Appariement de schémas semi-structurés

Projet de Recherche

Réalisé par : Tanguy SMODIS,
Étudiant en M1 Informatique à
l'Université Claude Bernard Lyon 1

Encadrant : Fabien DUCHATEAU,
Enseignant-Chercheur au LIRIS et
Maître de conférence à l'UCBL

UNIVERSITÉ DE LYON

Université Claude Bernard Lyon 1

Contexte du projet de recherche

Depuis l'apparition du web, le modèle de données relationnel très populaire jusqu'alors a été mis à mal par le volume, la vitesse de génération et la variété des informations circulant sur le web. On appelle ce phénomène "Big Data" et les besoins toujours croissants ont poussé l'innovation vers de nouveaux systèmes de gestion de données dont ceux orientés documents.

Ces systèmes sont moins contraignants que le modèle relationnel afin de pouvoir répondre aux besoins en performances, en disponibilité et en flexibilité. Ils sont par conséquent moins structurés, on dit ces modèles "flexibles" voire "sans schéma" bien qu'il existe une organisation cachée des données exploitables pour l'appariement.

L'appariement de schémas de données consiste à régler les problèmes qui surviennent quand on cherche à intégrer ou fusionner des données provenant de différentes bases pour être capable de les exploiter de manière uniforme[1].

On pourrait prendre l'exemple des hôpitaux d'une métropole pour lesquels on souhaite agréger les données, les termes utilisés pour désigner les soignants peuvent varier (docteurs, médecins, praticiens), la façon de représenter une opération menée par un soignant sur un patient (les deux sont référencés par des identifiants, l'un est inclus et l'autre référencés, les deux sont inclus). L'appariement doit tenir compte de ces possibles différences d'hétérogénéité et assembler ces structures en un tout cohérent qui conserve un maximum d'informations.

ID_Operation	Docteur	Patient
9	8	Ho WARD
22	34	Lalo DEGUZMAN
128	26	Charles McGILL

ID_Operation	Médecin	Patient
6	James McGILL	Ho WARD
32	Kimberley WEXLER	Lalo DEGUZMAN
326	Michael EHRMANTRAUT	Charles McGill

Il existe déjà de nombreuses études[2][3] concernant l'appariement de schémas dans les bases de données utilisant un modèle relationnel, mais peu de travaux portent sur les sources semi-structurées. Cela peut se comprendre par la relative jeunesse des modèles orientés documents dont l'apparition remonte à l'émergence du web et du phénomène "Big Data" et l'absence explicite d'un schéma ce qui complexifie le processus d'intégration. En relationnel, les différences d'hétérogénéité structurelle se limitent aux concepts table, attribut et clé étrangère tandis qu'en document, il y a clé-valeur, liste, sous-document et des liens de référence ou d'imbrication ce qui est structurellement plus riche.

[1] Philip A Bernstein, Jayant Madhavan, and Erhard Rahm. Generic schema matching, ten years later. Proceedings of the VLDB Endowment, 4(11):695–701, 2011.
[2] Ding, G., Sun, S. & Wang, G. Schema matching based on SQL statements. Distrib Parallel Databases 38, 193–226 (2020).
[3] Philip A. Bernstein, Sergey Melnik, Michalis Petropoulos, and Christoph Quix. Industrial-strength schema matching. 2004.

Detailed schema

```
{
   "Publication": "FoodNetwork",
   "Comments": [
      {
         "ID": 907,
         "Content": "Love this recipe!"
      },
      {
         "ID": 909,
         "Content": "I've had better"
}]}
```

No Schema

```
{
   "FoodNetwork": [
      [
         907,
         "Love this recipe!"
      ],
      [
         909,
         "I've had better"
      ]
]}
```

Autre exemple extrait d'un article[4], on peut voir ici que les deux fichiers n'ont pas la même structure (dictionnaire de dictionnaire à gauche, dictionnaire de tableaux à droite). Sur la version No Schema il n'y a pas de nom pour les données, on ne sait donc pas le type de données qu'on stocke. Il n'y a pas de règles définissant ce qui fait partie du schéma et ce qui fait partie des données, à gauche la "FoodNetwork" est une valeur, à droite c'est une clé; à gauche les identifiants sont encodés dans des objets, à droite dans des tableaux.

Ce projet de recherche s'oriente sur l'appariement de schémas de données semi-structurés basés sur des formats tels que le BSON/JSON utilisé notamment par MongoDB et d'autres systèmes de gestion de base de données orientés documents.

Ce projet est mené dans le cadre de l'UE Ouverture à la Recherche de l'université Claude Bernard Lyon 1.

Objectifs du projet

- Dresser l'état de l'art sur l'appariement de schémas JSON.
- Rédiger un rapport présentant les verrous scientifiques liés à ce genre de schémas.
- Présenter des solutions possibles aux-dits verrous scientifiques.
- Définir un banc d'essai (benchmark) avec ses métriques d'évaluation et ses jeux de données pour l'appariement de schémas JSON.

[4] Kunal Waghray. JSON Schema Matching: Empirical Observations, 2019,

Travail à effectuer

1. Dresser l'état de l'art et rapport des verrous

1.1. Lecture avec prises de notes des différents articles proposés par l'enseignant encadrant.

1.2. Recherche éventuelle d'autres articles concernant l'appariement de schémas JSON dans l'optique d'élargir l'horizon des verrous et des solutions.

1.3. Globaliser et comparer les notes prises lors de la lecture des-dits articles afin de mettre des termes sur les différents verrous et solutions.

1.4. Rédaction syntaxique de l'état de l'art.

2. Développement du benchmark

2.1. Conception d'un jeu de données.

2.2. Établissement des métriques d'évaluation.

2.3. Test du benchmark.

2.4. Rédaction d'un compte-rendu des tests.

Planning prévisionnel