Réconciliation des données représentées dans des tableaux : détection de données redondantes ou similaires ?

Contacts : Fatiha.sais@lri.fr, juliette.dibie_barthelemy@agroparistech.fr et patrice.buche@supagro.inra.fr

Actuellement, de plus en plus de sources d'informations sont disponibles et accessibles via le Web. Des travaux ont été déjà menés sur la construction semi-automatique d'un entrepôt thématique de données extraites à partir de documents hétérogènes collectés à partir du Web. Ces travaux se sont focalisés sur l'extraction et l'intégration des parties les plus structurées des documents, c'est-à-dire, celles qui sont représentées par des tableaux (Buche et al. 2009, Hignette et al. 2009).
Les tableaux extraits de sources variées (articles scientifiques, rapports de projets, mémoire de thèse, …) et de formats variés (documents PDF et HTML notamment) sont annotés sémantiquement avec un vocabulaire unique défini dans une ontologie de domaine. Les annotations sont exprimées sous la forme de graphes RDF. L'annotation de tableau est effectuée à deux niveaux de granularité : (i) annotation de sous ensembles de colonnes par des relations sémantiques déclarées dans l'ontologie de domaine et (ii) annotation de chaque cellule du tableau par un sous-ensemble flou à support numérique ou symbolique selon la valeur contenue dans la cellule. Une fois que les tableaux sont annotés, un opérateur (un expert du domaine) les ajoute un à un à l'entrepôt de données.
Il arrive souvent que les auteurs des documents scientifiques présentent les mêmes résultats expérimentaux dans différentes publications. Sans détection a priori de redondance, ces tableaux sont ajoutés à l'entrepôt de données et considérés comme des tableaux deux à deux différents. Cependant, conserver cette redondance d'informations peut contribuer à la dégradation de la qualité des données de l'entrepôt et par conséquent des résultats des services qui les exploitent, e.g. veille sanitaire, veille scientifique, des applications statistiques, etc.

Nous avons proposé une méthode de détection de tableaux redondants dans l'entrepôt de données (Khefifi et al. 2011) en nous inspirant des travaux existants sur la réconciliation de données (Saïs et al. 2009). L'approche proposée permet de détecter la redondance au niveau des instances de relations sémantique associées aux lignes des tableaux. La limite de cette approche est, d'une part, de pouvoir distinguer les cas où les tableaux sont réellement redondants des cas où ils ne sont que similaires et, d'autre part, de pouvoir identifier les redondances lorsque les données ne sont pas reproduites avec le même niveau de granularité (donnée détaillée/données moyennées). Pour ce faire, il faudra proposer une approche plus globale qui pourra exploiter d'autres informations sur les tableaux comme leur titre, les auteurs des publications dans lesquelles ils ont été trouvés. La question des actions à entreprendre une fois la redondance détectée devra également être traitée (dans quel tableau effectuer la suppression, comment gérer la traçabilité de la suppression, …).

Le but du stage est d'étudier les aspects théoriques et pratiques d'une telle distinction entre redondance et similarité. Le stagiaire aura à identifier les nouveaux critères permettant cette distinction, adapter la méthode de détection de redondances pour prendre en compte ces nouveaux critères et implémenter un prototype. Des tests pourront être réalisés en s'aidant de l'outil @Web co-développé par les équipes INRA de Mét@Risk (AgroParisTech) et de l'UMR IATE.

Compétences requises :
- Web Sémantique, Ontologies, RDF, logique et raisonnement.
- Bases de données, Logique floue, Java/XML.

Co-encadrement : Patrice Buche, Juliette Dibie Barthélemy et Fatiha Saïs

Lieu du stage : INRIA-Saclay Île-de-France (Orsay)
Rémunération : oui.
Durée du stage : 6 mois