MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

Interrogations de moteurs de recherche par des requêtes formulées en langage naturel

Ludovic Bonnefoy, Romain Deveaud et Eric Charton[1]

1 : LIA / Université d'Avignon, 339 chemin des Meinajaries, 84911 Avignon
Contact : ludovic.bonnefoy@etd.univ-avignon.fr,
romain.deveaud@etd.univ-avignon.fr, eric.charton@univ-avignon.fr

Résumé
L'utilisation de requêtes écrites en langage naturel est un des enjeux important du secteur de la recherche d'information pour les prochaines années. Malgré le fait que ce domaine commence à être couvert par quelques grands noms de l'informatique, les réalisations disponibles à ce jour ne peuvent être considérées que comme des prototypes compte tenu des résultats actuellement visibles. Ceci peut-être expliqué par la grande diversité du langage naturel et par les nombreux sens qui peuvent être donnés à un même mot ou une même expression. Les différents critères sémantiques naturellement présents dans les phrases permettent notamment de combiner la recherche d'information classique - extrayant des documents comparés à des termes, ou mots-clés - avec des « entités » ou des « concepts » pouvant être apparentés à des catégories sémantiques (par exemple : « personne »,« organisation »,« date »...). Nous proposerons dans cet article un système de catégorisation et d'extraction de mots-clés à partir de phrases formulées en langage naturel.

Abstract
Nowadays, requesting a search engine with natural language requests is a significant issue in the information retrieval research field, and some of its biggest actors begin to take it seriously. Some prototypes are actually available, but the error rate, inferred by the huge diversity of natural language and the different semantics of words or expressions, is still too large. Sentences naturally contain semantic criteria such as "entities", "concepts" or "categories" which can be combined with standard information retrieval in order to filter the documents with these semantic categories (e.g. "person", "organisation", "date"...). In this article we propose a categorization and keywords extraction system for natural language sentences.

Mots-clés : Sémantique, catégorisation, langage naturel, recherche d'information.

Keywords: Information retrieval, semantic, categorization, natural language.

1. Introduction

De nos jours, les moteurs de recherche sont un outil pleinement utilisable par les personnes averties et habituées, malgré la volonté d'en améliorer l'accessibilité. En effet, le processus consistant à passer d'une interrogation à un enchainement de mots-clés pertinents retranscrivant correctement la pensée initiale n'est pas un exercice aisé, du moins pour les personnes qui ne sont pas familières avec l'informatique ou internet, parmi lesquelles nous pouvons par exemple compter les enfants ou certaines personnes âgées.
Il serait en effet idéal de pouvoir simplement formuler une question à un moteur de recherche et que celui-ci puisse donner la réponse ou du moins un ensemble de documents dans lesquels une ou des réponses se trouveraient. Des sociétés telles que Google[1], Powerset[2] (propriété de Microsoft) ou Hakia[3] sont actuellement fortement investies dans le développement de solution

[1] http ://www.google.com
[2] http ://www.powerset.com
[3] http ://www.hakia.com

sémantiques à l'interrogation des moteurs de recherche et l'enrichissement communautaire en est un élément central, au moins pour les deux premiers protagonistes. Ils ont en effet recours à de nombreux sites dont les informations sont éditées par des internautes contributeurs, Wikipédia[4] étant le plus célèbre d'entre eux.
C'est également notre cas, puisque le système que nous proposons s'interface NLGbAse[5], une base de données classifiées provenant de Wikipédia qui peut être interrogée par le biais de trois moteurs de recherche différents. Le premier d'entre eux met en $\frac{1}{2}$uvre un algorithme calculant la *similarité cosinus* entre l'ensemble des mots-clés entrés et les documents issus de Wikipédia ; le deuxième est semblable au premier en tous points, à l'exception que l'on peut affiner la recherche en précisant une catégorie, ainsi seuls les documents classifiés comme appartenant à la catégorie spécifiée seront relevés. Le troisième applique quant à lui un algorithme de compacité[6] permettant de trouver une entité précise étant d'une catégorie donnée, proche d'un ou plusieurs mots donnés, dans le document Wikipédia se rapportant à une entité nommée donnée, ce qui permet notamment de pouvoir proposer une réponse factuelle à une requête.
Ces outils constituent le système de recherche d'information sur lequel nous appliquons les sorties de notre propre système ; ce dernier peut, à partir d'une phrase en langage naturel - de préférence une question, donner les différents mots-clés et catégories attendus par les moteurs de recherche de NLGbAse, et ainsi obtenir une liste de résultats - et éventuellement des réponses factuelles - pertinents suite à une interrogation en langage naturel.

2. Catégorisation des phrases

2.1. Utilisation de règles

Notre approche pour catégoriser les questions fonctionne avec un ensemble de règles. Cet ensemble est relativement restreint car nous avons pu remarquer qu ?avec une dizaine de règles environ on pouvait couvrir une majorité des cas mais qu'ensuite chaque petit gain se traduisait par la production d'un nombre croissant exponentiellement de nouvelles règles. Les règles que nous avons formulées se basent sur les pronoms interrogatifs des questions. En voici la liste : Who, Whom, Whose -> pers Where, Whence, Wither -> Si un nom propre ou un objet est trouvé à la question alors la catégorie sera celle du nom propre ou de l'objet (ex : où à étudié Patrick Sébastien, il y a peu de chance de trouver dans la fiche de ce lieu une mention de cette personnalité et il est à priori plus judicieux de proposer sa fiche à l'utilisateur). Dans le cas contraire -> loc. How -> On à la même chose que pour Where à l'exception près que si aucune des deux premières conditions n'est satisfaite alors on aura unk Pour ce pronom là nous avons ajouté quelques précisions : si on a far,few,great,little,many,much,tall,wide,high,big,old on attribut la valeur amount. Très bientôt cela sera affiné afin de correspondre aux conditions d'Ester.
What -> Même principe avec pour valeur par défaut unk Là aussi nous avons précisé certains mots pouvant suivre et déterminer la catégorie : day,month,... qui donneront date.

2.2. Utilisation des noms propres

Comme nous l'avons vu certaines règles ne nous permettent pas de trancher directement (par exemple What). Nous devons compléter notre analyse par un autre moyen. Le premier que nous avons mis en place concerne les questions qui comportent des noms propres. Tout d'abord la question est passée dans un analyseur syntaxique qui nous permet de trouver les noms propres dans la question (en général les mots commençant par une majuscule). En général ce nom propre est l'objet de la question ou alors l'objet est l'une de ses caractéristiques (Par ex : Quel est la date de naissance de Bruce Dickinson ? ou encore : Qui est le coéquipier de Batman ?). Nous devons donc trouver un moyen d'associer une catégorie à ce nom propre. Pour cela nous adressons une requête à un script issu de NLGbAse. Comme vu plus haut NLGbAse à associé une catégorie à chaque fiche de Wikipédia. Nous demandons donc à ce script de nous renvoyer la catégorie de la fiche correspondant à ce nom propre de la manière suivante : Si une fiche porte exactement le même nom alors la catégorie sera celle de cette fiche. Si ce n'est pas le cas mais que des fiches ont

[4] http ://www.wikipedia.org
[5] http ://www.nlgbase.org
[6] Référence nécessaire

un nom similaire alors la catégorie sera celle de la plus pertinente. Enfin si ce n'est pas le cas non plus nous appliquons la méthode suivante. Une recherche par tf/idf est faite. La solution la plus évidente aurait été de prendre pour catégorie celle du document ayant le meilleur score. Nous avons essayé une autre approche. Nous allons prendre la catégorie qui a le plus fort score. Pour cela chaque catégorie va se voir attribuer comme score la somme des scores des documents ayant cette catégorie. De ce fait la catégorie qui rassemble le plus de pertinence sera sélectionnée.
Cependant nous sommes conscients qu'il arrive parfois que cette stratégie ne soit pas idéale comme par exemple : Quel est le nom de la voiture de Batman ? Il y a des chances que cette méthode donne pers comme catégorie attendue alors que la solution idéale aurait probablement été prod. Cependant elle n'est pas totalement mauvaise car nous devrions trouver l'information dans la fiche de Batman donc même avec une étiquette pers (mais l'accès à la réponse est moins direct).

2.3. Utilisation des noms communs

Cependant toutes les questions ne comportent évidemment pas de noms propres mais uniquement des noms communs. C'est pourquoi nous avons du trouver un moyen de traiter ces questions.
Notre approche est assez simple mais efficace. Tout d'abord le parser que nous utilisons est capable (du moins en général) de trouver l'objet de la question. C'est celui ci que nous allons étudier. En effet par exemple pour la question "Que sont les généraux ?" l'objet de la question est "généraux". L ?enjeu est d'arriver à associer "généraux" à l'étiquette fonc.mil (fonction militaire).
Pour arriver à cela nous utilisons WordNet et ses hyperonymes ainsi que sa capacité à fournir déjà une classe pour chaque mot. En effet WordNet associe déjà à la totalité des termes une étiquette. Par exemple à "general" WordNet associe noun.person. L'étiquette que fournit WordNet est bien souvent satisfaisante, cependant son jeu d'étiquette ne correspond pas aux exigences d'Ester auquel notre projet doit se plier.
La première étape pour arranger ça fut d'associer "à la main" des étiquettes à des mots qui prendront le dessus sur celles de WordNet. En effet par exemple pour "general" nous ne voulons pas pers mais fonc.mil. Cependant faire ce travail sur tous les mots demanderait un investissement titanesque. Pour palier à ce problème pour chacune des catégories non traitées par WordNet et dont nous avons besoin, nous avons réfléchi aux mots les plus généraux possibles pour chaque catégorie. Ensuite nous vérifions que les hyponymes de ces mots sur WordNet correspondent bien à la même catégorie. Si ce n'est pas le cas alors nous allons sélectionner tout les hyponymes pour lesquels c'est le cas et répétons cette opération.
Ensuite l'algorithme est assez simple. C'est une fonction récursive qui va vérifier si le mot ne fait pas parti des mots étiquetés. Si ce n'est pas le cas cette vérification va être faite pour son hyperonyme. Les conditions d'arrêt sont soit on trouve un mot associé à une étiquette soit on arrive sur l'hyperonyme de plus haut niveau. Dans le premier cas le mot de départ se voit associer cette étiquette et donc la catégorie recherchée aussi. Dans le deuxième cas c'est l'étiquette associée au mot de départ qui prévaut et qui est donc la catégorie recherchée.
Au final même si il est évident que nous ne couvrons pas tout les cas en seulement trois heures il est possible de couvrir un très grand nombre de cas puisque pour chaque mot étiqueté tout ses hyponymes reçoivent cette même étiquette.

3. Extraction des mots clés