ÉLÉMENT DE PORTFOLIO 03

Vidéo

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : Inférence Interactive de Schema JSON

URL de l'élément : `https://dropsu.sorbonne-universite.fr/s/7nf6RQz7FTRR8zK`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Il s'agit d'une vidéo accompagnant la démonstration d'un outil présenté dans une conférence internationale [1]. Cette démonstration est l'implantation de l'approche d'inférence interactive de schéma dont les fondements théoriques ont été présentés dans [2]. Elle s'appuie également sur un effort d'implantation ayant permis la réalisation des expérimentations des publications décrivant l'inférence de schémas paramétriques [3] qui a suscité beaucoup d'intérêt auprès de la communauté puisque de nombreuses extensions ont récemment été publiées, les plus importantes étant [5] et [4] publiés dans les conférences majeurs du domaine (SIGMOD et EDBT).

3 PRÉSENTATION DE CET ÉLÉMENT

JSON est le format de données semi-structurées le plus utilisé. Il tient son succès de son expressivité et de sa flexibilité puisqu'il permet de décrire des données complexes sans définir de schéma préalablement. Cette liberté s'accompagne de nombreux défis lorsqu'il s'agit d'exploiter de manière correcte et efficace les données représentées dans ce format qui sont souvent fortement imbriquées et volumineuses. L'inférence de schéma *a posteriori* est une solution attractive pour exploiter de manière efficace ces données puisqu'elle permet d'extraire une description succincte de la structure sous-jacente des données et ainsi de formuler des requêtes et des programmes d'extraction de données sensés.

Or, une même collection JSON peut être décrite suivant différents niveaux d'abstraction et différents usages font appels à différents niveaux de précision. Il devient alors plus judicieux d'offrir la possibilité de choisir de manière fine le niveau de précision souhaité pour chaque fragment de données, ce qui, d'un point de vue théorique, revient à définir une technique d'inférence non déterministe, guidée par l'interaction avec l'utilisateur et efficace en présence de grands volumes de données.

Pour répondre à ces contraintes, nous avons proposé une approche interactive s'appuyant sur des techniques de réécriture d'arbres [2] et garantissant, de manière formelle, la correction des schémas obtenus par simple réécriture de schémas plus précis. L'avantage est de pouvoir utiliser seulement les schémas, pendant l'interaction, sans avoir à accéder aux données qui sont souvent volumineuses et empêcheraient toute interaction fluide avec l'utilisateur.

La réalisation logicielle présentée dans [1] exploite cette technique de réécriture et démontre la faisabilité de l'approche sur des schémas inférés de données réelles de taille importante et présentant de nombreuses variations structurelles qui tendent à confirmer l'utilité de l'approche dans le cas pratique.

Une extension récente visant à générer automatiquement des requêtes à partir de l'interaction de l'utilisateur a été mise en œuvre et a permis de démontrer l'utilité d'une approche de découverte des données, guidée par les schémas.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Mohamed-Amine Baazizi, Clément Berti, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Human-in-the-Loop Schema Inference for Massive JSON Datasets. In *EDBT 2020 - 23nd International Conference on Extending Database Technology*, pages 635–638, Copenhagen, Denmark, March 2020. OpenProceedings.org.

[2] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. A Type System for Interactive JSON Schema Inference (Extended Abstract). In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 101 :1–101 :13, Patras, Greece, July 2019.

[3] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Parametric schema inference for massive JSON datasets. *The VLDB Journal*, 28(4) :497–521, August 2019.

[4] Hanâ Lbath, Angela Bonifati, and Russ Harmer. Schema inference for property graphs. In *EDBT 2021-24th International Conference on Extending Database Technology*, pages 499–504, 2021.

[5] William Spoth, Oliver Kennedy, Ying Lu, Beda Hammerschmidt, and Zhen Hua Liu. Reducing ambiguity in json schema discovery. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1732–1744, 2021.