Sujet de stage 2013 : développement d'un simulateur dédié à l'intégration de données sémantique

Stagiaire : Oliver Conus
oliver.conus@etu.univ-lyon1.fr

Encadrants : Fabien Duchateau, Nicolas Lumineau et Mohand-Said Hacid
Université Lyon 1, LIRIS, UMR5205, Lyon, France
{fduchate,nluminea,mhacid}@liris.cnrs.fr

01/09/13 au 31/12/13

1 Contexte et objectifs

Ce sujet de stage s'inscrit dans le cadre du projet franco-norvégien Aurora KOGAR (*Knowledge Gardening in the Web of Data*) et du soutien Bonus Qualité Recherche de l'Université Claude Bernard Lyon 1.
Plus spécifiquement, le projet s'intéresse au partage et à la diffusion d'informations par des institutions culturelles (musées, bibliothèques). Ces dernières peuvent être vues comme des pairs autonomes sur un réseau mettant à disposition certaines collections de données. Un processus automatique d'intégration permet de découvrir des liens sémantiques (correspondances, ou mappings) entre ces collections de données. L'outil développé servira à des expérimentations scientifiques, pour tester différentes stratégies de découverte de correspondance, les performances du réseau P2P, la complétude, la stabilité des résultats d'une requête, la robustesse du réseau, etc.

Objectifs du stage :

- Créer des jeux de données
- Améliorer l'outil de découverte de correspondances
- Programmer un simulateur P2P
- Réaliser des expérimentations

Des présentations d'avancée de l'outil sont à envisager avec l'équipe norvégienne (par visio-conférence).

2 Création de jeux de données

Dans le domaine culturel, la mise à disposition d'œuvres ne manque pas. Dans le Web des Données, on parle **d'entités**. Ainsi, l'œuvre *le seigneur des anneaux* est une entité, tout comme son auteur *JRR Tolkien*. Une entité possède des attributs (ou propriétés). Par exemple, l'entité *le seigneur des anneaux* possède un attribut précisant son type (*une œuvre*), un attribut précisant sa date de rédaction (*1954*), etc.

La difficulté ne réside pas dans la création ou récupération d'entités, mais dans un choix judicieux (les sources doivent avoir des entités équivalentes) et dans la création (souvent manuelle ou semi-automatique) des correspondances. On peut évidemment envisager plusieurs domaines (e.g., un domaine *livres*, un domaine *films*).

Quelques pistes pour trouver des jeux de données :

- Recherche (moteurs de recherche, articles scientifiques)
- Yahoo met à disposition un moteur de recherche RDF[1], basé sur CommonsCrawl et le Web Data Commons dataset[2]. Les schémas pour chaque type d'entité (e.g., Recipe[3], TVepisode) proviennent de `http://schema.org`.
- Travaux de Naimdjon/Fabien avec la collection FRBR de la bibliothèque nationale norvégienne
- Créer un outil qui extrait des entités de DBpedia à partir d'une page Wikipedia, qui liste par exemple tous les films de science fiction

Pour démarrer et tester le prototype, on pourra utiliser des fichiers CSV sur les publications d'articles scientifiques (cf Fabien). On avait évoqué également le e-commerce (avec un scénario), mais ce n'est pas le domaine de l'héritage culturel initialement prévu dans le projet KOGAR.

3 Découverte de correspondances

L'outil d'intégration permet la découverte de correspondances entre les sources de données. Si possible, l'outil d'intégration et le simulateur P2P seront "séparés" au niveau programmation, afin de pouvoir utiliser l'un ou l'autre individuellement.

La découverte de correspondances s'effectue à deux niveaux :

- Modèle (e.g., modèle conceptuel, schéma, ontologie)
- Données (e.g., instances)

[1] Yahoo RDF search engine, `http://glimmer.research.yahoo.com/`
[2] Web Data Commons, `http://webdatacommons.org/`
[3] Schema for Recipe, `http://schema.org/Recipe`

Une réunion sera nécessaire pour expliquer le fonctionnement des approches de découverte de correspondances en détail.

Hypothèses :

- Les institutions culturelles (musées, bibliothèques, etc.) peuvent mettre à disposition leurs collections de données (e.g., sur des livres, des films, des oeuvres).
- Les sources de données sont très hétérogènes
- Certaines sources de données sont fortement disponibles (e.g., DBpedia, Freebase, OpenCyc, Amazon)
- Le réseau doit être capable de stocker certains connaissances (les correspondances et des informations sur les mesures de similarité)

Intuition : l'outil de découverte de correspondances va découvrir des correspondances entre les sources de données des institutions culturelles (couche basse) et les sources de données fortement disponibles (couche haute). Face à l'hétérogénéité importante des sources, il est difficile de configurer automatiquement un outil. Une approche générique[4] [5] a été développée, mais devra être légèrement adaptée (cf Fabien). De plus, il faudra spécifier le formalisme d'une correspondance, et les règles applicables comme la transitivité (cf encadrants).

4 Simulation P2P

Pour rappel, l'équipe norvégienne doit implémenter un vrai réseau P2P (éventuellement basé sur JXTA[6]). L'équipe française utilise donc un simulateur pour les premiers tests, par exemple PeerSim ou JADE (cf Nicolas). Les simulations devront se faire sur les jeux de données et les outils de recherche de correspondances préalablement définis. Afin d'effectuer des simulations réalistes, il sera nécessaire de définir les protocoles de communication entre les pairs simulés de manière à tenir compte de différents types de topologie. Afin d'améliorer les interactions entre nœuds voisins, il est pertinent de définir une topologie en adéquation avec les correspondances liant sémantiquement les pairs.
En plus des données, chaque pair simulé doit pouvoir accéder à des outils de découverte de correspondance et stocker les informations concernant les correspondances sémantiques existantes et découvertes. La découverte de correspondances se base sur des mesures de similarité qu'il sera nécessaire d'évaluer pour ensuite déterminer la ou la combinaison de mesure qu'il est pertinent d'utiliser. Enfin, se pose le problème de l'accès et la maintenance des correspondances. En effet, un pair entrant dans le réseau doit pouvoir accéder efficacement aux pairs et donc aux correspondances permettant d'interagir avec le réseau. De plus, l'obsolescence des correspondances doit être considérée pour éviter le stockage et la maintenance de correspondances non pertinente suite à l'évolution des données stockées sur les pais.

[4] Article : `http://liris.cnrs.fr/~fduchate/papers/duchateau-data13.pdf`
[5] Slides : `http://liris.cnrs.fr/~fduchate/presentations/duchateau-presentation-data13.pdf`
[6] JXTA, `https://jxta.kenai.com/`

Intuitivement, il sera nécessaire d'indexer de manière répartie les pairs/correspondances/groupe de correspondances/thème pour permettre à un pair entrant d'identifier efficacement un voisinage pertinent dans le réseau.

5 Calendrier

Le travail pendant ce stage de 4 mois peut-être réparti ainsi (à rediscuter) :

- **01/09 au 15/09** : Compréhension du stage, recherche des jeux de données, comparatif PeerSim/JADE
- **15/09 au 01/10** : Prise en main de l'outil de découverte de correspondances
- **01/10 au 15/10** : Prise en main du simulateur, définition des protocoles de communication
- **15/10 au 15/11** : Programmation des simulations
- **15/10 au 15/12** : Rédaction d'un rapport avec tests comparatifs et perspectives

Travail à plus long terme :

- Enrichissement sémantique : quand on veut de l'information sur une entité donnée, il faut fusionner les occurrences/mentions de cette entité stockées sur les différentes sources de données.
- Gestion des connaissances : découverte des incohérences, degré de confiance associé à un pair, etc.