Réduction du coût d'évaluation d'une règle relationnelle

Agnès Braud*, Teddy Turmeaux**

*CENTRIA, FCT/UNL, 2829-516 Caparica, Portugal
braud@fct.unl.pt
**LIFO, Université d'Orléans, rue Léonard de Vinci, BP 6759,
F-45067 Orléans Cedex 2, France
Teddy.Turmeaux@lifo.univ-orleans.fr

Résumé. De nombreuses tâches en Fouille de Données visent à extraire des connaissances exprimées sous la forme d'un ensemble de règles. Les algorithmes dédiés à ces tâches engendrent des règles dont l'adéquation aux données doit être évaluée. On se place dans le cadre où cette évaluation est réalisée directement en lançant des requêtes de dénombrement sur la base de données, et où cette base est relationnelle. Les requêtes comptent les données qui s'apparient avec la règle, calcul qui peut être extrêmement coûteux. Dans cet article, nous étudions l'impact d'une approche d'échantillonnage visant à réduire le coût de l'évaluation des règles relationnelles en tenant compte des spécificités structurelles des requêtes induites.

1 Introduction

Nous nous intéressons à l'apprentissage/extraction de règles basées sur un formalisme du premier ordre, dont l'utilisation en Apprentissage est étudiée dans le cadre de la Programmation Logique Inductive (PLI) [Muggleton et Raedt, 1994]. Dans ce contexte, un des problèmes principaux provient du coût du test d'appariement d'une clause avec les données qui permet de mesurer l'accord aux données. En Fouille de Données (FD), le premier ordre permet de traiter les bases de données relationnelles sans les aplatir et donc sans perdre d'informations sur la structure. Cependant, le problème de coût est alors encore plus critique étant donné les volumes de données à traiter, et les systèmes de PLI rencontrent des difficultés pour passer à l'échelle.

La plupart des systèmes de PLI procèdent à une reformulation des données sous la forme utilisée habituellement par le système (en général des clauses Prolog). Une fois les données reformulées dans le formalisme adéquat, le système peut être appliqué sans modification. Nous voyons cependant plusieurs avantages au fait de laisser la charge de l'évaluation au SGBD. Premièrement, il n'est pas nécessaire de passer par une représentation intermédiaire pour traiter les données. Deuxièmement, les SGBD sont sans cesse améliorés et disposent de techniques pour accélérer les accès aux données (index, plans d'exécution par exemple). Ensuite, les SGBD restent la seule solution viable pour traiter des volumes de données importants. Enfin, cela permet d'envisager l'intégration de modules de FD au sein même des SGBD existants.

Nous nous plaçons donc dans ce cadre et nous étudions une méthode visant à accélérer l'éxecution par un SGBD des requêtes permettant d'évaluer les règles relationnelles engendrées par les systèmes. Différentes optimisations ont déjà été proposées, à la fois dans le domaine