RÉSUMÉ

Les graphes de connaissances jouent aujourd'hui un rôle important pour représenter et stocker des données, bien au-delà du Web sémantique ; beaucoup d'entre eux sont obtenus de manière automatique ou collaborative, et agrègent des données issues de sources diverses. Dans ces conditions, la création et la mise à jour automatique d'une taxonomie qui reflète le contenu d'un graphe est un enjeu crucial.

Or, la plupart des méthodes d'extraction taxonomique adaptées aux graphes de grande taille se contentent de hiérarchiser des classes pré-existantes, et sont incapables d'identifier de nouvelles classes à partir des données. Dans ce mémoire, nous proposons une méthode d'extraction de taxonomie expressive applicable à grande échelle, grâce à l'utilisation de plongements vectoriels. Les modèles de plongement vectoriel de graphe fournissent une représentation vectorielle dense des éléments d'un graphe, qui intègre sous forme géométrique les régularités des données : ainsi, deux éléments sémantiquement proches dans le graphe auront des plongements vectoriels géométriquement proches.

Notre but est de démontrer le potentiel du regroupement hiérarchique non-supervisé appliqué aux plongements vectoriels sur la tâche d'extraction de taxonomie. Pour cela, nous procédons en deux étapes : nous montrons d'abord qu'un tel regroupement est capable d'extraire une taxonomie sur les classes existantes, puis qu'il permet de surcroît d'identifier de nouvelles classes et de les organiser hiérarchiquement, c'est-à-dire d'extraire une taxonomie expressive.

Pour l'extraction de taxonomie sur les classes existantes, nous proposons deux méthodes capables d'associer des classes existantes à des groupes d'entités en tenant compte de la structure d'arbre qui existe entre ces groupes ; cela permet de transformer l'arbre de clustering issu du regroupement hiérarchique en une taxonomie. La première de ces méthodes consiste à trouver une injection optimale des classes vers les clusters en résolvant un problème d'optimisation linéaire. La seconde est un lissage de la méthode précédente, conçue pour mieux tenir compte du bruit dans les données. Nous appliquons ces deux méthodes à DBpedia, et montrons qu'elles sont toutes deux capables de surpasser une méthode basée sur un regroupement supervisé.

Pour l'extraction de taxonomie expressive, nous présentons une méthode d'extraction d'axiomes capable de tirer profit d'un arbre de clustering pour obtenir des exemples positifs et négatifs pertinents, et induire des axiomes à partir de ces exemples. Nous y ajoutons un mécanisme de tirage aléatoire capable d'augmenter récursivement la spécificité des entités traitées, et donc de construire progressivement une taxonomie complète. Sur DBpedia, notre approche

est capable de reconstituer la taxonomie de référence sur les classes existantes, mais aussi de décrire ces classes au moyen d'axiomes logiques et d'identifier de nouvelles classes pertinentes.

ABSTRACT

Knowledge graphs are the backbone of the Semantic Web, and have been succesfully applied to a wide range of areas. Many of these graphs are built automatically or collaboratively, and aggregate data from various sources. In these conditions, automatically creating and updating a taxonomy that accurately reflects the content of a graph is an important issue.

However, among scalable taxonomy extraction approaches, most of them can only extract a hierarchy on existing classes, and are unable to identify new classes from the data. In this thesis, we propose a novel taxonomy extraction method based on knowledge graph embeddings that is both scalable and expressive. A knowledge graph embedding model provides a dense, low-dimensional vector representation of the entities of a graph, such that similar entities in the graph are embedded close to each other in the embedding space.

Our goal is to show how these graph embeddings can be combined with unsupervised hierarchical clustering to extract a taxonomy from a graph. We first show that unsupervised clustering is able to extract a taxonomy on existing classes. Then, we show that it can also be used to identify new classes and organize them hierarchically, thus creating an expressive taxonomy.

For the non-expressive taxonomy extraction task, we introduce two methods for mapping existing classes to clusters of entities. The first of these methods solves a linear optimization problem in order to find an optimal injective function from classes to clusters. The second one can be seen as a smoothed version of the first one, designed to better handle noise and uncertainty in the data. In both cases, the resulting mapping is used to transform the clustering tree into a taxonomy. We run experiments with these two methods on DBpedia, and show that they both outperform a method based on supervised clustering.

For the expressive extraction task, we propose an axiom extraction method that leverages the clustering tree to define positive and negative samples, and induces new axioms from these samples. Since samples are chosen based on the similarity of their embeddings, this method effectively narrows down the search space to relevant subsets of the full graph. We also add a resampling mechanism, which allows us to extract increasingly specific axioms. We try our method on DBpedia, and show that the predicted taxonomy is able to rebuild the reference taxonomy with good precision, and that it can also identify new relevant classes and describe them with logical axioms.