

Prédiction de liens dans les graphes de connaissances avec les concepts de plus proches voisins

Sébastien Ferré*

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

Résumé. La nature ouverte des graphes de connaissances implique souvent qu'ils soient incomplets. La prédiction de liens consiste à inférer de nouveaux liens entre entités sur la base des liens existants. La plupart des approches existantes s'appuient sur l'apprentissage de vecteurs de traits latents pour l'encodage des entités et des relations. En général cependant, les traits latents ne sont pas facilement interprétables. Les approches à base de règles sont interprétables mais un ensemble de règles différent doit être appris pour chaque relation. Nous proposons une nouvelle approche qui n'a pas besoin de phase d'apprentissage et qui peut fournir des explications intelligibles pour chaque inférence. Elle repose sur le calcul de Concepts de plus proches voisins (*Concepts of Nearest Neighbours*, CNN) pour identifier des entités similaires fondées sur des motifs de graphe communs. La théorie de Dempster-Shafer est ensuite utilisée pour tirer des inférences à partir des CNN. Nous évaluons notre approche sur FB15k-237, un *benchmark* classique en prédiction de liens, où elle obtient de meilleures performances que les approches existantes.

1 Introduction

Les graphes de connaissances (GC) sont de plus en plus utilisés pour représenter et partager les données sur le Web. Le Web sémantique définit des standards pour la représentation (RDF), l'interrogation (SPARQL) et le raisonnement (RDFS/OWL) sur ces GC. Des milliers de GC sont disponibles : ex., DBpedia, Wikidata (préc. Freebase), YAGO, WordNet. La nature ouverte des GC implique souvent qu'ils soient incomplets, et différentes techniques d'apprentissage automatique ont été étudiées pour les compléter.

La tâche de *prédiction de liens* (Nickel et al., 2016) consiste à prédire les (parties d')arcs manquants. Supposons que le directeur du film *Avatar* soit absent du GC, on souhaite le prédire, c-à-d. l'identifier parmi tous les nœuds du GC. Le principe est de trouver des régularités dans les connaissances existantes et de les exploiter afin de classer les nœuds du GC. Plus haut est un nœud dans le classement, meilleure est la prédiction. La prédiction de liens a été introduite pour les réseaux sociaux avec un seul type d'arc (Liben-Nowell et Kleinberg, 2007) puis a ensuite été étendue à des données multi-relationnelles et appliquée aux GC (Nickel et al., 2016). Comparé à la classification supervisée, la prédiction de liens affronte plusieurs défis. Premièrement, il y a autant de problèmes de classification qu'il y a de relations, qui comptent souvent dans les centaines. Deuxièmement,