Fouille de données dans les bases relationnelles pour l'acquisition d'ontologies riches en hiérarchies de classes

Farid Cerbah*

*Dassault Aviation
Département des études scientifiques
78, quai Marcel Dassault 92552 Saint-Cloud Cedex
farid.cerbah@dassault-aviation.fr

Résumé. De par leur caractère structuré, les bases de données relationnelles sont des sources précieuses pour la construction automatisée d'ontologies. Cependant, une limite persistante des approches existantes est la production d'ontologies de structure calquée sur celles des schémas relationnels sources. Dans cet article, nous décrivons la méthode RTAXON dont la particularité est d'identifier des motifs de catégorisation dans les données afin de produire des ontologies plus structurées, riches en hiérarchies. La méthode formalisée combine analyse classique du schéma relationnel et fouille des données pour l'identification de structures hiérarchiques.

1 Introduction

Dans les entreprises qui ont à produire et à gérer des données techniques très spécialisées pour la définition de produits complexes, comme dans les secteurs de l'aéronautique et de l'automobile, les entrepôts de données reposent pour une large part sur des bases de données relationnelles. Du fait de leur caractère structuré, ces entrepôts sont des sources à privilégier dans les processus de construction d'ontologies. Cependant, entreprendre un travail d'acquisition d'ontologies à partir de telles sources de données sans disposer d'une aide logicielle adaptée peut s'avérer très vite rédhibitoire.

La thématique d'acquisition d'ontologies à partir de bases de données relationnelles n'est pas nouvelle. Plusieurs méthodes et outils ont été développés pour tirer parti de ces données structurées, avec souvent pour objectif d'assurer l'intégration de bases de données hétérogènes. Cependant, on constate qu'une limite persistante des méthodes proposées est la dérivation d'ontologies de structure calquée sur les schémas des bases de données sources. Ces résultats peuvent difficilement convaincre des utilisateurs attirés par le pouvoir d'expression des formalismes du web sémantique et qui ne peuvent se satisfaire « d'entrepôts sémantiques » ressemblant fortement à leurs bases de données relationnelles. Une attente légitime est d'obtenir en retour des modèles qui rendent mieux compte de la structure conceptuelle sous-jacente aux données stockées.

La dérivation d'ontologies faiblement structurées est le propre des méthodes qui se contentent d'exploiter les méta-données définies dans les schémas sans examiner les données. Une analyse même sommaire de bases de données existantes montre que des motifs de catégorisation