Sujet de stage M2

Cartographie d'un datalake apprise par renforcement avec budget contraint

Ce stage est en co-encadrement entre les équipes LFI et BD du LIP6 et réalisé dans le cadre d'un projet LIP6

Contacts BD : Hubert Naacke hubert.naacke@lip6.fr , Camelia Constantin camelia.constantin@lip6.fr
Contact LFI : Jean-Noël Vittaut jean-noel.vittaut@lip6.fr

Contexte

Les datalakes permettent de centraliser et d'analyser différents types de données avec des structures hétérogènes (incluant des données non structurées ou semi-structurées comme par exemple des images, fichiers audio et vidéo et documents), provenant de sources variées. De grands volumes de données sont stockées et peuvent être explorées dans leur format d'origine, l'accès aux données avant même qu'elles soient nettoyées, transformées ou structurées offrant ainsi un plus grand potentiel de traitement de données. Les activités subséquentes d'apprentissage et d'analyse sont facilitées et les résultats d'analyses peuvent être obtenus plus rapidement.

L'accès à des données de qualité est crucial pour les processus d'analyse de données, la qualité des résultats étant étroitement dépendante de la qualité des données utilisées, en absence de mécanismes appropriés pour les adapter à des fins d'apprentissage et d'analyse, ces données pouvant être introuvables ou peu fiables. Afin de les rendre plus accessibles, des outils de préparation des données ont vu le jour. Ce processus de préparation de données peut s'avérer particulièrement complexe ce qui rend impossible en pratique la préparation et le nettoyage de l'ensemble des données d'un datalake. C'est pourquoi il devient essentiel d'établir au préalable une **cartographie** (une représentation des thématiques contenues dans les fichiers) des données afin d'identifier celles qui seront nettoyées.

Tâches et sujet de stage M2

Le but du stage est de **concevoir une nouvelle méthode pour établir la cartographie d'un datalake** pour permettre à un utilisateur d'identifier les fichiers en lien avec une thématique posée, tout en satisfaisant une contrainte de budget donnée. Le travail à réaliser est composé de trois tâches :

T1 : Approche comparative. Nous considérerons que l'ensemble des collections obtenues par la solution ConnectionLens représente une cartographie qui nous servira de référence. L'objectif de cette tâche est de comparer les cartographies obtenues lorsque seule une partie des fichiers est lue. Cette tâche a pour but de mesurer l'impact d'un accès partiel au datalake sur la qualité de la cartographie obtenue. Il s'agira de caractériser les critères de sélection pour choisir un fichier à lire et pour choisir la portion des données à lire dans un fichier. Une fonction permettant d'évaluer la qualité relative d'une cartographie par rapport à une référence sera définie.

T2 : Apprentissage d'une stratégie d'exploration des fichiers. Cette tâche s'appuie sur le résultat de la tâche précédente : il est possible d'obtenir une cartographie de qualité satisfaisante en accédant à une faible partie du datalake, cependant il est nécessaire de choisir habilement la partie des données lues. Ainsi, l'objectif de cette tâche est de définir une méthode pour apprendre, par renforcement, une stratégie permettant d'explorer un sous-ensemble du datalake. Le but de cette stratégie est d'atteindre les parties apportant des informations utiles pour la cartographie et d'éviter de lire des données moins utiles. L'apprentissage par renforcement a déjà été utilisé avec succès dans le domaine de l'analyse de données exploratoire [1] et peut être une piste de départ pour la cartographie d'un datalake. De part la nature hiérarchique du datalake et l'aspect incrémental de l'exploration, un apprentissage de politique par une méthode de la famille Monte-Carlo Tree Search (MCTS) est envisagé [2]. Cet apprentissage permet d'effectuer un compromis en l'*exploitation* des informations recueillies lors des précédentes explorations du datalake ; et l'*exploration* qui permet de s'intéresser aux parties les moins explorées. L'impact du budget sur la qualité de la stratégie sera évalué. Le bénéfice en termes de budget économisé par rapport à d'autres stratégies sans apprentissage par renforcement sera mesuré.

T3 : Découverte et prise en compte de l'intégrité référentielle. Dans [3] les auteurs proposent de calculer des jointures entre les tables d'un SGBD relationnel pour former des tuples capturant les informations d'intégrité référentielle entre les données. Les tuples obtenus servent d'exemple pour entraîner un modèle de mots existant (Fasttext). La représentation obtenue permet, par clustering, de regrouper des termes liés à des concepts latents communs. L'objectif de cette tâche est d'étendre la méthode permettant d'extraire des collections d'entités en la complétant avec la méthode proposée dans [3] lorsque des informations d'intégrité référentielle entre des fichiers ont été détectées. Un avantage intéressant de cette approche est la possibilité d'anonymisation offerte par la représentation statistique du contenu du datalake. Une méthode pour évaluer le niveau de confidentialité garanti par cette approche sera proposée.

Divers

Le stage se déroulera au LIP6 (métro Jussieu) pour une durée de 5 à 6 mois, à partir de mars 2022.
L'indemnité mensuelle de stage est d'environ 600 euros.

Bibliographie

[1] Ori Bar El, Tova Milo, Amit Somech. Automatically Generating Data Exploration Sessions Using Deep Reinforcement Learning. ACM SIGMOD Conference 2020: Pages 1527–1537 https://dl.acm.org/doi/abs/10.1145/3318464.3389779
[2] C. Browne, E. Powley. A survey of Monte Carlo Tree Search Methods. IEEE Transactions on Intelligence and AI in Games. 2012; 4(1): 1–49 https://ieeexplore.ieee.org/document/6145622
[3] Riccardo Cappuzzo, Paolo Papotti, Saravanan Thirumuruganathan: Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks. ACM SIGMOD Conference 2020: 1335-1349 https://dl.acm.org/doi/10.1145/3318464.3389742