Optimisation et analyse interactive de données : le Problème du Voyageur de Données

Directeurs de thèse

Nicolas LABROCHE (nicolas.labroche@univ-tours.fr), Patrick MARCEL (patrick.marcel@univ-tours.fr) et Vincent T'KINDT (tkindt@univ-tours.fr)

Équipe d'accueil

Laboratoire d'Informatique Fondamentale et Appliquée de Tours (EA 6300 LIFAT) – Equipe Recherche Opérationnelle, Ordonnancement et Transport (ERL CNRS 7002 ROOT) et équipe Bases de Données et Traitement des Langues Naturelles (BDTLN).

Sujet

L'ERL CNRS Recherche Opérationnelle, Ordonnancement et Transport (ROOT, *cf.* https://lifat.univ-tours.fr/teams/root/) et l'équipe Bases de Données et Traitement des Langues Naturelles (BDTLN) proposent un financement de thèse de doctorat institutionnelle à temps plein pour un début première quinzaine d'octobre 2020. La thèse sera basée à 50% sur Tours et à 50% sur Blois.

L'équipe ROOT est spécialisée dans les domaines de l'ordonnancement et du transport pour lesquels les outils de la Recherche Opérationnelle sont utilisés. L'équipe BDTLN est spécialisée dans les domaines des bases de données et notamment l'analyse interactive de données.

L'analyse interactive de données est un processus itératif consistant à effectuer une action (par exemple une requête sur des données), recevoir le résultat et décider de l'action suivante à effectuer. L'automatisation de cette tâche rencontre un certain nombre de verrous : comment déterminer parmi la multitude de données le chemin d'analyse à suivre, comment enchainer au mieux les différents types d'actions (requêtes, calcul de modèles, etc.) comment déterminer qu'un résultat est intéressant pour un objectif d'analyse donné, comment raconter, sous forme de narration de données (data storytelling) le résultat d'une analyse, etc.

Le problème qui nous intéresse dans le cadre de cette thèse, est de déterminer un ensemble de requêtes à exécuter en séquence de sorte à maximiser l'intérêt du résultat de ces requêtes par rapport au besoin initial de l'utilisateur. Il est également nécessaire de prendre en compte la durée d'exécution de l'ensemble de ces requêtes de sorte à ce que l'obtention des résultats soit fait dans un temps raisonnable pour l'utilisateur. La problématique soulevée ainsi dans le domaine des bases de données fait ressortir un problème d'optimisation pour lequel les outils de la Recherche Opérationnelle sont pertinents. Une analyse préliminaire fait ressortir une première modélisation de ce problème d'optimisation sous la forme d'un problème de voyageur de commerce (PVC) avec des contraintes particulières :

- les villes du PVC sont les requêtes d'analyse,
- les distances entre villes correspondent au coût cognitif de passer d'une requête à l'autre dans la construction de la narration. Le coût cognitif total (donc la distance totale entre ville) doit être minimisé,
- contrairement au PVC classique :
  - il est ici possible de ne pas visiter toutes les villes. Il faudra donc envisager de rejeter des villes (requêtes), faisant ainsi ressortir une problématique de type sac à dos (knapsack). Chaque ville étant dotée d'une valeur numérique représentant le gain espéré vis-à-vis de la tâche d'analyse à réaliser, il faudra donc sélectionner les villes

maximisant le gain total,

- chaque ville aura également une durée de visite qui représente la durée d'exécution de la requête. La somme des durées de visite ne doit pas dépasser un budget imparti.

Ce problème d'optimisation est NP-difficile et n'a pas fait l'objet d'études dans la littérature consacrée. Notons que d'autres modélisation pourront être proposées, par exemple, en prenant en compte une contrainte globale sur la diversité des requêtes sélectionnées.

L'objectif de cette thèse sera donc d'étudier et modéliser finement le problème d'optimisation posé, proposer des algorithmes exacts et heuristiques issus de la Recherche Opérationnelle (RO), en les évaluant dans le contexte de l'automatisation d'analyse interactive de données. Nous pourrons envisager, selon le profil du candidat, d'utiliser des techniques de Machine Learning appropriées à l'exploration de données, couplées aux algorithmes d'optimisation issus de la RO.

Le candidat recruté devra avoir de solides connaissances théoriques et pratiques en bases de données, particulièrement sur l'expression et l'optimisation de requêtes. Il devra également maîtriser les outils de la Recherche Opérationnelle (complexité, méthodes exactes et heuristiques, programmation mathématique). Des connaissances en machine learning seront un plus.

Merci aux candidats d'envoyer CV détaillé et lettre de motivation aux trois adresses emails indiquées ci-dessus, avant le 10 septembre 2020.