Intégration Holistique des Graphes basée sur la Programmation Linéaire pour l'Entreposage des Open Data

Alain Berro*, Imen Megdiche*, Olivier Teste*

* IRIT UMR 5505, Université de Toulouse
CNRS, INPT, UPS, UT1, UT2J, 31062 Toulouse Cedex 9
{berro, megdiche, teste}@irit.fr

Résumé. Dans cet article, nous proposons une approche holistique pour l'intégration des graphes d'Open Data. Ces graphes représentent une classification hiérarchique des concepts extraits des Open Data. Nous nous focalisons sur la conservation de hiérarchies strictes lors de l'intégration afin de pouvoir définir un schéma multidimensionnel à partir de ces hiérarchies et entreposer par la suite ces sources de données. Notre approche est basée sur un programme linéaire qui résout automatiquement la tâche de matching des graphes tout en maximisant globalement la somme des similarités entre les concepts. Ce programme est composé de contraintes sur la cardinalité du matching et de contraintes sur la structure des graphes. A notre connaissance, notre approche est la première à fournir une solution optimale globale pour le matching holistique des graphes avec un temps de résolution raisonnable. Nous comparons également la qualité des résultats de notre approche par rapport à d'autres approches de la littérature.

1 Introduction

L'intégration des données est un problème de recherche largement étudié depuis plusieurs années (Rahm et Bernstein, 2001). La difficulté dans ce domaine réside dans la complexité et la diversité des sources de données à traiter et dans l'estimation de la qualité et de la performance des approches automatiques. L'intégration repose sur une tâche appelée matching qui a pour rôle de déterminer les meilleures correspondances entre les éléments des sources de données. Dans la littérature, nous distinguons deux axes de matching selon le nombre de sources à traiter à savoir le pair-wise matching (deux sources de données) et le matching holistique (plusieurs sources de données). Pour le pair-wise matching, le défi consiste à trouver les meilleures correspondances pour deux sources de données (de petite ou moyenne taille). A ce défi s'ajoute le problème de performance lorsque les sources sont de grandes tailles. Quant au matching holistique, les défis consistent à maintenir une bonne qualité des correspondances et à garantir une performance acceptable vis-à-vis de la taille et du nombre de sources traitées.

Les données ouvertes ou Open Data (OD) tabulaires statistiques constituent des sources d'informations intéressantes à intégrer dans les entrepôts de données vu la diversité des scénarios d'analyses qui peuvent en découler. Toutefois, ces données ont trois principales caractéristiques qui complexifient leur intégration : l'hétérogénéité syntaxique et sémantique, la