Construction de graphes de connaissances historiques à l'aide d'algorithmes de graphes, LLMs et RAG

Mots-clés : LLM, RAG, GNN, bases de données graphe (Neo4j)

Début du stage : Février/Mars 2026
Durée et gratification du stage : 6 mois, 3 600 € (environ 600 euros par mois)
Lieu : LIP6 (Sorbonne Université), Équipe Bases de Données http://www-bd.lip6.fr/)
Encadrants du stage : Camelia CONSTANTIN (camelia.constantin@lip6.fr, LIP6-Équipe BD), Cédric du MOUZA (dumouza@cnam.fr, CNAM), Raphaël FOURNIER-S'NIEHOTTA (Raphael.Fournier@lip6.fr, LIP6-Équipe ComplexNetworks)

Collaborations : travaux de recherche communs entre historiens et informaticiens, réalisés dans le cadre du projet ANR LAURA, collaborations pendant le stage avec différents partenaires du projet (University of Perugia-Italie, CNAM, Université Panthéon-Sorbonne, Archives Nationales)

Possibilité de poursuivre en thèse : financement de thèse existant
Contexte : Ces recherches portent sur les bases prosopographiques ayant pour objet la période médiévale. La prosopographie est une méthode des sciences sociales (sociologie, histoire) dans laquelle on cherche à analyser un groupe à partir d'une étude systématique des itinéraires singuliers des individus qui le composent. Pour cela les chercheurs collectent tous les faits (factoïdes) possibles sur chaque individu. En histoire médiévale, ces données sont rares, discontinues, incertaines et souvent d'une qualité médiocre. En effet, les experts de ces disciplines gèrent la qualité et l'incertitude dans le temps et l'espace. Ainsi, les personnes sont désignées par plusieurs noms, les lieux changent de noms et de frontières avec le temps ou selon l'auteur et un parcours de diplomation peut changer en fonction de l'époque, du lieu ou de la classe sociale de la personne. En raison de cette complexité, de nombreuses règles restent opaques pour les historiens médiévistes. En organisant ces données sous forme de graphe de connaissances, nous pouvons représenter des concepts, des personnes, des lieux ou des objets sous forme des nœuds et des interactions entre ceux-ci, comme telles que des affiliations ou des localisations comme arêtes avec des propriétés. Cela permet une organisation sémantique des données qui aide à mieux comprendre les contextes et les relations complexes entre les entités mentionnées dans les textes et de visualiser clairement l'interconnexion entre les différents éléments d'un ensemble de données.
Problématique : La construction de graphes de connaissances à partir de sources de données ambiguës soulève plusieurs difficultés majeures liées à la nature imparfaite, hétérogène et souvent imprécise des textes ou bases d'origine. Les principales sources de complexité résident dans l'ambiguïté des entités (lorsqu'un même nom peut désigner plusieurs objets distincts ou, inversement, lorsqu'une entité apparaît sous des formes lexicales variées (abréviations, translittérations, synonymes). S'ajoutent à cela l'imprécision ou l'incomplétude des informations (par exemple, des dates approximatives ou des localisations vagues), qui favorisent la **duplication d'entités** lors de l'intégration : plusieurs nœuds représentant en

réalité la même entité peuvent être créés. Les textes peuvent également contenir des relations implicites difficiles à extraire automatiquement, ou des contradictions entre sources multiples.

Parmi ces problèmes, certains peuvent être atténués par l'usage de **modèles de langage de grande taille (LLMs)**, capables d'intégrer des indices sémantiques complexes et de désambiguïser les entités en tenant compte du contexte global. Les LLMs améliorent la normalisation des alias multilingues, la reconnaissance des relations implicites et la cohérence sémantique entre textes hétérogènes. Cependant, plusieurs défis demeurent : les modèles ne résolvent pas les ambiguïtés lorsque deux entités très similaires apparaissent dans le texte sans marqueurs de distinction, ni les contradictions entre sources (le LLM peut les reconnaître mais ne garantit pas de choisir la bonne version, ou peut donner une confiance excessive à l'une sans justification). Ils n'éliminent pas non plus les duplications induites par des informations incomplètes ou imprécises, et peuvent introduire de nouveaux biais, notamment par hallucination d'entités ou surestimation de leur confiance. Dans des domaines historiques, le LLM peut manquer de données de formation spécifiques, ce qui réduit sa performance. Ces limites exigent l'intégration de méthodes complémentaires, telles que la reconnaissance d'entités nommées (NER), qui impose un typage explicite et stable des entités (personne, organisation, lieu, date, etc.), en permettant ainsi de filtrer les entités erronées ou inventées. En s'appuyant sur des lexiques, ontologies ou dictionnaires de référence, elle facilite la normalisation et l'alignement des entités, limitant ainsi la création d'alias ou de doublons. Des méthodes complémentaires, comme *entity linking* ou l'utilisation de règles symboliques peuvent aider à l'alignement des entités obtenues par des LLMs.

L'ajout de réseaux de neurones de graphes (**GNN**) améliore la désambiguïsation d'entités lorsque le contexte textuel seul est insuffisant, en s'appuyant sur les voisins et les motifs relationnels pour identifier la bonne correspondance, en exploitant la structure relationnelle du graphe. Ils renforcent également la détection de doublons et la fusion d'entités similaires en apprenant des représentations qui intègrent à la fois les attributs et les connexions locales.
Objectif du stage : Afin de pouvoir améliorer le liage d'entités, il est important d'avoir le maximum d'information pour chaque entité, et notamment ses liens avec les autres entités. L'objectif de ce stage est la mise en œuvre d'une architecture RAG-GNN intégrée, destinée à la construction, à la détection des duplicats et à la fusion d'entités d'un graphe de connaissances construit à partir de données prosopographiques Studium[1] ambigües. Ces données existent sous forme de fiches où les mêmes individus ou lieux apparaissent plusieurs fois avec une description très différente suivant la source (donc des propriétés et relations différentes) voire des noms parfois très différents. 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. D'autres jeux de données comme KnowledgeNet[2] pourront être également utilisés.
Méthodologie : La méthodologie combine **des** modèles de langage préentraînés, recherche contextuelle**,** et apprentissage de représentations de graphes pour la construction et la consolidation de graphes de connaissances à partir de données historiques ambiguës. Dans une première étape, un **modèle de langage de grande taille (LLM)** est utilisé pour extraire automatiquement des entités, relations et attributs (dates, lieux, personnes) à partir des textes. Cette extraction est renforcée par des méthodes de **reconnaissance d'entités nommées (NER)**, assurant une détection typée et une segmentation fiable des mentions, conformément aux principes décrits dans [1]. Le graphe ainsi obtenu sera stocké dans une base de données graphe (Neo4j) avec traçabilité des chunks sources et va constituer une base brute soumise à

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

un processus d'enrichissement et de validation des liens de duplication(sameAs)). Un module de **Retrieval-Augmented Generation (RAG)** sera utilisé dans une première phase pour une décision sameAs ou notSame pour chaque paire d'entités candidate, accompagnée de preuves textuelles. Le LLM reçoit le contexte enrichi (entité A, entité B, leurs propriétés, leurs relations adjacentes dans le graphe et leurs *chunks* sources) pour prendre une décision de fusion, en fournissant la provenance factuelle pour la décision de résolution d'entité [2]. Les **Graph Neural Networks (GNNs)** sont ensuite utilisés pour exploiter la structure relationnelle du graphe et apprendre des représentations topologiques capables d'identifier les clusters d'entités équivalentes [3]. Le GNN sélectionne et connecte des sous-graphes d'un graphe de connaissances qui sont ensuite convertis en entrées textuelles pour un LLM [4], permettant la mise en place d'une **boucle de rétroaction LLM–GNN** afin d'affiner progressivement les décisions de fusion : les inférences structurelles issues du GNN guident le LLM dans ses réévaluations contextuelles, tandis que les jugements linguistiques du LLM enrichissent les représentations structurelles apprises par le GNN.

Bibliographie

[1] Ji et al., "A Survey on Knowledge Graphs: Representation, Acquisition and Applications" *(*AI Open*,* 2021*)*.

[2] Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" *(*NeurIPS*,* 2020)

[3] Ivanisenko et al., An Accurate and Efficient Approach to Knowledge Extraction from Scientific Publications Using Structured Ontology Models, Graph Neural Networks, and Large Language Models, Int J Mol Sci. 2024 Nov 3;25(21)

[4] Mavromatis et al. GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning, ArXiv, 2024