Méthodes d'indexation et de recherche d'information pour le Web Sémantique

Description de sujet

1. Contexte scientifique et motivations

La recherche d'information sémantique (RIS) a pour but de dépasser les limites d'une recherche classique par mots-clés. De plus en plus de travaux s'intéressent à l'exploitation de ressources sémantiques (ontologies, terminologies, thésaurii, etc.) pour améliorer l'accès à l'information.
Ces travaux sont à l'intersection de deux communautés : Recherche d'Information (RI) et Web Sémantique (WS). Ils s'intéressent entre autres à des questions d'annotation et indexation sémantique, désambiguïsation, extraction d'information (termes, entités nommées), expansion de requêtes, etc.
Néanmoins, ces deux communautés divergent dans les approches proposées. Ainsi la communauté WS s'intéresse plutôt à des questions de raisonnement et d'inférences en mettant l'accent sur les aspects formels. Très peu de travaux proposent d'intégrer des outils de traitement automatique de la langue et l'annotation sémantique reste assez fruste. La communauté RI reste cantonnée à des analyses distributionnelles et ne profite pas des raisonnements possibles sur les ontologies utilisées.
Le but de ce travail est de rapprocher ces deux communautés en proposant des méthodes hybrides d'indexation et de recherche d'information sémantique.
Ce stage fait suite au projet de fin d'étude portant sur une étude comparative de méthodes de recherche d'information sémantique [1]. Les comparaisons effectuées dans le cadre de ce projet serviront de *baseline* pour les travaux futurs. La mise en place d'une plateforme open source permet aisément d'intégrer de nouvelles fonctionnalités et faciliterait les développements futurs.

2. Objectifs du stage

Les objectifs de ce projet consistent à :

1. Établir un état de l'art des méthodes de recherche d'information sémantique de la communauté WS.
2. Analyser les liens et jonctions entre ces travaux et ceux déjà recensés de la communauté RI.
3. Proposition d'une ou plusieurs méthodes hybrides.
4. Mettre en place ces propositions.
5. Analyser l'apport des méthodes proposées.

A l'issue du stage, une plateforme d'indexation et de recherche d'information devrait être mise en place sous licence GPLv3. Cette plateforme devrait permettre d'ajouter aisément de nouvelles composantes et de pouvoir calculer différentes mesures d'évaluation (rappel, précision, etc.)

3. Corpus et ressources

Les corpus et ressources mis à disposition sont :

- ceux déjà utilisés dans le cadre du PFE : corpus de recettes de cuisine et ontologie associé [1].
- ressources proposés par le groupe uam en collaboration avec KMi : ces ressources comportent les corpus de TREC 9 et TREC 2001 avec des ontologies qui couvrent les domaines des requêtes [2].

4. Références

[1] Bannour I. «Mise en oeuvre d'une plateforme open source de comparaison de méthodes de recherche d'information sémantique», Rapport PFE ENSI, 68 pages, 2011.

[2] Uren, V., Sabou, M., Motta, E., Fernandez, M., Lopez, V., Lei, Y. (2011) «Reflections on five years of evaluating semantic search systems». International Journal of Metadata, Semantics and Ontologies, 5(2), p.87-98.