Université Claude Bernard Lyon 1

Appariement de Schémas Relationnels

Cahier des charges

Lisa IAMRACHE p2312973
Tamazgha SOUTOU p1813609
Steven WAKENHUT p2207505

I. Présentation du contexte :

Ce projet s'inscrit dans le cadre de l'UE M1if11, qui vise à initier les étudiants au monde de la recherche après une première immersion dans le monde de l'entreprise à la fin de la licence. Ce travail est supervisé par Monsieur Fabien Duchateau, Enseignant-chercheur à l'Université Claude Bernard Lyon 1, qui a proposé le sujet “appariement de schémas relationnels”.

L'appariement de schémas, également connu sous le nom de "Schema matching" en anglais, fait référence à la mise en correspondance de schémas issus de différentes sources de données.
Un schéma représente une structure formelle qui permet d’organiser des données, par exemple un schéma relationnel (SQL), un schéma DTD ou XSD (XML), un schéma de validation JSON, etc.
Une "correspondance" est une relation entre un ou plusieurs éléments d'un schéma et un ou plusieurs éléments d’un autre schéma.
La figure 1 montre un exemple d’appariement entre 2 sources, avec 5 correspondances identifiées (lignes rouges).

Dans ce projet, nous nous intéressons aux schémas relationnels, qui a pour objectif de détecter des correspondances entre des schémas de bases de données relationnelles hétérogènes. Ces correspondances sont essentielles pour l'intégration de données et la maintenance de bases de données. Le projet a pour objectif de calculer un score de ressemblance entre deux schémas relationnels.

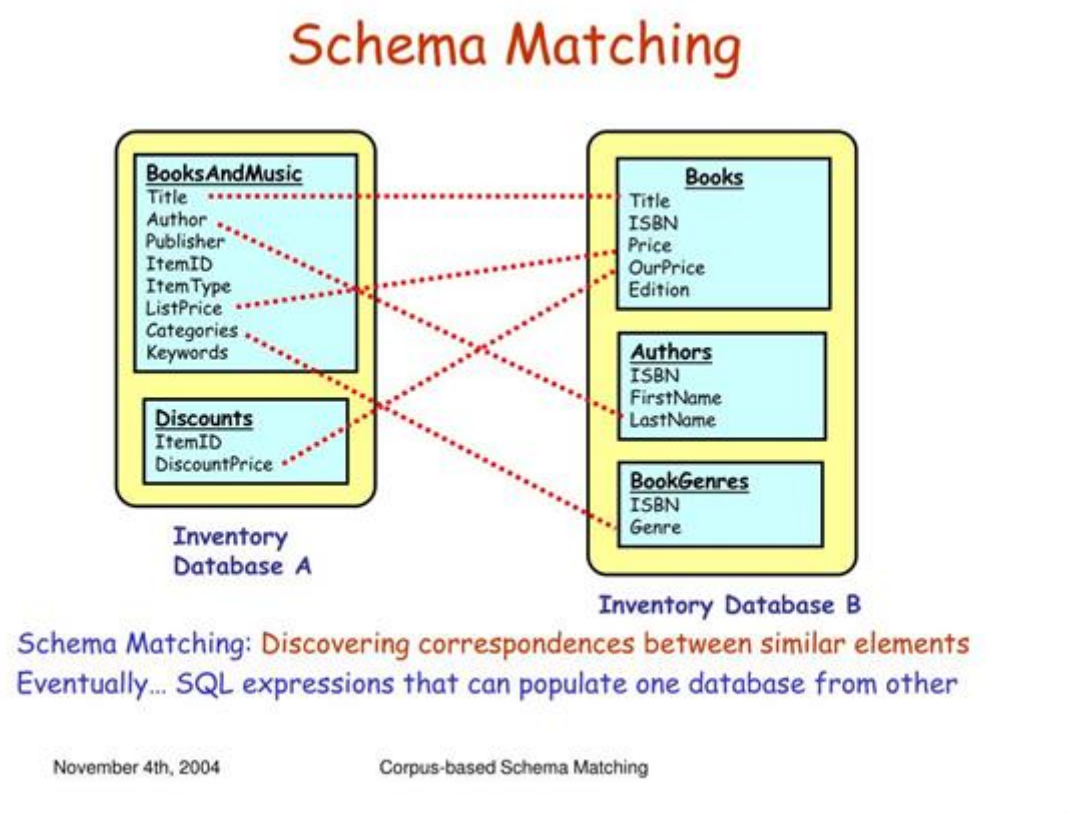

Figure 1. Exemple d’appariement de schémas entre deux sources A et B. Un exemple de correspondance est “ListPrice” avec “Price”.[4]

II. Objectifs :

Le projet vise à atteindre les objectifs suivants :

1. État de l'art : Réaliser un court état de l'art sur les méthodes et techniques existantes pour l'appariement de schémas, et éventuellement pour mesurer la similarité entre schémas relationnels [1, 2].

2. Critères et mesures de similarité : définir des critères de similarité, tels que le nom des relations, le nombre d'attributs en commun, etc. Également, sélectionner des mesures de similarité, comme l'égalité stricte, la distance de Levenshtein, Jaro-Winkler, pour comparer deux schémas relationnels [3].

3. Métriques de Comparaison : Concevoir des métriques de comparaison entre deux schémas relationnels en utilisant les critères et mesures de similarité définis précédemment. Par exemple, calculer le pourcentage de tables avec des noms similaires. Les métriques seront suffisamment flexibles (ex, poids, seuils) afin de tester différentes configurations.

4. Développement du Script Python : Mettre en place un script Python capable de prendre en entrée un ensemble de fichiers SQL contenant les définitions de schémas (sources de données) et de calculer un score global de similarité entre chaque paire de schémas.

III. Livrables :

Le projet devra fournir les livrables suivants :

- Un rapport d'état de l'art sur la similarité entre schémas relationnels.
- Une spécification détaillée des critères de similarité et des mesures à utiliser.
- Une définition des métriques de comparaison entre schémas relationnels.
- Une implémentation (script Python) pour le calcul de la similarité entre schémas.
- Une documentation claire et concrète pour l'utilisation du script.
- Une analyse et présentation des résultats obtenus avec des exemples concrets.

IV. Calendrier et Échéances :

Les échéances du projet sont les suivantes :

Dates	Tâches
01/11/23 à 15/11/23	Réalisation de l'état de l'art.
16/11/23 à 05/12/23	Définition des critères et mesures de similarité.
06/12/23 à 31/12/23	Mise en place des métriques de comparaison.
01/01/24 à 30/01	Développement du script Python.
01/02/24 à 27/02/24	Tests, documentation et préparation du rapport.
À voir plus tard	Présentation des résultats.

V. Références :

[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.

[3] Najlah Gali, Radu Mariescu-Istodor, and Pasi Fränti. Similarity measures for title matching. In 2016 23rd International Conference on Pattern Recognition, pages 1548–1553. IEEE, 2016. Lien ICPR2016.

[4] URL de la figure : schema matching