Définir les catégories de *DBpédia* avec des règles d'associations et des redescriptions

Justine Reynaud*, Esther Galbrun* Mehwish Alam** Yannick Toussaint*, Amedeo Napoli*

*LORIA (CNRS - INRIA - Université de Lorraine)
Campus Scientifique BP 239 – 54506 Vandœuvre-lès-Nancy
prenom.nom@loria.fr,
** Semantic Technology Lab, ISTC-CNR,
Rome, Italy.
mehwish.alam@istc.cnr.it

Résumé. *DBpédia*, qui encode les connaissances de *Wikipédia*, est devenue une base de référence pour le web des données. Les ressources peuvent y être répertoriées par des catégories définies manuellement, dont la sémantique n'est pas directement accessible par des machines. Dans cet article, nous proposons de remédier à cette lacune au moyen de méthodes de fouille de données, à savoir la recherche de règles d'associations et de motifs apparentés. Nous présentons une étude comparative de ces variantes sur une partie de *DBpédia* et discutons le potentiel des différentes approches.

1 Introduction

Le foisonnement des bases de connaissances sur le web pose de nouveaux enjeux quant à leur construction, leur enrichissement et leur interrogation. Nous prenons ici l'exemple de *DBpédia*. Dans cette base de connaissances, les ressources peuvent appartenir à une ou plusieurs *catégories*, générées manuellement. Cependant, les catégories sont définies en extension : on sait *comment* les ressources sont groupées, mais pas *pourquoi* elles sont groupées ainsi. Dans cet article, nous souhaitons définir les catégories en intension. C'est-à-dire que nous nous intéressons à des méthodes permettant d'expliciter les critères de regroupement.

Savoir caractériser ces catégories permettra non seulement d'enrichir *DBpédia* par cette nouvelle connaissance, mais aussi de corriger la base : les ressources assignées à tort à une catégorie ou inversement, les ressources non assignées à une catégorie à laquelle elles sont sensées appartenir pourront être détectées automatiquement.

La base de connaissances de *DBpédia* contient un ensemble de triplets RDF, dont les sujets correspondent à des articles de *Wikipédia*, associés à des paires (`predicat`, `objet`) qui représentent les catégories auxquelles ces articles sont rattachés et d'autres informations.

Afin de caractériser ces catégories en terme des autres informations, nous recherchons des définitions en exploitant des techniques de fouille de données. Chaque paire distincte (`predicat`, `objet`) est représentée par un attribut Booléen, tandis que chaque article est représenté par une entité, associée à un sous ensemble d'attributs représentant les catégories