le cnam cedric

Fusion d'entités dans des graphes de connaissances

Mots-clés : graphes de connaissances, apprentissage, algorithmes de graphe

Contexte : Un graphe de connaissances est une structure de données qui organise et représente des informations sous forme de nœuds (ou entités) et d'arêtes (ou relations). Ces nœuds peuvent représenter des concepts, des personnes, des lieux ou des objets, tandis que les arêtes décrivent les relations entre eux, telles que des affiliations, des localisations ou des propriétés. Cette représentation permet de visualiser clairement l'interconnexion entre les différents éléments d'un ensemble de données. On peut les utiliser pour stocker des informations extraites de collections de données textuelles, comme des articles de presse, des collections de fiches historiques ou encore des logs d'un site de vente en ligne. 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. Dans un corpus de nouvelles, par exemple, un graphe de connaissances peut relier des articles différents à travers des personnes communes, des lieux ou des événements mentionnés, offrant ainsi une vue d'ensemble des thèmes récurrents et des liens entre les nouvelles. Les graphes de connaissances offrent de plus une manière efficace de gérer l'information et de détecter par une analyse automatisée des tendances ou des patterns au sein de grands ensembles de données textuelles.

Problématique : La construction d'un graphe de connaissances à partir de données textuelles implique plusieurs étapes, principalement basées sur des techniques d'apprentissage automatique et de traitement du langage naturel (NLP). Initialement, le processus extrait des entités représentant des noms de personnes, lieux, organisations, ou d'autres concepts pertinents présents dans le texte. L'extraction de relations, détermine par la suite l'interaction ou l'association entre les entités. Des modèles d'apprentissage supervisé ou semi-supervisé sont souvent employés pour entraîner des systèmes à reconnaître et classifier ces relations de manière précise. Les techniques de NLP comme l'analyse syntaxique et la compréhension de texte assistée par des modèles linguistiques contextuels (comme BERT ou GPT) jouent un rôle crucial dans la compréhension des nuances du langage et l'extraction précise de l'information. Enfin la dernière étape est le liage d'entités (entity linking) qui consiste à identifier et à lier des entités à leurs équivalents dans le graphe de connaissances, ce qui peut être difficile lorsque les informations disponibles sont fragmentaires ou ambiguës. Par exemple, si une entité apparaît isolément ou dans des contextes très variés, il devient complexe de déterminer avec précision à quel nœud du graphe elle devrait être associée. Le manque de relations récurrentes ou significatives complique la tâche car les algorithmes dépendent souvent de la répétition et de la cohérence des relations pour faire des associations précises. De plus, le problème de désambiguïsation des entités, où plusieurs nœuds peuvent correspondre à une seule mention dans le texte, est encore plus important quand les entités n'ont pas de relations clairement définies qui aideraient à les distinguer. Ceci explique que cette étape dans le processus d'enrichissement d'un graphe de connaissances est l'étape avec les moins bonnes performances

en terme de précision et de rappel. Lorsqu'un liage n'est pas détecté, l'entité est insérée dans le graphe de connaissances qui pourrait être ainsi amené à contenir de nombreuses occurrences d'une même entité qui n'ont pas été fusionnées avec pour conséquence une augmentation de la taille du graphe et une dégradation de sa qualité pour des futures utilisations.

Objectif de la thèse : L'objectif de cette thèse est de développer des méthodes innovantes pour améliorer cette étape cruciale du liage d'entités dans les graphes de connaissances et de proposer des stratégies efficaces pour détecter et fusionner les entités redondantes ou séparées à tort par le biais d'un post-traitement avancé. Ce travail se concentrera sur l'exploitation combinée d'algorithmes de graphes et de techniques d'apprentissage automatique pour relever ces défis.

Méthodologie : Pour atteindre cet objectif, nous envisageons de travailler dans un premier temps sur l'enrichissement du graphe de connaissances à partir des données existantes 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. Les algorithmes de graphe peuvent être utilisés pour analyser la structure du graphe et identifier des motifs ou des clusters récurrents. Par exemple, des techniques comme la recherche de chemins ou l'analyse de la centralité peuvent révéler des entités influentes ou des connexions inattendues entre différentes parties du graphe. Ces découvertes peuvent ensuite être utilisées pour prédire de nouvelles relations ou pour renforcer les connexions existantes. En exploitant des modèles de langage et des ontologies, on peut déduire des relations sémantiques qui ne sont pas immédiatement évidentes, comme les liens entre des concepts similaires ou contextuellement liés. Quant aux techniques d'apprentissage automatique, elles permettent de généraliser à partir des exemples existants pour prédire de nouvelles entités et relations. Par exemple, des modèles prédictifs peuvent être entraînés sur des parties du graphe pour identifier des motifs de relations, et ces modèles peuvent ensuite être utilisés pour inférer des relations similaires dans des régions non explorées du graphe. L'apprentissage profond, en particulier, peut être utilisé pour traiter et interpréter de grandes quantités de données textuelles pour extraire et intégrer de nouvelles informations pertinentes dans le graphe.

Dans un deuxième temps nous allons identifier dans le graphe construit précédemment les entités qui devraient être fusionnées. Cette étape sera essentielle pour résoudre les problèmes de redondance et d'ambiguïté qui surviennent lorsque des entités similaires ou identiques sont traitées comme distinctes. L'utilisation d'algorithmes de graphes, tels que la détection de communautés ou les techniques de clustering basées sur la similarité structurelle et sémantique, permettra de regrouper les entités qui partagent des attributs ou des contextes similaires, facilitant ainsi leur fusion. On exploitera aussi des techniques de machine learning qui offrent des outils puissants pour comprendre et interpréter les données à un niveau plus profond, permettant de détecter des similarités peu évidentes entre les entités. Par exemple des modèles comme Word2Vec, BERT ou GPT, entraînés sur de vastes corpus de texte, peuvent être utilisés pour évaluer la similarité sémantique entre les descriptions ou les attributs des entités. Les entités dont les vecteurs de caractéristiques sont très proches peuvent être fusionnées. Des classifieurs peuvent aussi 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.

Jeux de données : Nous allons valider nos approches sur la base de données prosopographiques Studium[1]. Dans ce jeu de données 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, conduisant à la multiplication des nœuds au sein de la base de connaissances. D'autres jeux de données comme KnowledgeNet[2] pourront être également utilisés.

Directeur de thèse : Cédric du Mouza, dumouza@cnam.fr

Lieu de la thèse : Laboratoire CEDRIC, CNAM

Collaborations pendant la thèse avec le LIP6 (Sorbonne Université, Paris VI) et LAMOP (Université Panthéon-Sorbonne, Paris I)

Bibliographie

Stefano Faralli, Andrea Lenzi, Paola Velardi: A Benchmark Study on Knowledge Graphs Enrichment and Pruning Methods in the Presence of Noisy Relationships. J. Artif. Intell. Res. 78: 37-68 (2023)

Fatiha Saïs: Knowledge Graph Refinement: Link Detection, Link Invalidation, Key Discovery and Data Enrichment. University of Paris-Sud, Orsay, France, 2019

Victor Telnov, Yuri Korovin: Machine Learning and Text Analysis in the Tasks of Knowledge Graphs Refinement and Enrichment. DAMDID/RCDL (Supplementary Proceedings) 2020: 48-62

Dalei Zhang, Qiang Yang, Zhixu Li, Junhua Fang, Ying He, Xin Zheng, Zhigang Chen: Tail Entity Recognition and Linking for Knowledge Graphs. APWeb/WAIM (1) 2020: 286-301

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, Xindong Wu: A Comprehensive Survey on Automatic Knowledge Graph Construction. ACM Comput. Surv. 56(4): 94:1-94:62 (2024)

[1] http://studium.univ-paris1.fr/

[2] https://paperswithcode.com/dataset/knowledgenet