

Appariement de schémas relationnels

Fabien Duchateau (prénom.nom@univ-lyon1.fr)

Contexte. L'appariement de schémas consiste à détecter les correspondances entre concepts équivalents de schémas hétérogènes [1]. Dans ce projet, nous nous intéressons aux schémas relationnels, et l'appariement exploite donc les relations (tables), attributs, ainsi que les contraintes de clé primaire ou de clé étrangère [2]. En particulier, on souhaite détecter des schémas relationnels qui se ressemblent fortement (e.g., dans un but d'intégration de données ou de refactoring de base de données). Par exemple, les 3 schémas ci-dessous (sous forme de code SQL) contiennent chacun une table avec des informations sur des animaux. En définissant des critères de similarité, il est possible d'évaluer la similarité entre ces schémas (e.g., le schéma de gauche est similaire à 63% avec celui du milieu). En perspective, on pourra étudier l'explicabilité de la similarité calculée.

```
CREATE TABLE Animal (  
  idA INTEGER PRIMARY KEY,  
  nomA VARCHAR(42),  
  espèceA VARCHAR(42)  
);
```

```
CREATE TABLE Animals (  
  id INTEGER PRIMARY KEY,  
  name VARCHAR(50),  
  specie VARCHAR(50)  
);
```

```
CREATE TABLE Animaux (  
  nom VARCHAR(100) UNIQUE,  
  espèce VARCHAR(100)  
);
```

Objectifs. Ce projet cherche à atteindre les objectifs suivants :

- Dresser un court état de l'art sur la similarité entre schémas relationnel;
- Définir des critères (e.g., nom des relations, nombre d'attributs en commun) et des mesures de similarité (e.g., égalité stricte, Levenhstein, Jaro-Winkler) pour comparer deux schémas relationnels;
- Définir des métriques de comparaison entre 2 schémas relationnels (e.g., pourcentage de tables avec un nom similaire), en utilisant les critères et mesures de similarité précédemment sélectionnées;
- Développer un script Python qui prenne en entrée un ensemble de fichiers SQL (contenant la création de schémas), et qui calcule un score global de similarité entre chaque paire de schémas.

Bibliographie.

- [1] P. A. Bernstein, J. Madhavan, and E. Rahm. Generic schema matching, ten years later. *Proceedings of the VLDB Endowment*, 4(11):695–701, 2011. [Lien VLDB2011](#).
- [2] M. Labreche, X. Lorca, A. Montarnal, S. Weill, J.-P. Adi, and T. Sébastien. A general approach for of schema matching problem: case of databases. *Preprint to KIS*, 2022. [Lien KIS2022](#).