Exploration utilisateur de lacs de données très hétérogènes

Nelly Barret*, Simon Ebel*, Théo Galizzi*, Ioana Manolescu*, Madhulika Mohanty*

*Inria et Institut Polytechnique de Paris
prenom.nom@inria.fr

1 Scénario : les lacs de données très hétérogènes

Les dernières décennies ont vu une importante production de données digitales. Ces données couvrent de nombreux domaines, e.g. la santé, l'environnement et la finance. Elles sont détenues et utilisées par différents acteurs, qui ne sont pas toujours les producteurs, e.g. les journalistes pour les applications de *data journalism*. L'hétérogénéité des données apporte de nombreux défis pour leur intégration, exploration et compréhension. Les **lacs de données**, tels que ceux décrits dans Hai et al. (2016, 2023); Nargesian et al. (2019); Sawadogo et Darmont (2021), sont des répertoires centralisés destinés à stocker et rendre accessible de grandes quantités de données, souvent structurées. ConnectionLens (Anadiotis et al. (2022)) est un lac de données intégrant des sources hétérogènes dans un paradigme *graphe*, afin de capturer précisément la structure que ces données (semi-, non-)structurées peuvent avoir. De plus, des techniques d'Extraction d'Information identifient, dans les valeurs (feuilles), des entités, e.g. personnes, ou entreprises ; très intéressantes pour les journalistes.

2 Contexte : intégration en graphe de données hétérogènes

ConnectionLens ingère n'importe quelles données structurées, semi-structurées et non-structurées comme suit. Les documents XML conduisent à un noeud pour chaque élément, attribut ou valeur ; les relations parent-enfant deviennent des arêtes. Pour un document JSON, chaque dictionnaire, liste, valeur (feuille) est converti en un noeud du graphe. Pour les graphes RDF, chaque triple `s,p,o` produit deux noeuds labellisés "s" et "o" connectés par une arête labellisée `p`. Pour les données CSV et relationnelles, chaque tuple et valeur deviennent des noeuds, connectés par des arêtes labellisées avec le nom des colonnes (si existant). Les documents texte sont segmentés en paragraphes, qui deviennent des noeuds, tous enfants d'un noeud racine commun. Les documents Office et PDF sont convertis en JSON puis ingérés.

Ensuite, de la reconnaissance d'entités nommées (*NER*) est appliquée sur chaque noeud feuille du graphe, ce qui mène à de (nouveaux) noeuds labellisés avec le nom des entités reconnues et sont connectés au noeud feuille dont ils ont été extraits via une arête. Quand deux noeuds entité sont identiques, i.e. ont le même label, ils sont fusionnés et un seul noeud est gardé dans le graphe. Ceci permet de facilement trouver des connexions *inter-sources*.

3 Nouveaux modules pour ConnectionStudio

Collaborant avec des journalistes, nous avons compris qu'ils trouvaient difficile de : (α) relier le graphe intégré aux documents qu'ils ajoutaient ; (β) trouver des mots clés intéressants à requêter ; (γ) télécharger des documents concrets à partir du graphe, qu'ils partageront entre collègues ; (δ) réparer les erreurs introduites dans les entités extraites par la NER. Nous avons donc développé **ConnectionStudio** (Barret et al. (2023)), une nouvelle plateforme qui étend ConnectionLens pour répondre aux souhaits (α) à (δ). Les contributions de ConnectionStudio sont : (i) un ensemble de statistiques calculés au nivau du lac ; (ii) des résumés structurels des fichiers ingérés sous la forme de diagrammes Entité-Relation ; (iii) une interface de requêtage simplifiée où les utilisateurs peuvent composer leurs requêtes à partir de blocs élémentaires de données ; (iv) une interface simplifiée pour corriger les valeurs erronnées et les erreurs faites par les modèles d'extraction. ConnectionStudio est accessible en ligne, avec des exemples et tutoriels, à `https://connectionstudio.inria.fr/fr/`.

4 Conclusion et perspectives

Les lacs de données tels que ConnectionLens (Anadiotis et al. (2022)) ont pour but d'aider les utilisateurs à explorer des sources hétérogènes. ConnectionLens adopte un paradigme graphe pour intégrer ces sources et extrait les entités menant à des opportunités de connexion des données. Via ConnectionStudio, nous décrivons de nouveaux paradigmes d'exploration et de découverte de données, se basant sur les souhaits des journalistes. Les utilisateurs peuvent ainsi *découvrir* le graphe, *requêter de manière simplifiée* les connections entre les sources, et faire du *nettoyage incrémental* du graphe. Nous planifions d'étudier comment ConnectionStudio aide les utilisateurs novices à explorer des graphes et inclure leurs retours.

Remerciements. Ce travail est partiellement financé par les bourses DIM RFSI PHD 2020-01, AI Chair SourcesSay (ANR-20-CHIA-0015-01) et CQFD (ANR-18-CE23-0003). Nous remercions aussi Camille Pettineo, qui a contribué à cet article en tant que journaliste.

Références

Anadiotis, A., O. Balalau, C. Conceicao, et al. (2022). Graph integration of structured, semi-structured and unstructured data for data journalism. *Inf. Systems 104*.

Barret, N., S. Ebel, T. Galizzi, I. Manolescu, et M. Mohanty (2023). User-friendly exploration of highly heterogeneous data lakes. In *CoopIS*.

Hai, R., S. Geisler, et C. Quix (2016). Constance : An intelligent data lake system. In *SIGMOD*, New York, NY, USA.

Hai, R., C. Koutras, C. Quix, et M. Jarke (2023). Data lakes : A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering*.

Nargesian, F., E. Zhu, R. J. Miller, K. Q. Pu, et P. C. Arocena (2019). Data lake management : challenges and opportunities. *Proceedings of the VLDB Endowment 12*(12).

Sawadogo, P. et J. Darmont (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems 56*, 97–120.