L'alignement des schémas hétérogènes : approche basée sur des embeddings

Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste

L'alignement des schémas hétérogènes : approche basée sur des embeddings

Oumaima El Haddadi*,**,***, Max Chevalier*, Bernard Dousset*
Ahmad El Allaoui**, Anass El Haddadi **, Olivier Teste *

*IRIT, SIG
Université Paul Sabatier, Toulouse, France
**LSA, SDIC
Université Abdelmaled Essaadi, Tetouan, Maroc
*** elhaddadi.oumaima@gmail.com

Résumé. Les espaces de données tels que les lacs de données, reposent généralement sur plusieurs jeux de données («datasets») provenant de différentes sources de données hétérogènes impliquant différents schémas qui doivent cohabiter. Pour interroger ces espaces de données on se retrouve alors face à différents schémas posant des problèmes de redondance et de complémentarité des données. La difficulté dans ce cadre est de gérer de façon dynamique cet ensemble de schéma hétérogène voire dynamique permettant malgré tout de retrouver les données pertinentes en réponse à un besoin d'analyse . Dans ce cadre, notre objectif est d'étudier particulièrement l'alignement automatique des différents schémas. Notre proposition repose sur des méthodes de plongements sémantiques (« embeddings» en anglais) afin d'identifier les alignements pertinents.

1 Introduction

Parmi les techniques proposées par la communauté scientifique pour arriver à aligner les données on trouve la mise en correspondance des schémas. Cette dernière est une technique permettant l'identification des objets qui sont sémantiquement et/ou contextuellement liés. Cette phase sert à trouver des correspondances entre les concepts provenant de différentes sources de données présentes dans notre espace de données et qui sont hétérogènes (schéma différents).Dans la littérature, les chercheurs dans ce domaine ont défini plusieurs objectifs à l'utilisation de correspondance des schémas (« schema matching » en anglais) : tels que l'exploration des données, la recherche des jointures entre les jeux des données, l'enrichissement d'un espace de données et l'alignement de données. (Aumueller et al., 2005), (Bernstein et al., 2011), (Rahm et Bernstein, 2001),(Alserafi, 2021). Notre objectif est d'aligner automatiquement les différents schémas provenant de plusieurs sources via l'utilisation des méthodes basées sur le plongements sémantiques. On présente dans ce résumé la méthodologie adopté et les perspectives à venir.

2 Méthodologie

Notre méthode consiste à incorporer les informations (attributs ou contenu) de schéma par l'utilisation de la méthode du plongement de graphe pour capturer les relations entre les éléments de données en ce qui concerne leur sémantique et le contexte qu'ils partagent. Cette méthode est composée, de trois étapes :

1-Préparation de graphe Cette étape consiste à transformer tous les schémas des jeux de données dans un seule graphe, nommée *S=(A,R)* avec *A* qui présente les attributs et *R* qui présente les relations entre les attributs. Notant aussi qu'un sous-graphe *S'=(A',R')* représente un schéma d'un jeu de donnée avec $A' \subset A$.

2-Préparation des vecteurs La première étape après la lecture des données est de générer des phrases «sentence» depuis un sous-graphe par des chemins aléatoires dans le graphe (ou « Random Walk ») afin de pouvoir appliquer les techniques de plongement. A la suite on va utiliser ces phrases pour entraîner le modèle de plongement que l'on souhaite utiliser (e.g. node2vec). En sortie nous obtiendrons un vecteur caractérisant chaque sous-graphe.

3-Correspondance des schémas Sur la base de ces vecteurs, on calcule la distance (e.g. distance cosinus) entre chaque attribut dans l'espace vectoriel afin d'obtenir la correspondance des schémas.

3 Conclusion et Perspectives

Pour évaluer l'efficacité de notre approche, nous envisageons de réaliser des tests sur des jeux de données provenant de deux domaines fonctionnels différents : concernant les films (IMBD et MovieLens) et les brevets. Par la suite, nous élargirons nos expérimentations en menant une étude comparative entre notre méthode et les autres méthodes identifiées dans la littérature (e.g. EmbdI (Cappuzzo et al., 2020), REMA (Koutras et al., 2020)). Cette étude pourra reposer sur d'autres jeux de données disponibles publiquement.

Références

Alserafi, A. Dataset proximity mining for supporting schema matching and data lake governance.

Aumueller, D., H.-H. Do, S. Massmann, et E. Rahm. Schema and ontology matching with COMA++. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data - SIGMOD '05*, pp. 906. ACM Press.

Bernstein, P. A., J. Madhavan, et E. Rahm. Generic schema matching, ten years later. *4*(11), 695–701.

Cappuzzo, R., P. Papotti, et S. Thirumuruganathan. Local embeddings for relational data integration. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 1335–1349.

Koutras, C., M. Fragkoulis, A. Katsifodimos, et C. Lofi. REMA : Graph embeddings-based relational schema matching. pp. 4.

Rahm, E. et P. A. Bernstein. A survey of approaches to automatic schema matching. *10*(4), 334–350.