Laboratoire LIP6

PLDAC - 2024

Annotation sémantique de fichiers CSV

Contact : Bernd AMANN (bernd.amann@lip6.fr)

1 ou 2 deux étudiants

Contexte

Les données tabulaires sous forme de fichiers CSV constituent le format d'entrée courant dans un pipeline d'analyse de données. Cependant, un manque de compréhension de la structure sémantique et de la signification du contenu peut entraver le processus d'analyse des données. L'acquisition de cette compréhension sémantique sera donc très utile pour l'intégration des données, le nettoyage des données, l'exploration des données, l'apprentissage automatique et les tâches de découverte des connaissances. Par exemple, la compréhension de la nature des données peut aider à évaluer les types de transformation appropriés sur les données.

L'appariement de données tabulaires à des graphes de connaissances (KG) [1] est le processus consistant à l'attribution d'étiquettes sémantiques provenant de graphes de connaissances (par exemple, Wikidata ou DBpedia) aux éléments du tableau. Cette tâche est cependant souvent Cette tâche est cependant souvent difficile dans la pratique car les métadonnées (par exemple, les noms des tables et des colonnes) sont manquantes, incomplètes ou ambiguës. des noms de tables et de colonnes) étant manquantes, incomplètes ou ambiguës.

Objectifs généraux

Ce projet PLDAC s'inscrit dans le contexte d'une collaboration avec la société Zeenea (https://zeenea.com/fr/) sur la construction de catalogues de données pour les lacs de données. L'objectif du projet est d'étudier les problèmes et les solutions pour l'appariement de données tabulaires à des glossaires spécialisés. Par rapport aux travaux existants qui considèrent des graphes de connaissances générales , les glossaires spécialisés sont souvent structurées sous forme d'hiérarchies de termes spécialisés par rapport un domaine d'application précis.

Ce projet pose plusieurs défis :

- Exploitation de métadonnées : Comment exploiter au mieux les informations contextuelles / métadonnées associées aux données (documentation, noms des tables, …) ?
- Génération de glossaires : Comment construire un glossaire automatiquement pour une collection de données (par exemple dans les archives Open Data)
- Choix des méthodes : Comment choisir les meilleures méthodes d'appariement pour une collection de données et un glossaire ?
- Evaluation : Comment évaluer la qualité des résultats obtenus ?

Travail à réaliser

- Evaluation et comparaison des méthodes existantes et en particulier les méthodes récentes fondées sur l'IA génériques (LLMs)
- Spécification et implémentation d'un prototype
- Mise en œuvre et validation expérimentale sur divers datasets

Ce projet pourra éventuellement être prolongé sous forme de stage M1.

Références bibliographiques (sélection):

[1] SemTab: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching
https://www.cs.ox.ac.uk/isg/challenges/sem-tab/

[2] Unicorn: A Unified Multi-tasking Model for Supporting Matching Tasks in Data Integration
https://dl.acm.org/doi/abs/10.1145/3588938

[3] Amin Beheshti et al. “CoreKG : a knowledge lake service”. en. In : Proceedings of the VLDB Endowment 11.12 (août 2018), p. 1942-1945. issn : 2150-8097. doi : 10.14778/3229863.3236230. url : https://dl.acm.org/doi/10.14778/3229863.3236230.

[4] Adriane Chapman et al. “Dataset search : a survey”. en. In : The VLDB Journal 29.1 (jan. 2020), p. 251-272. issn : 0949-877X. doi : 10.1007/s00778- 019-00564-x. url : https://doi.org/10.1007/s00778-019-00564-x (visité le 24/08/2022).

[5] Xiang Deng et al. “Turl : Table understanding through representation learning”. In : ACM SIGMOD Record 51.1 (2022). Publisher : ACM New York, NY, USA, p. 33-40

[6] Ahmed Helal et al. “A demonstration of KGLac : a data discovery and enrichment platform for data science”. In : Proceedings of the VLDB Endowment 14.12 (juill. 2021), p. 2675-2678. issn : 2150-8097. doi : 10.14778/3476311.3476317. url : https://doi.org/10.14778/3476311.3476317