Exploration interactive de datasets JSON volumineux à base de schémas

Encadrant :

Mohamed-Amine Baazizi (mohamed-amine.baazizi@lip6.fr)

Ces dernières années ont connu un engouement particulier pour JSON qui s'est imposé comme format de représentation des données massives. Réseaux sociaux, forums, API sont autant d'exemple d'applications où le format de données utilisées est JSON. La majorité de ces applications n'utilisent pas de schémas car l'accent est mis sur la flexibilité et l'agilité. Ceci prive les data-scientist d'une information capitale, celle de la structure des données sous-jacentes. Par conséquent, il devient très compliqué de formuler des requêtes correctes et des programmes d'extraction performants.

En dépit de ce constat, très peu de techniques ont été mises en oeuvre pour l'extraction de schémas à partir de données JSON. Récemment, nous avons proposé dans [1] et [2] une approche qui pallie à ce manque. Cette technique permet d'inférer un schéma global reflétant toutes les variations structurelles d'un jeu de données JSON. L'originalité de notre approche est qu'elle permet de traiter de manière efficace de larges volumes en des temps relativement courts mais surtout de produire un schéma compact et compréhensible par les utilisateurs. La compacité est certes une propriété désirée mais engendre dans certains cas une perte d'information (cf. [3]). Par ailleurs, l'inférence de schémas précis va à l'encontre de l'objectif principal qui est de rendre le schéma exploitable qui requiert forcément de produire des schémas succincts. Pour combler ce fossé, nous envisageons de proposer une technique interactive permettant aux utilisateurs de raffiner les schémas obtenus et de choisir le niveau de précision/d'abstraction souhaité. L'idée est d'avoir des schémas sur lesquels on peut faire des zoom avant/arrière selon le besoin.

Les objectifs de ce travail sont :

1- Prendre en main le système développé dans [1] et [2].
2- Proposer une technique d'inférence d'un espace de schémas précis et étudier les liens de raffinement/abstraction entre ces schémas.
3- Implanter cette technique en utilisant le système développé dans [1] et [2].
4- Expérimenter l'approche sur des jeux de données existants.
5- Crawler de nouveaux jeux de données et étudier l'impact de la nouvelle approche.
6- Proposer une technique de visualisation permettant d'assurer une interaction fluide entre utilisateurs et système d'inférence.

Les prérequis :

1- Capacité d'analyse et de résolution de problèmes, travail en équipe.
2- Programmation impérative et fonctionnelle (la maitrise de Scala serait un plus).
3- Conception d'applications web.

Références

[1] Mohamed-Amine Baazizi, Houssem Ben Lahmar, Dario Colazzo, Giorgio Ghelli et Carlo Sartiani. Inférence de Schémas pour Données JSON Massives. BDA 2016, Poitiers.

[2] Mohamed-Amine Baazizi, Houssem Ben Lahmar, Dario Colazzo, Giorgio Ghelli et Carlo Sartiani. Schema Inference for Massive JSON Datasets. EDBT 2017, Venice (To appear).

[3] http://webia.lip6.fr/~baazizi/research/json/pldac/json-schema-extensions.pdf