Recherche de Tables Associées : IA et Modèles de Langage au Service des Données

Sujet PLDAC 2024 (1 ou 2 étudiants)

Encadrants :

- Bernd Amann (bernd.amann@lip6.fr)
- Allaa Boutaleb (allaa.boutaleb@lip6.fr)
- Rafael Angarite (rafael.angarita@lip6.fr)

Contexte et Motivation

La recherche de tables associées, ou "related table search", est une tâche qui vise à identifier les tables de données qui sont pertinentes par rapport à une table de référence (requête) donnée. Cette recherche peut être motivée par divers objectifs, tels que l'enrichissement de tables (union, jointure), la découverte de tables dans un datalake ou l'amélioration de la qualité des données (données manquantes). La notion de relation/association entre la table requête et les tables cherchées peut être large et englober la similarité du schéma, le chevauchement au niveau des enregistrements (tuples), les descriptions des données, etc. Dans ce projet nous nous intéressons en particulier à deux types des relations pour la recherche:

- Tables unionables : Ce sont des tables qui peuvent être combinées avec une table de requête pour ajouter plus de lignes, en partageant généralement des attributs du même domaine. Les colonnes des tables unionables doivent avoir une forte similarité sémantique.
- Tables joignables : Ces tables peuvent être jointes à une table de requête en fonction des valeurs qui se chevauchent dans une ou plusieurs colonnes. Elles doivent avoir des éléments de schéma communs avec un fort chevauchement de données.

La recherche de tables associées est une tâche complexe qui présente plusieurs défis. Un des défis est de tenir compte de la variété des schémas des tables, qui peuvent être inconsistantes, incomplètes ou manquantes. De plus, il faut tenir compte des relations sémantiques entre colonnes, qui ne sont pas toujours explicitées par les noms de colonnes. Enfin, il faut pouvoir gérer de grands volumes de données de manière efficace et scalable.

Plusieurs approches ont été développées pour la recherche de tables associées [2, 6, 7]. Ces méthodes incluent l'analyse des similarités des métadonnées (noms et description de tables et colonnes), l'évaluation des similarités des valeurs dans les colonnes (distance Jaccard), ainsi que l'utilisation de techniques d'apprentissage automatique pour apprendre des représentations vectorielles des tables. Les approches récentes exploitent des modèles de langage pré-entraînés (PLM) tels que BERT, permettant de mieux intégrer le contexte et la sémantique des données. Par ailleurs, des techniques d'indexation sont souvent utilisées pour améliorer l'efficacité de la recherche.

Objectifs du projet

L'objectif général du projet est d'analyser et d'améliorer les méthodes de recherche de tables associées en tenant compte de plusieurs aspects clés.

1. Le premier objectif du projet est d'analyser différentes méthodes de recherche de tables associées pour mieux comprendre les forces et faiblesses de chaque méthode par rapport à différents benchmarks. Les dimensions possibles à comparer sont la *robustesse* (sensibilité par rapport à l'ordre des colonnes et des lignes et ou la présence de métadonnées), le *coût* (calcul, mémoire, disque), la *performance* (indexation, interrogation), la *scalabilité* (nombre de tables, colonnes, lignes) des méthodes et la *qualité* (précision, recall, MAP@k, F1Score) des résultats.

2. Le deuxième objectif est d'étendre les méthodes existantes, en permettant aux utilisateurs de mieux préciser leurs besoins et leurs intentions lors de la recherche de tables liées. Cette extension peut consister à spécifier des tâches plus complexes qui intègrent des *connaissances extérieures* (contexte utilisateur, graphes de connaissances) et des *préférences* par rapport aux attributs/colonnes dans le processus de recherche.

Travail à réaliser

Le projet vise à analyser et améliorer les méthodes de recherche de tables liées, en s'appuyant sur le benchmark LakeBench [1].

Le travail à réaliser comprend les étapes suivantes :

1. Prise en main de LakeBench : Il s'agit de se familiariser avec le benchmark LakeBench (`https://github.com/RLGen/LakeBench`), qui comprend des tables provenant de diverses sources, telles que des données gouvernementales ouvertes, des données économiques et des graphes de connaissances. Le benchmark comprend des ensembles de données pour des tâches telles que l'identification de l'unionabilité, de la joignabilité et des sous-ensembles de tables.

2. Définition d'un protocole d'expérimentation : Cela implique le choix des mesures pour évaluer l'efficacité des différentes méthodes et la définition des expériences menées pour comparer différentes méthodes dans des conditions cohérentes et en utilisant une analyse statistique appropriée.

3. Expérimentation et analyse des résultats : Cela impliquera l'exécution des différents algorithmes sur les ensembles de données LakeBench, la mesure de leurs performances en utilisant les mesures établies et l'analyse des résultats afin d'identifier les forces et les faiblesses de chaque approche. Cela comprend la compréhension de la manière dont les algorithmes fonctionnent en fonction de facteurs tels que la taille des ensembles de données et les types de données. L'analyse doit tenir compte à la fois de l'efficacité et de l'efficience des différentes méthodes.

4. Sélection et extension d'une méthode existante : Cette tâche implique l'exploration de techniques pour combiner différentes notion de relation/association entre les tables. Ces techniques peuvent consister à mieux exploiter les métadonnées associées aux données (descriptions textuelles, schémas) et des informations contextuelles décrites par un graphe de connaissances [6].

Mots-clés

Data Discovery, Data Integration, Information Retrieval, Related Table Search, NLP, Benchmarks, PLM (Pre-trained Language Models), Semantic Similarity, Context-aware Search.

Équipes d'accueil et Contact

- Equipe Bases de Données du LIP6 (Paris): `http://www-bd.lip6.fr/`

References

[1] Yuhao Deng, Chengliang Chai, Lei Cao, Qin Yuan, Siyuan Chen, Yanrui Yu, Zhaoze Sun, Junyi Wang, Jiajun Li, Ziqi Cao, Kaisen Jin, Chi Zhang, Yuqing Jiang, Yuanfang Zhang, Yuping Wang, Ye Yuan, Guoren Wang, and Nan Tang. LakeBench: A Benchmark for Discovering Joinable and Unionable Tables in Data Lakes. *Proceedings of the VLDB Endowment*, 17(8):1925–1938, April 2024.

[2] Yuyang Dong, Kunihiro Takeoka, Chuan Xiao, and Masafumi Oyamada. Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 456–467, Chania, Greece, April 2021. IEEE.

[3] Yuyang Dong, Chuan Xiao, Takuma Nozawa, Masafumi Enomoto, and Masafumi Oyamada. DeepJoin: Joinable Table Discovery with Pre-Trained Language Models. *Proceedings of the VLDB Endowment*, 16(10):2458–2470, June 2023.

[4] Michael Günther, Maik Thiele, Julius Gonsior, and Wolfgang Lehner. Pre-Trained Web Table Embeddings for Table Discovery. In *Fourth Workshop in Exploiting AI Techniques for Data Management*, pages 24–31, 2021.

[5] Xuming Hu, Shen Wang, Xiao Qin, Chuan Lei, Zhengyuan Shen, Christos Faloutsos, Asterios Katsifodimos, George Karypis, Lijie Wen, and Philip S. Yu. Automatic Table Union Search with Tabular Representation Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3786–3800, Toronto, Canada, 2023. Association for Computational Linguistics.

[6] Aamod Khatiwada, Grace Fan, Roee Shraga, Zixuan Chen, Wolfgang Gatterbauer, Renée J. Miller, and Mirek Riedewald. SANTOS: Relationship-based Semantic Table Union Search. *Proceedings of the ACM on Management of Data*, 1(1):9:1–9:25, May 2023.

[7] Aamod Khatiwada, Harsha Kokel, Ibrahim Abdelaziz, Subhajit Chaudhury, Julian Dolby, Oktie Hassanzadeh, Zhenhan Huang, Tejaswini Pedapati, Horst Samulowitz, and Kavitha Srinivas. TabSketchFM: Sketch-based Tabular Representation Learning for Data Discovery over Data Lakes, December 2024.

[8] Kavitha Srinivas, Julian Dolby, Ibrahim Abdelaziz, Oktie Hassanzadeh, Harsha Kokel, Aamod Khatiwada, Tejaswini Pedapati, Subhajit Chaudhury, and Horst Samulowitz. LakeBench: Benchmarks for Data Discovery over Data Lakes, July 2023.

[9] Yi Zhang and Zachary G. Ives. Finding Related Tables in Data Lakes for Interactive Data Science. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1951–1966, Portland OR USA, June 2020. ACM.