Enrichissement de graphes de connaissances avec des algorithmes de graphes et d'apprentissage

Encadrants : Cédric du MOUZA (dumouza@cnam.fr, lab. CEDRIC, CNAM Paris)
Raphaël FOURNIER-S'NIEHOTTA (Raphael.Fournier@lip6.fr, LIP6-ComplexNetworks, Sorbonne Université)
Camelia CONSTANTIN (camelia.constantin@lip6.fr, LIP6-Équipe BD, Sorbonne Université)

Contexte. Un graphe de connaissance (*knowledge graph*) organise des informations sous forme de nœuds et d'arêtes. Les nœuds représentent des *entités*, personnes, lieux ou objets, et les arêtes expriment la nature des *relations* entre celles-ci (affiliations, localisations, propriétés, ...). Cette structure peut représenter des informations d'articles de presse, de fiches historiques voire de *logs* de sites de vente en ligne, offrant ainsi une modélisation sémantique efficace. Grâce à l'analyse automatisée, ces graphes permettent de détecter et visualiser des tendances et des motifs (*patterns*) dans de grands ensembles de données textuelles.

Problématique. La construction d'un graphe de connaissances à partir de données textuelles implique plusieurs étapes reposant principalement sur l'apprentissage automatique et le traitement du langage naturel (*Natural Language Processing* ou NLP). D'abord, le processus extrait du texte les entités représentant des personnes, des lieux, des organisations ou d'autres concepts pertinents. Ensuite, l'extraction de relations détermine l'interaction entre ces entités. Des modèles supervisés ou semi-supervisés sont utilisés pour reconnaître et classifier ces relations avec précision. Les techniques NLP, telles que l'analyse syntaxique et la compréhension contextuelle à l'aide de modèles linguistiques (comme BERT ou GPT), jouent un rôle essentiel dans l'interprétation des nuances du langage et l'extraction d'informations exactes. La dernière étape est le liage d'entités (*entity linking*), qui vise à identifier et associer les entités extraites à leurs équivalents dans le graphe. Cette tâche peut être complexe lorsque les informations sont fragmentaires ou ambiguës. Par exemple, une entité apparaissant isolément ou dans des contextes variés rend difficile son association à un nœud précis, surtout si les relations ne se répètent pas. De plus, la désambiguïsation se complique lorsque plusieurs nœuds correspondent à une même mention en l'absence de relations clairement définies. Ces difficultés expliquent pourquoi le liage d'entités est souvent l'étape la moins performante. Lorsqu'un liage n'est pas détecté, l'entité est insérée plusieurs fois dans le graphe, augmentant sa taille et dégradant sa qualité pour les usages futurs. Il est donc crucial de développer des méthodes innovantes et des stratégies efficaces pour détecter et fusionner les entités redondantes, notamment via un post-traitement avancé.

Objectif scientifique. L'objectif de la thèse est d'améliorer l'enrichissement de graphes de connaissances en utilisant des techniques avancées issues des domaines des algorithmes de graphes, de la sémantique et de l'apprentissage automatique. Cette démarche vise à améliorer la qualité et l'utilité du graphe en découvrant et en intégrant des informations qui ne sont pas explicitement présentes mais qui peuvent être inférées à partir des relations et des attributs existants.

Méthodologie. Nous combinerons des algorithmes de graphes avec des techniques d'apprentissage automatique, pour analyser la structure et identifier des motifs ou des clusters récurrents. La recherche de chemins ou l'analyse de la centralité révèlent des entités influentes ou des connexions inattendues, utiles pour prédire de nouvelles relations ou pour renforcer les connexions existantes. En utilisant des modèles de langage et des ontologies, on peut déduire des relations sémantiques non évidentes, comme des liens entre concepts similaires ou contextuellement liés. Les techniques d'apprentissage automatique pourront prédire de nouvelles entités et relations : entraînés à identifier des motifs relationnels sur des parties du graphe, ils infèrent des relations similaires dans des zones inexplorées. Ensuite, nous identifierons les entités à fusionner dans le graphe, afin de résoudre les problèmes de redondance et d'ambiguïté lorsque des entités identiques sont considérées comme distinctes. L'utilisation d'algorithmes de graphes, tels que la détection de communautés ou le *clustering* basé sur la similarité structurelle et sémantique, permet de regrouper des entités similaires, facilitant leur fusion. Nous exploiterons également des techniques de *machine learning* pour comprendre et interpréter les données en profondeur, détectant des similarités subtiles entre les entités. Des modèles comme Word2Vec, BERT ou GPT, entraînés sur de vastes corpus de texte, évalueront la similarité sémantique entre les descriptions ou attributs des entités. Des classifieurs pourront également être entraînés pour prédire si deux entités doivent être fusionnées, en utilisant des caractéristiques dérivées de leurs attributs et de leurs connexions dans le graphe.

Données. Nous allons valider nos approches sur la base de données prosopographiques Studium[1] dans laquelle les mêmes individus ou lieux apparaissent plusieurs fois avec une description ou nom très différents suivant la source (donc des propriétés et relations différentes) conduisant à la multiplication des nœuds dans la base de connaissances. D'autres jeux de données comme KnowledgeNet2[2] pourront être également utilisés.

[1] http://studium.univ-paris1.fr/
[2] https://paperswithcode.com/dataset/knowledgenet

Justification de l'approche scientifique. L'enrichissement des graphes de connaissances s'appuie sur une approche hybride combinant algorithmes de graphes, NLP et apprentissage automatique, qui permet d'extraire et d'intégrer des informations implicites en exploitant les structures et relations existantes. Un défi majeur réside dans le liage d'entités, étape souvent imprécise qui entraîne une fragmentation due au manque de relations explicites et à la diversité des formulations. L'objectif est d'améliorer cette phase en détectant et fusionnant les entités redondantes via un post-traitement reposant sur des modèles de similarité et des techniques de *clustering*. Ainsi, l'approche renforce la qualité structurelle et sémantique du graphe, le rendant plus exploitable pour la recherche et la gestion des connaissances. Pour atteindre ces objectifs, nous combinons : (i) des algorithmes de graphes pour identifier motifs récurrents, connexions et communautés, (ii) des modèles d'apprentissage profond (BERT, GPT) pour évaluer la similarité sémantique, et (iii) des modèles prédictifs entraînés pour inférer de nouvelles relations.

Adéquation à l'institut. Ce projet propose une approche exploratoire pour la construction et l'enrichissement de graphes de connaissances en combinant méthodes d'intelligence artificielle (IA), sur graphes (*GraphML*) ou autour du texte, avec des algorithmes classiques de graphes, se plaçant ainsi au cœur des thématiques de recherche de SCAI. Les contributions attendues visent à développer et adapter des méthodes IA faisant un usage hybride d'algorithmes de graphe et d'apprentissage automatique, tout en apportant des solutions pour enrichir les graphes de connaissances dans le domaine des bases de données. Les résultats de cette recherche seront appliqués à la base Studium de l'Université Panthéon-Sorbonne, dans le cadre d'une collaboration existante entre des membres du LAMOP et les encadrants. Outre des publications de haut niveau en informatique, nous envisageons une valorisation dans le cadre des journées du Consortium Européen Héloïse[3] pour, entre autres, valider l'applicabilité de la méthode à travers différents cas d'usage. Enfin, la diffusion des résultats pourra se faire au sein de SCAI dans le cadre d'ateliers ou séminaires (ISIR, LIP6), pour des chercheurs et étudiants intéressés par l'utilisation d'algorithmes de graphes et d'IA sur graphes.

Rôle et compétences scientifiques des encadrants. Les méthodes de ce projet s'appuient sur l'expertise complémentaire des encadrants, en optimisation de requêtes *big data*, en analyse de grands graphes, gestion de données et apprentissage sur graphes. L'équipe ComplexNetworks étudie les interactions à large échelle, en développant des modèles d'apprentissage sur graphes [4, 5] et de raisonnement dans des graphes temporels [3]. L'équipe Bases de Données du LIP6 se focalise sur le partitionnement de graphes multicouches (structure et types d'arêtes) [2] et son application au calcul distribué d'embeddings [1], ainsi que sur la détection de sujets émergents dans les archives scientifiques [8]. L'équipe ISID du CNAM, experte en représentation et qualité des données, travaille en Humanités Numériques (projets Mastodons QualHis, ANR Daphné[4]) et sur l'enrichissement de bases de connaissances à partir de textes [6, 7]. Une collaboration entre les Laboratoires LIP6 (SU), CEDRIC (CNAM) et LAMOP (Université Panthéon-Sorbonne) s'inscrit dans le projet ANR Laura en cours de soumission.

Profil de l'étudiant recherché. Titulaire d'un M2 ou ingénieur, avec de solides bases en informatique et en apprentissage automatique et idéalement de bonnes notions de graphes. La maîtrise d'un langage de programmation (comme Python) est indispensable.

Date limite pour candidater : 01/05/2025.

Références

[1] Y. Bai, C. Constantin, and H. Naacke. Leiden-fusion partitioning method for effective distributed training of graph embeddings. In *ECML PKDD 2024*, volume 14947, pages 366–382. Springer, 2024.

[2] C. Constantin, C. du Mouza, and Y. Li. A label-based edge partitioning for multi-layer graphs. In *52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019*, pages 1–10, 2019.

[3] V. David, R. Fournier-S'niehotta, and N. Travers. NeoMaPy : A Parametric Framework for Reasoning with MAP Inference on Temporal Markov Logic Networks. In *CIKM 2023*, pages 400–409. ACM, 2023.

[4] Y. Karmim, M. Lafon, R. Fournier-S'niehotta, and N. Thome. Supra-Laplacian Encoding for Transformer on Dynamic Graphs. In *NeurIPS'24*, 2024.

[5] Y. Karmim, E. Ramzi, R. Fournier-S'niehotta, and N. Thome. ITEM : improving training and evaluation of message-passing based gnns for top-k recommendation. *Transactions on Machine Learning Research (TMLR)*, 2024.

[6] M. Prieur, C. du Mouza, G. Gadek, and B. Grilhères. Evaluating and improving end-to-end systems for knowledge base population. In *ICAART*, pages 641–649, 2023.

[7] M. Prieur, C. du Mouza, G. Gadek, and B. Grilhères. Shadowfax : Harnessing textual knowledge base population. In *SIGIR*, 2024.

[8] H. Rahimi, H. Naacke, C. Constantin, and B. Amann. ANTM : aligned neural topic models for exploring evolving topics. *Trans. Large Scale Data Knowl. Centered Syst.*, pages 76–97, 2024.

[3] Héloïse : `https://heloise.hypotheses.org/`, congrès d'historiens spécialisés en prosopographie.
[4] `https://daphne-anr.huma-num.fr/`