

Extraction de Dépendances Fonctionnelles Approximatives: une Approche Incrémentale

Ekaterina Simonenko* et Noël Novelli**

* LRI-CNRS UMR 8623

Université Paris-Sud XI; F-91405 Orsay Cedex

ekaterina.simonenko@lri.fr

** LIF-CNRS UMR 6166 – Case 901

Université d’Aix-Marseille; Faculté des Sciences de Luminy;

F-13288 Marseille Cedex 9

noel.novelli@lif.univ-mrs.fr

Résumé. La découverte de dépendances fonctionnelles (DF) à partir d’une relation existante est une technique importante pour l’analyse de Bases de Données. L’ensemble des DF exactes ou approximatives extraites par les algorithmes existants est valide tant que la relation n’est pas modifiée. Ceci est insuffisant pour des situations réelles où les relations sont constamment mises à jour.

Nous proposons une approche incrémentale qui maintiens à jour l’ensemble des DF valides, exactes ou approximatives selon une erreur donnée, quand des tuples sont insérés et supprimés. Les résultats expérimentaux indiquent que lors de l’extraction de DF à partir d’une relation continuellement modifiée, les algorithmes existants sont sensiblement dépassés par notre stratégie incrémentale.

1 Contexte

Les Dépendances Fonctionnelles (DF) représentent les contraintes d’intégrité les plus courantes et les plus importantes en Bases de Données (Mannila et Räihä (1994)). Une DF entre 2 attributs (X, Y) notée $X \rightarrow Y$ est vraie dans une relation si les valeurs de Y sont totalement déterminées par les valeurs de X (Codd (1970)). Le problème de l’extraction de DF est le suivant : “Étant donnée une relation r , trouver toutes les DF qui sont valides dans r ”. Les Dépendances Fonctionnelles Approximatives (DFA) généralisent les DF et sont définies comme “les DF qui sont presque valides dans r , i.e. quelques tuples doivent être retirés de la relation r pour que la DF $X \rightarrow Y$ soit vraie dans r ” (Kivinen et Mannila (1995)). Des DFA apparaissent dans les relations s’il existe une dépendance naturelle entre les attributs mais certains tuples contiennent des erreurs ou représentent une exception. Comme des erreurs peuvent être présentes dans les BD, les DF approximatives sont particulièrement intéressantes.

Récemment, la taille des bases de données a augmenté significativement voire de façon infinie pour les flux de données, rendant les algorithmes existants inefficaces. Ces approches ne peuvent considérer que des relations figées. Quand un tuple est ajouté ou supprimé, l’ensemble des DF valides doit être recalculé.

Les algorithmes les plus efficaces pour l’inférence de DF sont TANE (Huhtala et al. (1998)),