Orange : http://assistance.orange.fr/la-recherchegeographique-avec-orange-et-wikipedia-3039.php. Ce moteur peut répondre directement (en se basant sur le contenu de l'encyclopédie en ligne Wikipedia) à des questions portant sur plusieurs centaines de milliers de sujets désignant :

- des personnes :

- des lieux :

- des épreuves sportives :

Une architecture technique adaptée

En se projetant à plus long terme dans la continuité de ce nouvel usage, l'équipe du moteur Orange a mis en place une architecture intégrant un moteur d'interprétation linguistique de requêtes et une base de connaissances. Le premier a pour rôle l'identification de la présence d'une question dans la requête et la transformation de cette dernière (si elle est détectée) en une requête structurée.

La base de connaissances a pour rôle d'agréger un ensemble le plus complet possible de connaissances permettant de répondre à un très grand nombre de questions.

L'interprétation linguistique de la requête est complexe. Elle doit prendre en compte toutes les formes de requêtes possibles pour poser une même question y compris celles qui contiennent des erreurs. Elle doit aussi filtrer toutes les requêtes qui ne nécessitent pas une réponse directe et éviter de faire des contre-sens toujours possibles pour une moteur aussi généraliste, auquel on peut tout demander. Il faut aussi pouvoir désambigüiser certains termes lorsque ceux-ci sont explicitement précisés dans la requête.

La question de la complétude de la base de connaissances est aussi complexe. Elle est fondée sur une alimentation intelligente à partir de plusieurs sources importante de données. Une source importante est le contenu encyclopédique Wikipedia. En 2009, Orange a signé un partenariat avec la fondation Wikimedia (http://fr.wikipedia.org/wiki/Wikimedia_Foundation#Partenariat_avec_Orange). Dans ce cadre le contenu Wikipedia est utilisé pour alimenter la base de connaissances. Par ailleurs, l'objectif est d'utiliser les contenus mis à disposition par des initiatives « open data » (http://www.data.gouv.fr/) qui encouragent ce type de pratique ainsi que ceux des partenaires d'Orange afin de constituer un capital de données attractif.

L'interpréteur linguistique de requêtes et la base de connaissances se basent sur des technologies innovantes que sont :

- **Le Traitement Automatique des Langues Naturelles** (TALN ou NLP pour Natural Language Processing en anglais). Cette technique permet d'analyser, de comprendre et de traduire les requêtes pour les agents logiciels du moteur. Ces techniques sont d'autant plus importantes que la requête est longue et contient du contexte. L'équipe du moteur Orange s'est doté d'un outil de TALN performant développé depuis plusieurs années par Orange Labs à Lannion (cf : http://www.atala.org/IMG/pdf/TAL-2008-49-2-01-Heinecke.pdf).

- **Les langages, protocoles et outils du Web Sémantique.** L'utilisation croissante de connaissances dans nos différentes briques de recherche a rapidement fait ressortir le besoin de représenter, lier et échanger les descriptions élémentaires associées provenant à la fois de sources internes et externes (Wikipedia, data.gouv.fr, etc.).

Nous nous sommes donc orienté vers les technologies du Web Sémantique, standardisés par le W3C, qui proposent :
(1) un langage de description de ressources, RDF (http://www.w3.org/TR/rdf-primer) (*Resource Description Framework*),
(2) Des langages de description de modèles flexibles, RDFS (http://www.w3.org/TR/rdf-schema/) (*Resource Description Framework*) et OWL (http://www.w3.org/TR/owl-overview/) (*Ontology Web Language*), ainsi

(3) qu'un protocole et un langage de requêtes pour échanger ces descriptions entre elles sur le réseau.

Courant 2011, la réflexion a donc évolué pour passer d'une base de connaissances relationnelle classique vers une base de connaissances implémentant les technologies du Web Sémantique afin de lier et enrichir les données fournies par les différents services et partenaires d'Orange.

Dans la suite de cet article, nous détaillons l'utilisation des ces deux techniques que sont le TALN et le Web Sémantique dans le cadre du développement du moteur de question/réponse Orange.

Un moteur linguistique à base de connaissances

Nous verrons par la suite que la base de connaissances utilisée pour trouver les réponses s'interroge dans un langage informatique appelé SPARQL. Cependant l'internaute pose sa question en langage naturel. Il n'est pas d'actualité de demander à l'internaute « grand public » de saisir ses questions en SPARQL. Au contraire, les utilisateurs ont pris l'habitude de taper leur requête au sein d'une boite de recherche unique sans devoir sélectionner différentes options. Avec les capacités grandissantes d'interprétations des moteurs, ces utilisateurs formuleront des requêtes de plus en plus longues, surtout lorsque certains moteurs offrent déjà des interfaces vocales très performantes. Le challenge est donc maintenant d'être capable de traduire la requête naturelle humaine vers une requête SPARQL comprise par une machine.

Cette traduction est possible grâce à une interprétation linguistique précise des requêtes et des connaissances sémantiques associées à chaque terme de la requête comme cela est décrit par les équipes d'Orange Labs dans ce document : http://asso-aria.org/coria/2011/207.pdf

La traduction s'effectue même si les requêtes comprennent des fautes comme nous l'avons constaté dans la question citée plus haut "*nom des habitands de Beaulieu dans l'ardèche*". Si la correction de requêtes est aujourd'hui très souvent proposée dans les principaux moteurs de recherche web couvrant un large vocabulaire, ceci est plus difficile lorsque l'on traite un domaine particulier alors que les requêtes peuvent être hors de ce domaine. Il est en effet crucial de ne pas chercher systématiquement à corriger ou à traduire toutes les requêtes soumises au moteur vers une requête pour laquelle le moteur de question/réponse a une réponse. Dans le cas contraire, les contre sens pouvant être faits sont rédhibitoires.

Une autre particularité est l'ensemble des formes de questions que le moteur peut gérer. Pour évaluer cela, nous l'avons comparé avec un moteur utilisant un Système de Gestion de Base de Données (SGBD) classique qui recherche simplement dans un index la requête soumise. Chaque entrée de cet index correspond à une concaténation de noms de propriété et de sujet ou de sujet puis de propriété tels qu'ils sont trouvés dans la base de connaissances. Pour les deux systèmes évalués, nous nous sommes restreints au domaine de la géographie. En leur soumettant un fichier log de plus de 10 millions de requêtes réelles, le moteur de question/réponse affiche une réponse pour quatre fois plus de questions avec un taux de précision supérieur à 98%.

Enfin le moteur de question/réponse est aussi capable de tenir compte de l'ensemble des mots de la requête pour la désambigüiser. Ainsi, la requête "*nom des habitants de beaulieu dans l'ardèche*" aura une réponse précise pour la ville de Beaulieu dans l'Ardèche, alors que la requête " *nom des habitants de beaulieu*" peut référer à au moins à onze villes françaises. Un traitement approprié est dans ce cas nécessaire.

Exemple de réponse quand la question est ambigüe :

Le Web Sémantique dans le cadre du moteur de question/réponse Orange

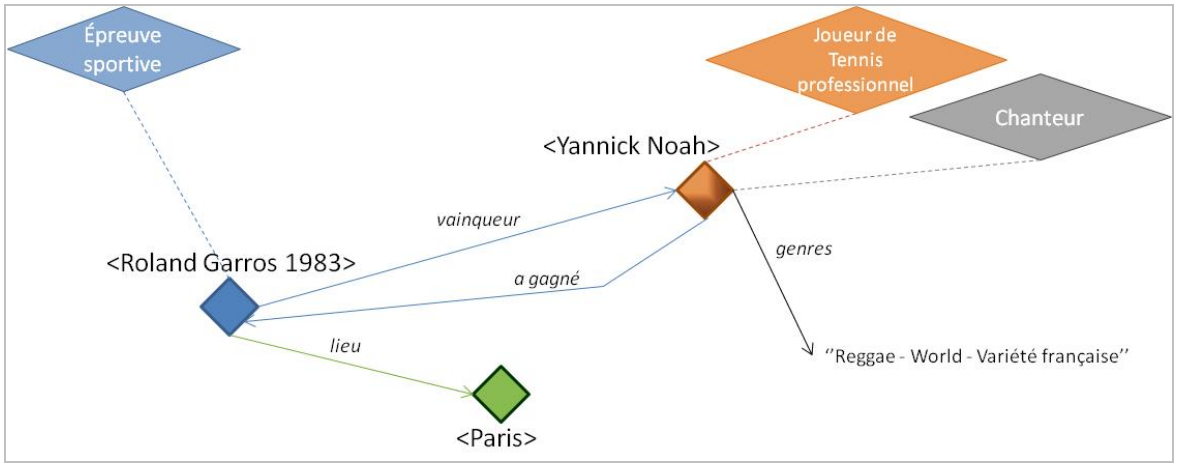
La base de connaissances sur laquelle s'appuie le moteur de question/réponse est constituée de triplets RDF (RDF est un acronyme désignant le standard « Resource Description Framework » - http://www.w3.org/TR/rdf-primer - du W3C) , *<sujet> <prédicat> <objet>*, qui à l'instar du *<sujet> <verbe> <objet>* de simple phrase en langage naturelle permettent de décrire n'importe quelle entité. Le *<sujet>* est nécessairement une entité alors que l'<objet> peut être soit une entité soit une valeur littérale (entier, chaine de caractère, etc.). Chaque entité est identifiée de manière uniforme par une URI ce qui permet de lier automatiquement les triplets mêmes s'ils sont distribués sur le réseau, ce qui permet de passer de silos de données à un graphe global de connaissances.

L'illustration ci-dessous donne un aperçu schématique de la modélisation de la connaissance nécessaire pour répondre à la requête "*qui a gagné roland garros ?*".

Elle montre notamment :

- Que le sujet <Roland Garros 1983> est de type « Epreuve sportive »
- Que le sujet <Yannick Noah> est de type « Joueur de Tennis professionnel »
- Que le sujet <Yannick Noah> est également de type « Chanteur »
- Que le lien "vainqueur" lie le sujet <Roland Garros 1983> au sujet <Yannick Noah>, ces trois entités forment ainsi un triplet : <Roland Garros 1983> vainqueur <Yannick Noah>. Le lien est une propriété ou un prédicat (les deux mots sont utilisés) qui caractérise le sujet.
- Qu'il existe dans la base de connaissances, de la même manière, un deuxième triplet : <Yannick Noah > a gagné <Roland Garros 1983>
- Que concernant le sujet <Yannick Noah> il existe un autre triplet qui caractérise son activité de « Chanteur » et qui donne l'information suivante : <Yannick Noah > genres "Reggae - World - Variété française''

A la différence des deux triplets cités précédemment, la troisième partie de ce triplet n'est pas elle-même un sujet mais est une valeur simple de type "chaîne de caractères". C'est-à-dire qu'il ne peut pas y avoir de liens du graphe qui partent depuis cette valeur. Nous venons de découvrir que l'extrémité finale d'un triplet (l'<objet>) peut être soit un <sujet> soit une <valeur>, auquel cas le graphe atteint une de ses extrémités à partir de laquelle il ne peut plus s'étendre. *A contrario*, depuis un <objet> de type <sujet>, le graphe peut continuer à s'étendre et à connecter/collecter de la nouvelle connaissance. Par exemple l'extrémité du triplet <Yannick Noah > a gagné <Roland Garros 1983> est le sujet <Roland Garros 1983> qui lui-même devient l'origine d'un nouveau triplet <Roland Garros 1983> lieu <Paris>.



Le moteur de question/réponse Orange tire parti des triplets de la base de connaissances pour fournir les réponses sémantiques. Ainsi, le nouveau triplet <Roland Garros 1983> lieu <Paris> permet au moteur de question/réponse, de fournir une nouvelle réponse sémantique à la question "*Dans quel lieu est Roland Garros 1983 ?*"

Dans le schéma précédent, nous pouvons distinguer deux types de connaissances :

- Le premier correspond aux connaissances entre les nœuds du graphe que sont les entités <Roland Garros 1983> ou <Yannick Noah> ou "Reggae - World - Variété française". Il s'agit de connaissances factuelles sur ces entités. Ce niveau est souvent appelé « couche RDF ».

- Le deuxième, que l'on appelle ontologique, associe une classe comme <Epreuve sportive >, <Chanteur> et <Joueur de Tennis professionnel> aux nœuds du graphe. Il s'agit d'un méta-niveau par rapport au niveau précédent. Ce niveau utilise les langages du Web Sémantique nommés RDFS et OWL . Ils permettent de structurer et de classifier la connaissance dans un domaine donné. Dans le code couleur utilisé ci-dessus, le code couleur bleu et orange correspondent à un domaine lié au sport, le code couleur gris au domaine des personnalités et le vert à la géographie. Ce schéma nous indique que les classes qui composent le modèle ontologique sont aussi des nœuds du graphe et peuvent donc être décrit avec les mêmes mécanismes que n'importe quelle autre entité.

Accès à la Base de Connaissance

Le W3C standardise également un langage de requête qui permet de questionner un graphe de connaissance. Ce langage nommé SPARQL (http://www.w3.org/TR/sparql11-query/) est utilisé par le moteur de question/réponse Orange pour préparer les réponses aux questions des internautes. Le W3C propose un mode d'accès à une base sémantique, basé sur le protocole HTTP, sous la forme d'un « *sparql endpoint* » (http://www.w3.org/wiki/SparqlEndpoints). Par exemple le « *sparql endpoint* » de la base sémantique de Dbpedia est public, il est accessible à l'url http://dbpedia.org/sparql.

Une requête SPARQL se base sur le format de triplets du graphe de connaissance. L'exemple de requête SPARQL ci-dessous extrait de la base Dbpedia les artistes nés à Nice ainsi que leurs dates de naissance. Le format de sortie est HTML. Plusieurs formats de sortie sont prévus pour

les « *sparql endpoint* » comme, par exemple, les formats XML ou JSON qui s'avèrent très pratiques pour incorporer des données sémantiques dans une page web (*mashup*).

```
SELECT ?name ?birth WHERE
{
 ?person <http://dbpedia.org/ontology/birthPlace>
<http://dbpedia.org/resource/Nice> .
 ?person <http://dbpedia.org/ontology/birthDate>  ?birth .
 ?person rdfs:label                               ?name  .
 ?person rdf:type
<http://dbpedia.org/ontology/Artist> .
 FILTER (lang(?name) = "fr")
}
ORDER BY ?name
```

Cette requête donne le résultat suivant, le format d'affichage est le format par défaut, c'est à dire HTML :

name	birth
"François Corteggiani"@fr	1953-09-21
"J. M. G. Le Clézio"@fr	1940-04-13
"Joann Sfar"@fr	1971-08-28
"Michèle Mercier"@fr	1939-01-01
"Philippe Bercovici"@fr	1963-01-05
"Yves Klein"@fr	1928-04-28
"Éric Fottorino"@fr	1960-08-26

Une requête SPARQL est formée de patrons de triplets qui sont recherchés dans le graphe de connaissance. Les variables recherchées commencent par le caractère '?'. Le point final à chaque triplet de la requête signifie un ET logique. En lisant les cinq lignes de la requête de haut en bas, on recherche :

Les triplets : ?person né à <Nice>
ET les triplets : ?person né le ?birth
ET les triplets : ?person ayant un nom ?name
ET les triplets : ?person de type Artist
ET filtrer les noms de langue française

L'instruction SELECT précise les variables que l'on veut garder pour la sortie, c'est-à-dire dans notre exemple *?name* et *?birth*.

L'instruction ORDER BY réalise un tri, en l'occurrence dans notre exemple le tri sur les étiquettes *?name*.

Cette requête SPARQL peut être essayée dans un navigateur en se positionnant sur l'url du « sparql endpoint » dbpedia soit http://dbpedia.org/sparql et en coupant-collant la requête SPARQL donnée ci-dessus. La base dbpedia est très complète mais un de ces inconvénients majeur est qu'elle est spécialisée pour les données américaines et anglaises du site Wikipedia.

Il existe un « *sparql endpoint* » pour la base de connaissances du moteur Orange qui est spécialisé pour les données françaises. Celui-ci est pour le moment réservé à un usage interne mais il n'est pas exclu qu'il devienne un jour ouvert au public.

Perspectives

Nous avons détaillé l'avantage pour le moteur Orange de bénéficier de l'apport d'un moteur de question/réponse. Cet apport passe par la maîtrise de deux technologies très complémentaires que sont les technologies du Web Sémantique et celles du TALN. L'approfondissement de ces technologies ouvre également de nombreuses perspectives comme nous l'avons laissé entrevoir tout au long de cet article :

- L'extension de la couverture thématique de notre moteur de question / réponse, pour l'instant réservé aux domaines géopolitique, sport et people, va profiter de la démarche de publication et de partage de données publiques (cf initiatives « open data » de type http://www.data.gouv.fr et http://datalift.org).

- Au-delà de la problématique d'agrégation des informations et des données qui permettent déjà de tirer le meilleur parti des multiples sources à notre disposition, nous travaillons actuellement sur des techniques d'inférence. Par exemple, à partir de ces deux triplets présents dans la base de connaissances :
<Karim Benzema> joue au <Real de Madrid>,
< José Mourinho> est l' entraineur du < Real de Madrid>
Il est possible d'inférer la nouvelle information :
< José Mourinho> est l'entraineur de <Karim Benzema>
Par ce même mécanisme, notre base de connaissance s'enrichira de nouvelles informations.

- La démocratisation de la « sémantique » dans le web avec HTML5 et *via* l'utilisation des tags sémantiques donnera de plus en plus de matière à notre base de connaissances (voir http://events.linkeddata.org/ldow2012/papers/ldow2012-inv-paper-2.pdf). Nous observons le même type d'évolution sur le web francophone grâce à notre observatoire interne. Il sera de plus en plus crucial de savoir interpréter les données structurées du web pour améliorer la qualité des réponses.

- Les usages liés aux nouveaux terminaux confortent la position sémantique du moteur. On constate par exemple sur nos moteurs mobiles que les requêtes s'allongent et s'approchent du langage naturel. Sur la télé, on peut imaginer de nouveaux usages liés à l'enrichissement de l'information affichée à l'écran par des accès aux données structurées liées.

- Même si nous sommes encore loin de pouvoir traiter l'ensemble des questions pouvant être posées en français sur n'importe quel sujet, notre volonté est d'enrichir notre interpréteur linguistique de requêtes pour pouvoir supporter les formes les plus courantes des questions que l'on trouvera dans nos logs de requêtes. Si aujourd'hui nous sommes très schématiquement limité à des forme de type prédicat sujet (*Quelle est la <population> de la <France> ?*), nous envisageons, par exemple, de pouvoir répondre à des questions de la forme :

- Quel est le plus grand pays d'Europe ?
- Est-ce que Tony Parker est plus grand que Joakim Noah ?
- Quels sont les artistes nés à Nice ?

Mais comme toujours, nous serons guidés par les usages envisagés et constatés. Ce seront donc nos utilisateurs les véritables initiateurs des futurs changements sur nos moteurs de recherche.

Par Guillaume Viland **(Leader Technique du projet Base de Connaissances du moteur Orange :** guillaume.viland@orange.com**), Michel Plu** **(Ingénieur de recherche senior en services d'accès à des contenus multimédia à Orange Labs Lannion :** michel.plu@orange.com**) et Séverine Aubry** **(Responsable de l'équipe de développement moteur web :** severine.aubry@orange.com**). Sur Twitter : @lemoteur**

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