Mesure et enrichissement sémantiques des modèles à base d'embeddings pour la prédiction de liens dans les graphes de connaissances

(Accessit du Prix de Thèse EGC 2025)

Nicolas Hubert*

* Université de Lorraine

La prédiction de liens est une composante fondamentale dans l'enrichissement des graphes de connaissances. Elle s'attache à inférer de nouvelles relations dans un graphe existant afin d'étendre notre compréhension du domaine représenté. Cette thèse se focalise sur des modèles générant des représentations vectorielles des entités et des relations, connues sous le nom d'embeddings. Se basant sur l'état actuel de la recherche sur les modèles à base d'embeddings, trois limitations principales sont identifiées :

— Manque de ressources sémantiquement riches. Les jeux de données traditionnellement utilisés en prédiction de liens manquent de profondeur et de variété en termes d'informations disponibles pour entrainer des modèles neuro-symboliques. Cette rareté de jeux de données enrichis de contenu sémantique entrave l'utilisation des modèles à base d'embeddings dans des applications nécessitant une compréhension nuancée et une interprétation des données.

— Cadre d'évaluation unidimensionnel. Les méthodes d'évaluation existantes, qui se concentrent largement sur les métriques basées sur le rang, offrent une vue limitée de la performance d'un modèle. Bien que ces métriques mesurent efficacement la précision de la prédiction de liens, elles ne tiennent pas compte d'autres aspects cruciaux tels que la richesse sémantique et la pertinence contextuelle des liens prédits.

— Manque de considérations sémantiques dans les approches basées sur l'apprentissage automatique. Les modèles actuels reposent essentiellement sur les principes de l'apprentissage automatique. Ces modèles ne peuvent donc pas capturer les nuances sémantiques et subtilités contextuelles présentes dans les graphes de connaissance. L'absence d'approche neuro-symbolique signifie que de tels modèles sont limités à ce qui peut être évalué quantitativement, négligeant souvent les couches sémantiques qualitatives des données.

Face à ces limitations, cette thèse explore les questions de recherche suivantes :

QR1. Comment concevoir des ressources sémantiquement enrichies pour encourager le développement de modèles neuro-symboliques, en particulier dans le cadre d'applications basées sur des graphes de connaissances telles que la prédiction de liens ?

QR1. Est-il possible de proposer de nouvelles métriques qui évaluent de manière qualitative différents aspects de ces modèles, notamment leur capacité à prédire des liens sémantiquement cohérents ?

QR1. En utilisant des jeux de données sémantiquement enrichis et de telles métriques, peut-on développer des approches d'entraînement neuro-symboliques capables de tirer pleinement parti de la richesse des connaissances disponibles ?

Cette thèse s'attaque aux limitations existantes en créant des ressources sémantiquement enrichies, telles que des graphes de connaissances, et en explorant des approches visant à améliorer la compréhension et l'évaluation sémantique des modèles basés sur les embeddings. Les principales contributions de cette recherche sont les suivantes :

1. Enrichissement et publication de jeux de données : Création de versions sémantiquement enrichies de jeux de données issus de l'état de l'art pour la prédiction de liens (QR1).
2. Conception d'une ontologie et d'un graphe de connaissances : Développement et mise à disposition d'une ontologie (EducOnto) et d'un graphe de connaissances (EduKG) dans le domaine de l'orientation scolaire (QR1).
3. Outil PyGraft : Développement, maintenance et diffusion de PyGraft, un outil Python open-source permettant de générer des ontologies et des graphes de connaissances synthétiques, entièrement personnalisables par l'utilisateur (QR1).
4. Nouvelle métrique sémantique : Introduction de Sem@K pour évaluer les capacités sémantiques des modèles à base d'embeddings (QR2).
5. Réévaluation des modèles à base d'embeddings : Réexamen approfondi des modèles d'embeddings de l'état de l'art sous l'angle de leur performance sémantique (QR2).
6. Développement d'approches neuro-symboliques : Conception de méthodes neuro-symboliques évaluées non seulement pour leurs capacités sémantiques, mais aussi pour leur utilisation de connaissances sémantiques lors de l'entraînement afin d'améliorer les performances prédictives (mesures basées sur le rang) (QR3).
7. Étude des approches neuro-symboliques : Analyse approfondie des méthodes neuro-symboliques et de leurs applications, en particulier dans les systèmes de recommandation (QR3).