Fouille de Données Relationnelles dans les SGBD

Cédric Udréa, Fadila Bentayeb

ERIC – Université Lumière Lyon 2
5 avenue Pierre Mendès-France – 69676 Bron Cedex – France
{cudrea,bentayeb}@eric.univ-lyon2.fr

Les travaux sur la fouille de données relationnelles prennent leur essor dans le domaine de la Programmation Logique Inductive (PLI). Bien qu'efficace en terme d'extraction de connaissances, la PLI est inadaptée pour traiter des bases de données relationnelles de grande taille. Dans cet article nous présentons une nouvelle approche qui apporte une solution efficace à la fouille de données relationnelles en intégrant les algorithmes de fouille, en particulier les algorithmes de construction d'arbres de décision, au sein des Systèmes de Gestion de Bases de Données (SGBD).

Notre approche permet d'effectuer les algorithmes de fouille sur des données provenant de plusieurs tables relationnelles sans limitation de taille en utilisant uniquement les outils offerts par les SGBD, en particulier les index bitmap de jointures. Ces derniers permettent d'une part, d'optimiser les temps de traitement et d'autre part, d'exploiter le caractère prédictif porté par la structure de la base de données.

Notre approche consiste à déterminer les effectifs des différentes populations grâce aux index bitmap de jointure qui constituent alors la base d'apprentissage. Les différents effectifs sont obtenus facilement par application des opérations logiques et des opérations de comptage sur les bitmaps (tableaux de bits) sans accéder aux données sources, réduisant les temps de traitement. D'autre part, les index bitmap de jointure apportent une solution au problème des données manquantes engendré par des jointures sur des tables liées par des relations de type 0–N. Nous considérons ces valeurs manquantes comme la négation des autres valeurs possibles. Notre solution consiste à ajouter un index bitmap de jointure artificiel possédant deux bitmaps, l'un correspondant à l'union des différentes valeurs de l'attribut de jointure, l'autre à la négation de cette union. Pour les n–uplets ayant une valeur manquante, leurs bits sont mis à 0 pour le bitmap correspondant à l'union des valeurs et à 1 pour le bitmap correspondant à la négation de l'union. L'index ainsi obtenu permet de différencier les n–uplets ayant une correspondance avec une table de ceux n'en ayant pas. Or cette information (appartenance ou non à une table) peut s'avérer prédictive dans le processu de fouille.

Afin de valider notre approche, nous avons implémenté l'algorithme ID3 (Induction Decision Tree) sous le SGBD Oracle 9i, sous la forme de packages de procédures stockées PL/SQL [1]. Les tests effectués sur des bases possédant des relations 0–N ont montré que notre méthode permet de considérer l'appartenance ou non à une table comme un élément prédictif. De plus, nous obtenons des temps de traitement acceptables.

Ce travail de recherche ouvre de nombreuses perspectives. Il est intéressant d'étudier les performances de notre approche sur des grandes bases de données réelles. Par ailleurs, l'exploitation du caractère prédictif des dépendances fonctionnelles et des contraintes d'intégrité dans le processus de fouille constitue aussi une voie de recherche prometteuse.

1. http://bdd.univ-lyon2.fr/download/relational_tree.zip