

Sujet : Profilage des ensembles de données dynamiques

Le profilage des données est une tâche fondamentale qui sous-tend la gestion des données et les tâches analytiques, y compris le traitement des données et l'apprentissage automatique. Un ingrédient clé du profilage des données est la découverte des contraintes/dépendances qui régissent un ensemble de données en explorant l'ensemble de données lui-même.

En raison de l'importance de ce problème dans la pratique, le monde universitaire et l'industrie ont construit divers outils de profilage pour aider les scientifiques des données dans leur quête de dépendances de données. Ces outils offrent un bon support pour la découverte de dépendances dans le contexte de statique où l'ensemble de données ne change pas. Dans des contextes dynamiques, comme le web, où les ensembles de données sont sujets à des changements (mises à jour, insertion et suppression), les solutions existantes ne sont pas adaptées, ce qui nécessite une nouvelle famille d'algorithmes et d'outils qui prennent en charge le support incrémental de la découverte de dépendances.

L'objectif de ce stage est d'étudier les classes d'algorithmes existantes qui conviennent à un cadre dynamique en examinant et en adaptant les algorithmes qui ont été proposés dans des cadres statiques.

Selon le background et les souhaits de l'étudiant, le stage peut être orienté vers l'analyse des propriétés mathématiques du problème, sa complexité et/ou la proposition de nouveaux algorithmes et la réalisation informatique d'une évaluation empirique.

L'étudiant bénéficiera d'une supervision étroite et acquerra une première expérience dans la conduite de recherche universitaire dans un domaine qui est pertinent à la fois pour les mathématiques et l'informatique, et qui est d'actualité à l'ère du Big Data.

Mots-clés : Big Data, profilage des données, gestion des connaissances et de l'information.

Contacte : Khalid Belhajjame (kbelhajj@gmail.com)

LAMSADE, Université Paris Dauphine

Référence :

Ziawasch Abedjan, Lukasz Golab, Felix Naumann, Thorsten Papenbrock: Data Profiling. Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2018. Accessible at <https://tinyurl.com/ysvbjwcr>