Fusion de données redondantes : une approche explicative

Fatiha Saïs[(1)] and Rallou Thomopoulos[(2)]

[(1)] LRI (CNRS UMR8623 & Université Paris Sud), Université Paris-Saclay,
Bât. 650 Ada Lovelace, F-91405 Orsay Cedex
[(2)] INRA (UMR IATE) & INRIA GraphIK, 2 place Viala, F-34060 Montpellier cedex 1

Résumé. Nous nous intéressons, dans le cadre du projet ANR Qualinca au traitement des données redondantes. Nous supposons dans cet article que cette redondance a déjà été établie par une étape préalable de liage de données. La question abordée est la suivante : comment proposer une représentation unique en fusionnant les "duplicats" identifiés ? Plus spécifiquement, comment décider, pour chaque propriété de la donnée considérée, quelle valeur choisir parmi celles figurant dans les "duplicats" à fusionner ? Quelle méthode adopter dans le but de pouvoir, par la suite, retracer et expliquer le résultat obtenu de façon transparente et compréhensible par l'utilisateur ? Nous nous appuyons pour cela sur une approche de décision multicritère et d'argumentation.

1 Introduction

Le *liage de données*, aussi appelé dans la littérature *réconciliation de données* (Ferrara et al. (2013); Saïs et al. (2009)), est le problème où l'on s'intéresse à détecter les descriptions référant au même objet du monde réel (e.g. même personne, même livre). La *fusion de données* est le problème posé suite à un liage de données si l'on souhaite fournir à l'utilisateur une représentation unique et homogène des données liées. La difficulté majeure rencontrée est celle des conflits et des inconsistances entre les valeurs d'une même propriété des données liées. Pour obtenir des données cohérentes et de bonne qualité, il faut résoudre ces conflits et choisir pour chaque propriété une ou plusieurs valeurs (cas des propriétés multi-valuées).

Des travaux ont étudié le problème de fusion de données dans le domaine des bases de données relationnelles (voir Bleiholder et Naumann (2009) pour un état de l'art). Dans cet article, nous nous intéressons au contexte du Web de données et étudions le problème de fusion de données RDF, distinct du cadre relationnel car caractérisé par la souplesse intrinsèque au modèle permettant la multi-valuation des propriétés, l'hypothèse du monde ouvert et la possibilité d'avoir plusieurs ontologies (schémas) décrivant les données. Récemment, des travaux (Saïs et Thomopoulos (2008); Flouris et al. (2012); Mendes et al. (2012)) se sont intéressés au problème de fusion de données RDF. Cependant aucun ne permet une bonne compréhension par l'utilisateur des résultats de la fusion. C'est sur cet aspect explicatif, visant à tracer et restituer les raisons ayant conduit à un résultat de fusion, que nous nous focalisons dans cet article.

La partie 2 décrit le problème de fusion et l'exemple illustratif. La partie 3 présente la méthode multicritère de fusion de données. La partie 4 introduit le processus explicatif qui s'appuie sur la construction d'arguments. La partie 5 conclut cette étude.