Moteur de questions réponses à partir de données du web sémantique

Michel Plu

France Telecom Orange Labs Lannion
Technopole Anticipa
2 avenue Pierre Marzin
22300 Lannion CEDEX
michel.plu@orange-ftgroup.com

1 Introduction

Une nouvelle tendance des moteurs de recherche sur le web est d'enrichir leur liste réponses en répondant directement aux questions posées dans les requêtes des utilisateurs. Par exemple à une requête comme "nom des habitants de Beaulieu dans l'Ardèche", un tel moteur de recherche affiche en première réponse "Belliloquois" même si celle-ci comporte des erreurs comme dans la figure ci-dessous.

Figure 2: Réponse à une question avec des erreurs sur le moteur orange.fr

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 sans cette fonction appelée par la suite moteur de questions réponses, l'utilisateur est censé rechercher sa réponse dans les extraits cités des documents proposés dans la liste réponses ou les parcourir pour éventuellement la trouver mais sans aucune garantie. Ce parcours est le plus souvent fastidieux et inconfortable pour l'utilisateur. Cette pénibilité est encore plus grande lorsque le terminal est petit comme un téléphone mobile ou lorsque le débit de la connexion réseau n'est pas suffisant pour télécharger rapidement les documents en réponse.

Une telle fonctionnalité de questions réponses a été déployée dans le moteur de recherche des portails www.lemoteur.fr, www.voila.fr et www.orange.fr pour répondre à des questions relatives à des lieux et à des personnes (cf. http://assistance.orange.fr/la-recherche-geographique-avec-orange-et-wikipedia-3039.php et http://assistance.orange.fr/la-recherche-biographique-avec-orange-et-wikipedia-3038.php). Ce moteur peut répondre à des questions portant sur plus de 160 000 sujets désignant des personnes ou des lieux. La démonstration

présentée est une extension de ce moteur à un ensemble de connaissances plus vaste (plus de 13 millions de relations) exportées sous formes de triplets RDFS (cf. http://www.w3.org/TR/rdf-schema/) et dont la sémantique des classes et des propriétés utilisées peut être décrite en OWL (cf. http://www.w3.org/2004/OWL/). Le moteur démontré supporte aussi plus de formes de questions grâce à l'utilisation du langage de requêtes SPARQL (cf http://www.w3.org/TR/rdf-sparql-query/) pour interroger la base de connaissances. Nous utilisons l'extension sémantique du SGBD oracle pour traiter les requêtes SPARQL (cf.http://www.oracle.com/technetwork/database/options/semantic-tech/index.html). La base de connaissances à été initialisée en chargeant les fichiers de l'ontologie du projet DBpedia (cf www.dbpedia.org) et leur extraction de valeurs de propriétés dans les pages françaises de Wikipedia (cf www.wikipedia.org).

2 Un moteur linguistique à base de connaissances

La particularité de ce moteur de questions réponses réside dans sa capacité à répondre à de multiples formes de requêtes qui ne sont pas traitées par les autres moteurs de recherches majeurs du web sur un vaste ensemble de sujets.

Tout d'abord il est capable de répondre à des questions même si celles-ci comprennent des fautes. 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 applicatif particulier alors que les requêtes peuvent être hors de ce domaine. Il est en effet crucial de ne pas chercher systématiquement à corriger toutes requêtes soumises au moteur vers une requête pour laquelle le moteur de question réponses à une réponse. Dans le cas contraire, les contre sens pouvant être faits sont rédhibitoires. Pour éviter cela, les corrections proposées doivent tenir compte du contexte de chaque mot corrigé en vérifiant des contraintes syntaxiques et sémantiques. Si le contexte n'est pas suffisant pour valider une correction, la correction n'est pas proposée afin de ne pas prendre de risque.

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 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 connaissance. Pour les deux systèmes évalués nous nous sommes restreints au domaine de la géographie. En leur soumettant un log de plus de 10 millions de requêtes réelles, notre moteur affiche une réponse pour quatre fois plus de questions avec un taux de précision supérieur à 98%.

Enfin notre moteur est aussi capable de tenir compte de l'ensemble des mots de la requête pour les désambigüiser. Ainsi la requête "nom des habitants de beaulieu dans l'Ardèche" aura la bonne réponse alors que la requête " nom des habitants de beaulieu" peut référer à au moins à onze villes françaises (cf http://fr.wikipedia.org/wiki/Beaulieu).

Ces capacités sont possibles grâce à une interprétation linguistique précise des requêtes et des connaissances sémantiques associées à chaque terme de la requête comme cela est décrit dans Plu et Heinecke (2011). Une bibliographie plus générale des interfaces de bases de données en langage naturelle peut être trouvée dans Androutsopoulos et al (1995). Pour

développer cet interpréteur nous avons utilisé la plateforme logicielle TILT (Heinecke et al 2008).

3 Vers un moteur de recherche du web sémantique

Ce type de moteur préfigure ce que devra être un moteur de recherche du web sémantique tel qu'il se dessine aujourd'hui où de multiples bases d'informations seront accessibles et publiées dans des formats permettant leur intégration et leur compréhension (Nigel Shadbolt, Wendy Hall, Tim Berners-Lee 2006). Mais pour que celles-ci soient retrouvées facilement, il est illusoire de penser que les usagers sauront tous utiliser les langages de requêtes spécifiques ou auront la discipline de remplir des champs pré formatés dans des interfaces utilisateurs ad hoc. 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. 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.

C'est ce mode naturel qu'il faudra supporter pour interroger le web sémantique comme nous proposons de le démontrer dans un domaine encore relativement restreint.

Références

Androustsopoulos I., Ritchie G.D., and Thanish P. (1995). Natural Languages Interfaces to Databases – An introduction. In Natural Langugae Engineering, vol 1, part 1, pages 29-81.

Heinecke J., Smits G., Chardenon C., Guimier De Neef E., Maillebuau E., Boualem M., «TiLT :plate-forme pour le traitement automatique des langues naturelles » (2008). *TAL, vol. 49 :2*, p. 17-41.

Plu, M., Heinecke, J. (2011). Moteur de questions réponses pour une base de connaissances. *Actes de la 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances* Brest.

Nigel Shadbolt, Wendy Hall, Tim Berners-Lee (2006). "The Semantic Web Revisited". *IEEE Intelligent Systems 2006; 21(3): 96-101*.

Summary. This demonstration presents a question answering system with linguistic features. This enables the correct processing of questions having multiple linguistic forms including some misspelling which are not supported by similar systems with the same scale. Provided answer comes from an RDF and OWL knowledge base. An optimized and restricted version of this system has been integrated and is now accessible from French web search engine for casual users (www.lemoteur.fr, www.voila.fr www.orange.fr). This web search engine has the third biggest internet audience in France.