Soutenance de Thèse

Institut de Recherche
en Informatique de Toulouse

Lundi 12 Octobre 2020
14h00
UT3 Paul Sabatier, IRIT, Auditorium J. Herbrand

Max HALFORD
Equipe PYRAMIDE, IRIT

Apprentissage statistique pour l'estimation de sélectivité en bases de données relationnelles

Jury :
- Lynda Tamine-Lechani – *Professeure*
- Chirine Ghedira Guegan – *Professeure*
- Aurélien Garivier – *Professeur*
- Nicolas Bousquet – *MCF*
- Philippe Saint-Pierre – *MCF*
- Franck Morvan – *Professeur*

Résumé : Les bases de données relationnelles sont couramment utilisées pour stocker et interroger des données. Les données y sont stockées dans des relations qui sont liées entre elles. Un utilisateur interroge ces données via des requêtes exprimées dans un langage déclaratif. Un des défis pour le Système de Gestion de Bases de Données (SGBD) est d'évaluer les requêtes le plus rapidement possible. Pour cette évaluation, il existe un large choix de plans d'exécution associé à une requête donnée. Le SGBD délègue le choix du plan d'exécution le plus efficace à un optimiseur. Dans ce processus, le modèle de coûts, quand à lui, a la responsabilité d'estimer le temps de calcul de chaque plan d'exécution considéré par l'optimiseur. Par conséquent, le modèle de coûts influe sur la qualité des plans d'exécution produit par l'optimiseur, et donc sur le temps de réponse perçu par l'utilisateur.

Le paramètre le plus influant dans le modèle de coûts est la sélectivité, qui permet d'estimer la quantité de données qui transite entre chaque opérateur d'un plan d'exécution. D'un point de vue statistique, ceci correspond à de l'estimation de densité, qui consiste à déterminer la quantité de données qui vérifie un ensemble de conditions. Il est d'usage courant d'utiliser des modèles avec des hypothèses simplificatrices pour des raisons de performances. Par exemple, il est souvent supposés qu'il n'existe pas de dépendances de valeurs entre différents attributs de relations. Dans la majeure partie des applications cette hypothèse se révèle non vérifiée et conduit à de grandes erreurs d'estimation. Le but de cette thèse est donc de proposer des modèles plus réalistes qui offrent un meilleur compromis entre la précision des estimations et la complexité calculatoire requise. Premièrement, nous proposons une méthode s'appuyant sur les réseaux Bayésiens dont la représentation sous forme d'arbre nous permet d'améliorer la précision des estimations avec une complexité linéaire. Par la suite, nous explorons l'usage de modèles d'apprentissage en ligne pour corriger un modèle d'estimation existant tout en s'adaptant à l'évolution des données. Nos résultats expérimentaux basés sur les bechmaks JOB et TPC-DS montrent des résultats très encourageants.

05 61 55 65 10
info@irit.fr
www.irit.fr