Titre :

Formalisation et extraction de points de vue de à partir de multiples ressources du Web.

Information :

Encadrant :	Clement Jonquet (LIRMM, UM2) – jonquet@lirmm.fr
	Philippe Lemoisson (Tetis, Cirad) – philippe.lemoisson@cirad.fr
Spécialités :	DECOL, AIGLE
Nombre d'étudiants :	3-4
Contexte:	Projet SIFR (Semantic Indexing of French Biomedical Data Resources)
Ou:	LIRMM
Quand:	2nd semestre 2012-2013

Mots clés :

Extraction de connaissances, points de vue, API web, web services, technologies web, web sémantique, ingénierie des connaissances, base de données

Technologies :

Java, technologies web de votre choix. Vous devrez écrire des clients de service web (SOAP ou REST), MySQL, ResTful web services, XML/JSON, RDF.

Résumé :

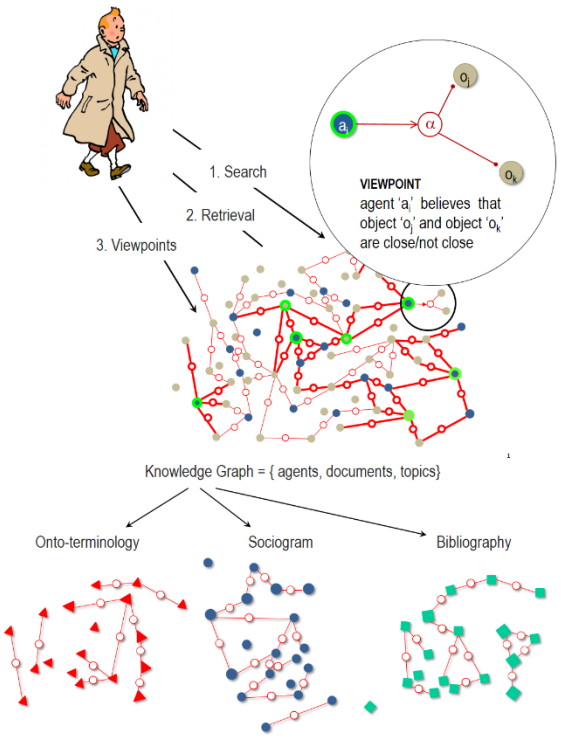
Depuis l'émergence du web 2.0, les utilisateurs du web sont énormément sollicités et participent massivement à la production de contenus. Ces sollicitations se transforment en émissions de point de vues divers et variés explicites ou non. Par exemple, un « I like » sur Facebook, ou un tag de photo sur Flickr ou tout simplement la rédaction d'un Tweet sur un sujet donné. De manière similaire, les données (scientifiques, multimédias, publiques, etc.) sont elles aussi annotées manuellement ou automatiquement pour les enrichir ou les indexer, ce qui représente autant de points de vue émis par les annotateurs. Le LIRMM et le CIRAD travaillent sur un formalisme, nommé Viewpoints, de représentation et d'exploitation de points de vue pour la création et l'évolution de connaissances au sein de communautés. Ce TER consiste à concevoir et implémenter des outils extracteurs de points de vue à partir de différentes ressources de données du web. Nous nous intéresserons entre autres à des données de publication scientifiques (Pubmed), des données sociales (Research Gate) et des données de projets (Cordis). Les technologies à utiliser pour développer les extracteurs sont libres de choix. Le format d'extraction sera spécifié (XML), de façon à pouvoir alimenter directement un graphe des points de vue construit à l'aide d'une API Java existante.

Contexte : l'approche « ViewpointS »

Nous proposons de considérer le web comme un réseau peuplé de trois types d'objets : i) les documents, ii) les personnes ou "agents" et iii) les concepts ou « topics » (ontologies/web sémantique), ces objets étant reliés entre eux par les points de vue des agents. Par exemple « je (moi ou Google) pense que ce document traite de ce topic ».

L'approche ViewpointS [2] met en œuvre un formalisme simple et puissant qui unifie les notions d'indexation manuelle par des métadonnées, d'indexation automatique, d'opinons relatives à la pertinence des réponses à des recherches d'information, d'opinons relatives aux proximités entre agents, documents et topics (topic est pris ici au sens de mot-clé ou plus généralement d'objet d'investigation).

Les *viewpoints* sont émis par les agents (humains ou artificiels); ils expriment la croyance de l'agent relativement à la proximité sémantique entre deux « objets » (agents, documents ou topics). Chaque *viewpoint* connecte donc trois objets : l'émetteur (nécessairement un agent) et deux objets quelconques. Cette relation crée un graphe bi-parti (objets/viewpoints) et fonde le calcul d'une distance sur l'ensemble des objets. C'est la métaphore de la synapse : deux objets seront d'autant plus proches qu'un grand nombre de viewpoints positifs les relient. Chaque recherche d'information est ramenée à un calcul de voisinage (suivant une distance) au sein du sous-graphe contenant les agents, les documents et les topics.



Les *viewpoints* peuvent être émis de façon constructive, ou de façon réactive à l'issue d'une recherche d'information avec le système: par exemple, le requêteur est incité à exprimer son point de vue (positif ou négatif) sur la proximité entre l'objet initial de la requête et chacun des objets obtenus en réponse. Les distances dans le graphe évoluent donc continuellement et reflètent de façon dynamique la connaissance implicite de la communauté. Ce graphe socialement construit et évolutif en fonction des créations/suppressions de viewpoints est le cœur des mécanismes d'indexation et de recherche d'information. Il permet de produire automatiquement des sociogrammes, des onto-terminologies, des bibliographies organisées en tant que sous graphe (ou vues) du graphe principalement. Il permet également de répondre aux questions : quel est le plus court chemin entre deux objets ? quel est le voisinage d'un objet ? etc.

Description du travail :

Dans ce TER nous nous intéresserons non pas à des *viewpoints* réactifs mais à la détection des *viewpoints* « cachés » dans des ressources de données (articles scientifiques, posts, projets, etc.) préalablement produites par des agents (scientifiques, organisations, outils logiciels) qui les annotent ou indexent par des topics (mot clés, conférences, concepts d'ontologies).

Nous nous intéresserons par exemple aux ressources de données suivantes :

- PubMed (http://www.ncbi.nlm.nih.gov/pubmed) – base de données de référence pour les publications scientifiques dans le domaine biomédical.

- Research Gate (www.researchgate.net) – réseau social (similaire a Facebook) de chercheur scientifiques.

- Cordis (http://cordis.europa.eu) – base de données des projets de recherches européens.

Pour chacune de ces ressources, il s'agira tout d'abord de proposer un modèle capable de traduire sous forme de *viewpoints* l'information disponible, par exemple :

le fait qu'une personne soit auteur d'un article,
le fait qu'un article soit indexé avec tel ou tel mot clés,
le fait qu'un article référence d'autres articles,
le fait qu'un article soit publié dans un journal scientifique,
le fait qu'un chercheur soit un « follower » d'un autre,
le fait qu'une organisation participe à un projet de recherche,
le fait qu'un article soit issu d'un projet de recherche,
le fait qu'un post soit « liké » par un chercheur, etc.

Il s'agira ensuite, pour chaque type de *viewpoint* intégré au modèle, de l'identifier en s'appuyant sur les ontologies et vocabulaires du web sémantique [1], par exemple : dc:author, foaf:knows, rdf:about, etc.

Il s'agira enfin, pour chacune de ces ressources, d'écrire un accesseur (wrapper) afin d'« aspirer » le contenu de la ressource (par exemple via un web service d'accès à la ressource) en le traduisant/indexant sous forme de viewpoints. Une API java est disponible pour manipuler les graphes de viewpoints et sera utilisée pour importer/exporter les viewpoints à partir des accesseurs.

Références :

[1] Fabien Gandon, Catherine Faron-Zucker, and Olivier Corby. *Le web sémantique - Comment lier les données et les schémas sur le web ?* Dunod, 2012.

[2] Philippe Lemoisson, Guillaume Surroca, and Stefano A. Cerri. Viewpoints: An Alternative Approach toward Business Intelligence. In *eChallenges e-2013 Conference*, page 8, Dublin, Irland, October 2013.

http://wwwis.win.tue.nl/infwet03/proceedings/8.pdf