Requêtes discriminantes pour l'exploration des données

Julien Cumin*, Jean-Marc Petit*, Fabien Rouge*,
Vasile-Marian Scuturici*, Christian Surace**, Sabina Surdu*

*INSA Lyon, LIRIS (UMR 5205 CNRS) 69621 Villeurbanne
prenom.nom@insa-lyon.fr
**LAM, CNRS, Marseille
prenom.nom@lam.fr

Résumé. À l'ère du Big Data, les profils d'utilisateurs deviennent de plus en plus diversifiés et les données de plus en plus complexes, rendant souvent très difficile l'exploration des données. Dans cet article, nous proposons une technique de réécriture de requêtes pour aider les analystes à formuler leurs interrogations, pour explorer rapidement et intuitivement les données. Nous introduisons les requêtes discriminantes, une restriction syntaxique de SQL, avec une condition de sélection qui dissocie des exemples positifs et négatifs. Nous construisons un ensemble de données d'apprentissage dont les exemples positifs correspondent aux résultats souhaités par l'analyste, et les exemples négatifs à ceux qu'il ne veut pas. En utilisant des techniques d'apprentissage automatique, la requête initiale est reformulée en une nouvelle requête, qui amorce un processus itératif d'exploration des données. Nous avons implémenté cette idée dans un prototype (iSQL) et nous avons mené des expérimentations dans le domaine de l'astrophysique.

1 Introduction

L'un des aspects encore relativement peu étudié du Big Data porte sur l'exploration interactive des données. Nous nous plaçons dans un cadre de données scientifiques relevant du Big Data et accessibles en SQL, comme c'est communément le cas dans de nombreux domaines, de l'astrophysique à la biologie. Un analyste exploitant ce type de données va passer un temps important à construire la requête de sélection (via SQL) des données qui ont un intérêt pour lui. Nandi et Jagadish (2011) précise que le temps passé pour construire une requête SQL est beaucoup plus important que le temps d'exécution de cette même requête. En partant de ce constat, un nouveau paradigme d'interaction a vu le jour, utilisant les données existantes dans la base de données pour guider le processus de construction de la requête. L'analyste va raffiner successivement la requête en fonction des résultats obtenus pour arriver à une solution convenable.

Sur les données scientifiques plus précisément, il est commun d'avoir des relations avec plusieurs centaines d'attributs, la plupart avec des valeurs numériques issues de mesures physiques. Les critères de sélection des données ne sont pas toujours facilement exprimables avec les attributs existants dans la table. Par exemple, un astrophysicien peut vouloir sélectionner