

Concepts de voisins dans les graphes RDF : Une extension Jena et une interface graphique

Nicolas Fouqué*, Sébastien Ferré*, Peggy Cellier*

*Univ Rennes, CNRS, INSA, IRISA
Campus de Beaulieu, 35042 Rennes cedex, France
Prenom.Nom@irisa.fr

Résumé. Les concepts de voisins définissent une forme symbolique de similarité entre les entités d'un graphe de connaissances. Partant d'une entité, chaque concept de voisins est un cluster d'entités voisines partageant un même motif de graphe centré sur l'entité. Dans ce papier démo, nous rappelons les définitions des concepts de voisins et nous présentons une extension de la librairie Jena dont l'API permet de calculer les concepts de voisins pour un modèle RDF(S) Jena. Nous présentons également une interface graphique permettant à un utilisateur d'effectuer ces calculs de façon simple et interactive.

1 Introduction

La méthode des k plus proches voisins (k-NN) (Mitchell, 1997) s'appuie sur la notion de *similarité* entre instances pour réaliser différentes tâches d'inférences, telles que la classification ou le raisonnement à partir de cas (De Mantaras et al., 2005). Par exemple, la classe d'une nouvelle instance x peut être décidée par un vote majoritaire parmi les k plus proches voisins de x , c'est-à-dire les k instances les plus similaires à x . Notre préoccupation est de pouvoir définir et calculer la similarité entre les nœuds d'un graphe de connaissances, typiquement un graphe RDF (Hitzler et al., 2009) mais pas seulement. Par similarité entre nœuds on entend ici la similarité des descriptions de ces nœuds, où la description d'un nœud X est constituée des relations arrivant et partant de X et récursivement des descriptions des nœuds adjacents. Par exemple, la similitude entre les pays France et Italie est que ce sont des républiques européennes parlant une langue latine et ayant les Alpes comme massif (entre autres choses).

Bisson (2000) distingue deux types de similarités : numériques et symboliques. Les similarités *numériques* ont l'avantage d'être faciles d'emploi mais elles n'offrent pas d'explications (*Pourquoi cette instance est plus similaire que telle autre ?*) et peuvent masquer des différences (*Deux instances ont la même similarité mais pour des raisons très différentes.*). De plus, peu de ces mesures ont été définies pour des données relationnelles (ex., RIBL (Horváth et al., 2001)). Les similarités *symboliques* évitent ces inconvénients en produisant des représentations symboliques généralisant plusieurs instances. Néanmoins, à notre connaissance, ces similarités ont seulement été utilisées pour la formation de concepts et de règles par généralisation (ex., Programmation Logique Inductive, PLI (Muggleton, 1995)), pas pour la recherche de plus proches voisins. De plus, les travaux existants considèrent généralement comme instances des