Découverte des dépendances fonctionnelles conditionnelles fréquentes

Thierno Diallo* et Noël Novelli**

*Université de Lyon, LIRIS, CNRS-UMR5205
7 av. Jean Capelle, 69621 Villeurbanne Cedex, France
thierno.diallo@insa-lyon.fr
**Université de la Méditerranée, LIF, CNRS-UMR6166
Fac. des Sc. de Luminy, 163 av. de Luminy, 13288 Marseille Cedex 9, France
noel.novelli@lif.univ-mrs.fr

Résumé. Les Dépendances Fonctionnelles Conditionnelles (DFC) ont été introduites en 2007 pour le nettoyage des données. Elles peuvent être considérées comme une unification de Dépendances Fonctionnelles (DF) classiques et de Règles d'Association (RA) puisqu'elles permettent de spécifier des dépendances mixant des attributs et des couples de la forme attribut/valeur.
Dans cet article, nous traitons le problème de la découverte des DFC, i.e. déterminer une couverture de l'ensemble des DFC satisfaites par une relation r. Nous montrons comment une technique connue pour la découverte des DF (exactes et approximatives) peut être étendue aux DFC. Cette technique a été implémentée et des expériences ont été menées pour montrer la faisabilité et le passage à l'échelle de notre proposition.

Mots clés: Dépendances entre données, Fouille de données, Théorie des bases de données.

1 Introduction

Les précédents travaux sur l'amélioration de la qualité des données s'appuient principalement sur les classes de contraintes traditionnelles telles que les DF ou encore les DF Approximatives (Kivinen et Mannila (1995)). Même si les mesures g_1, g_2, et g_3 pour les DF Approximatives capturent certaines erreurs, leur expression n'est pas assez forte ou pas assez fine pour capturer les données incohérentes de manière précise. En effet, ces mesures permettent de détecter des erreurs sur les DF et non au niveau des classes de valeurs. Récemment, Bohannon et al. (2007) ont étendu les DF aux DFC pour pallier ce problème.

Dans cet article nous traitons le problème de l'inférence des DFC i.e. trouver une couverture de l'ensemble des DFC satisfaites dans une relation. Disposer de ces techniques de découverte permet entre autres d'améliorer la performance des outils de nettoyage de données basées sur les DFC. A notre connaissance deux contributions ont été faites sur la fouille de DFC (Chiang et Miller (2008); Fan et al. (2009)). Chiang et Miller (2008) présentent un outil