Stage Ingénieur 2010/2011

Titre : Etude comparative de méthodes de recherche d'information sémantique

Contexte scientifique et motivations du stage

L'utilisation d'ontologies dans le cadre d'une recherche d'information a pour but de dépasser les limites d'une recherche classique par mots clés.

Malgré la profusion de propositions de systèmes de recherche d'information sémantique, il est encore difficile de dresser un bilan. Ceci est principalement dû à un manque d'une base de test commune (*benchmark*) qui peut permettre de comparer les différentes méthodes et algorithmes entre eux.

Ce stage fait suite à un précédent travail de Master[1]. La partie expérimentation de ce dernier n'a pas été finalisée et souffre de plusieurs lacunes des protocoles de test.

L'objectif de ce stage est double : (1) participer à enrichir le *benchmark* constitué (cf. , (2) établir une comparaison "solide" des différents algorithmes de l'état de l'art.

Les scénarios de comparaison porteront sur :

- Choix des unités d'index
- Choix de pondération
- Intégration des connaissances ontologies

Objectifs

- Une étude comparative de deux moteurs de recherche *open source* qui serviront de baseline : Lucene et Terrier[2].
- Choisir les méthodes et algorithmes qui seront testés.
- Définir les différents scénarios de comparaison.
- Établir le protocole de comparaison et étoffer le benchmark déjà construit.

1. Imen Ouardani, Indexation et recherche d'information sémantique, Master2R Université Paris Dauphine, 51 pages

2. le(a) stagiaire est libre de proposer d'autres moteurs de recherche.

– Analyser les résultats et dresser un bilan des travaux examinés.

Corpus et ressources

Le corpus fourni par la compétition scientifique internationale Computer Cooking Contest 1 (CCC) sera exploité. La base comporte 1 489 recettes qui sont des documents textuels en XML faiblement structurés et comportent les éléments suivants : titre de la recette (<TI>), liste d'ingrédients (<IN>) et préparation (<PR>).

Nous nous intéressons essentiellement au contenu textuel, un précédent travail a "nettoyé" ce corpus du balisage XML.

La ressource sémantique qui sera utilisée est une ontologie formalisée en OWL qui décrit le domaine de cuisine. Elle a été développé par des chercheurs de l'équipe Orpailleur du LORIA. Cette ontologie a été construite à partir du corpus de recettes et du thésaurus de cuisine "The Cooking Thesaurus". L'ontologie n'est pas publique et le(a) stagiaire devra s'engager à ne pas la communiquer (ou partie) à des tiers.

Lieu du stage : LIPN (http ://www-lipn.univ-paris13.fr/), Université Paris 13

Encadrante : Haïfa Zargayouna (MCF, UP13)

Conditions Stage de 4 mois, financé sur projet.

Prérequis

– Programmation en Java indispensable ;
– Connaissance de OWL, RDF, XML, appréciées.
– Intérêt pour le Web Sémantique et la Recherche d'Information