Détection de redondances dans les tableaux guidée par une ontologie

Rania Khefifi*, Patrice Buche**,
Juliette Dibie-Barthélemy***, Fatiha Saïs*

* LRI/INRIA Saclay, 4 rue Jacques Monod, F-91893 Orsay Cedex, France
{Rania.Khefifi, Fatiha.Sais}@lri.fr
**INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2, France
LIRMM, CNRS-UM2, F-34392 Montpellier, France
Patrice.Buche@supagro.inra.fr
***INRA - Mét@risk & AgroParisTech, 16 rue Claude Bernard,
F-75231 Paris Cedex 5, France
Juliette.Dibie@agroparistech.fr

Résumé. Nous nous intéressons dans cet article à la réconciliation d'annotations floues associées à des tableaux de données par une méthode d'annotation sémantique, qui est guidée par une ontologie de domaine. Etant donnés deux tableaux, la méthode consiste à détecter leurs instances de relation redondantes. Elle s'appuie sur les connaissances déclarées dans l'ontologie, ainsi que sur des scores de similarité entre les annotations floues représentées par des sous-ensembles flous numériques ou par des sous-ensembles flous symboliques.

1 Introduction

L'ouverture sur le Web permet de publier des documents sans aucun contrôle sur leur contenu. Cet absence de contrôle a des avantages : la richesse et la diversité des informations disponibles sur le Web ; mais elle a également des inconvénients, notamment l'hétérogénéité et la redondance de ces informations. Dans le domaine de l'intégration de données, des travaux ont été menés sur la construction d'entrepôts thématiques de données extraites à partir de sources hétérogènes publiées sur le Web. Certains travaux, comme celui de Hignette et al. (2009), se sont en particulier focalisés sur l'extraction et l'intégration de données structurées représentées dans des tableaux. Les tableaux extraits proviennent de différentes sources (articles scientifiques, rapports de projets, mémoires de thèses, etc.) dans des formats hétérogènes (documents HTML, documents PDF ...). Leur intégration dans un entrepôt thématique repose sur une méthode d'annotation sémantique de tableaux, guidée par une ontologie de domaine, qui permet de traiter le problème de l'hétérogénéité sémantique du point de vue du vocabulaire utilisé pour décrire les données. Cette méthode génère automatiquement des annotations floues. Une annotation floue représente une instance de relation sémantique de l'ontologie reconnue sur une ligne d'un tableau. L'annotation sémantique n'empêche cependant pas l'intégration de données redondantes dans l'entrepôt. La présence de redondance dans l'entrepôt dégrade la