Sujet de thèse

Accès sémantique à l'information textuelle en domaine ouvert

Encadrement : Adeline Nazarenko (directrice) et Haïfa Zargayouna (co-encadrante)

Laboratoire : LIPN

Mots clefs :
recherche d'information sémantique, annotation sémantique, passage à l'échelle, web sémantique, alignement d'ontologies

Objectif

L'accès sémantique à l'information a pour but de dépasser les limites d'une recherche classique par mots-clés qui prend mal en compte les problèmes de polysémie et de synonymie inhérente aux vocabulaires des langues naturelles. De plus en plus de travaux s'intéressent à l'exploitation d'une ontologie pour améliorer l'accès à l'information.

Ces travaux sont à l'intersection de courants de recherche : Recherche d'Information Sémantique (RIS) et Web Sémantique (WS). Les moteurs de RIS reposent sur des modèles classiques de recherche d'information qui ont été largement éprouvés. Ces modèles reposent sur la proximité et le poids distributionnel des unités choisies pour l'indexation. L'indexation sémantique repose sur la prise en compte d' « unités sémantiques » représentant des concepts plutôt que des mots et sur des calculs de voisinage dans l'ontologie. La RIS permet de présenter les résultats par ordre de pertinence. Les moteurs WS permettent un raisonnement plus fin sur la sémantique des documents explicitée dans un modèle sémantique (l'ontologie). Ils reposent sur une instantiation du modèle sémantique et permettent la mise en place de raisonnements sur ces représentations appelées « annotations sémantiques ». La recherche consiste à exploiter ces annotations et retourne des résultats généralement booléens (les documents répondent ou non à la requête).

Ces deux courants divergent donc dans les approches proposées. Ainsi le WS s'intéresse plutôt à des questions de raisonnement et d'inférences en mettant l'accent sur les aspects formels. La RI, qui reste cantonnée à des analyses distributionnelles, ne profite pas des raisonnements possibles sur les ontologies utilisées mais est capable de gérer de grandes masses de documents et de retourner des résultats triés par ordre de pertinence. Dans les deux cas cependant, les ontologies sont plutôt exploitées sous l'hypothèse du « monde fermé » et leur exploitation reste largement tributaire à leur couverture et leur qualité.

L'objectif de ce travail est de rapprocher ces deux courants en proposant une hybridation des méthodes d'indexation et de recherche d'information sémantique. Deux pistes sont possibles :

- Appliquer un moteur du WS (comme Corese [3]) en guise de filtrage des résultats d'un moteur RIS : une telle approche permettrait de tirer profit des raisonneurs pour affiner les résultats de la recherche.
- Proposer une indexation vraiment hybride qui intègre les annotations formelles à un modèle de RI existant. Des travaux tels que [2] ont montré des résultats encourageants mais ils reposent sur un modèle sémantique assez restreint et il s'agirait de mettre en place ce type d'approche à plus grande échelle.

De plus avec la multiplication des ontologies en ligne et les *Linked Data*, la vision monolithique où une seule ontologie sert à améliorer l'accès à l'information est remise en question. Le passage à l'échelle pose la problématique d'annotation sémantique d'une manière plus accrue avec des verrous tels que la sélection et l'alignement d'ontologies. Le but étant d'avoir la plus large couverture sémantique tout en garantissant des raisonnements cohérents et en raisonnant dans un « monde ouvert ».

L'évaluation de ce travail profitera des efforts de constitution de bancs d'évaluation pour la RI sémantique en spécifiant les fonctionnalités sémantiques à évaluer et en participant à la constitution de ressources appropriées pour l'évaluation [5]. Une collaboration est envisagée avec le groupe IRG (Information Retrieval Group)[1] et KMi (The Knowledge Media Institute) qui ont déjà entamé un premier travail dans ce sens et ont proposé de réutiliser des corpus de TREC 9 et TREC 2001 avec des ontologies (une vingtaine) qui couvrent les domaines des requêtes [4].

Contexte

Une plateforme d'indexation et de recherche d'information sémantique (Terrier SIR) [1] devrait être mise en place avant la rentrée 2012 sous licence GPLv3. Cette plateforme permet d'ajouter aisément de nouvelles composantes et de calculer différentes mesures d'évaluation (rappel, précision, etc.). Elle servira de plateforme test pour les questions de passage à l'échelle et d'intégration d'ontologies dans un processus de RI.

Ce sujet de thèse s'inscrit dans le cadre des travaux de l'équipe RCLN sur l'accès sémantique à l'information [6] et à l'arrière-plan de projets collaboratifs comme Legilocal qui comporte un volet d'annotation sémantique et de modélisation de documents.

Ces questions constituent également un enjeu important pour le Labex "Fondements empirique de la linguistique", auquel participe l'équipe RCLN, notamment pour les questions d'accès au contenu textuel de l'axe "Analyse sémantique computationnelle".

Références

[1] I. Bannour, H. Zargayouna (2012) «Une plate-forme open-source de recherche d'information sémantique» In COnférence en Recherche d'Information et Applications (CORIA), Mars 2012.

[2] P. Castells, M. Fernandez, and D. Vallet. (2007) «An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval». In IEEE Trans. on Knowl. and Data Eng. 19, 2 , 261-272.

[3] O Corby, R Dieng and C. Faron-Zucker «Querying the Semantic Web with Corese Search Engine» In Prestigious Applications of Intelligent Systems PAIS, ECAI, 2004.

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

[5] Haïfa Zargayouna (2011) «Quelle évaluation pour la Recherche d'Information Sémantique.» *Animation de table ronde* In Atelier Recherche d'Information SEmantique RISE@CORAI 2011.

[6] Haïfa Zargayouna (2005) «Indexation sémantique de documents XML» Thèse de doctorat. Université Paris-Sud, Déc. 2005.

1 Information Retrieval Group à Polytechnic School de Autónoma University of Madrid (http://ir.ii.uam.es/)