Système adaptatif d'aide à la génération de requêtes de médiation

Dimitre Kostadinov – Verónika Peralta – Assia Soukane – Xiaohui Xue

Laboratoire PRiSM, Université de Versailles
45 avenue des Etats-Unis
78035 Versailles Cedex
France
{prenom.nom}@prism.uvsq.fr

RÉSUMÉ. *Les systèmes de médiation sont aujourd'hui très développés et connus. Cependant, leur mise en oeuvre pose un certain nombre de problèmes, en particulier la génération de requêtes en fonction du contenu des sources et des besoins des utilisateurs. Ce problème est d'autant plus crucial lorsque les sources sont nombreuses et hétérogènes. Nous proposons un outil qui permet de générer automatiquement les requêtes de médiation dans un contexte relationnel et XML et d'adapter ces requêtes aux besoins des utilisateurs en termes de qualité.*

ABSTRACT. *Nowadays, mediation systems are well-known and there exists a great number of implementations. However, their implementation poses several problems, specially, the query generation according to source contents and user's needs. Furthermore, the problem is particularly important when there is a high number of heterogeneous sources. We propose a tool to automatically generate the mediation queries, both in a relational and in a XML context, and to adapt the queries to user quality needs.*

MOTS-CLÉS : *Système de médiation, génération de requêtes, adaptabilité, qualité des données, hétérogénéité de sources.*

KEYWORDS: *Mediation systems, query generation, user adaptability, data quality, source heterogeneity.*

1. Introduction

De nos jours, les systèmes de médiation sont de plus en plus développés et connus. Leurs composants essentiels sont : le schéma global, les mappings du schéma global avec les sources, les fonctions de réécriture de requêtes et les fonctions de composition des résultats. Tous ces composants prennent en compte l'hétérogénéité qui est un des principaux problèmes pour lesquels les systèmes de médiation sont construits. D'autres problèmes de conception émergent lors de l'utilisation de ces médiateurs. Parmi ces problèmes on distingue la définition du schéma global et la définition des mappings qui relient ce schéma global aux sources de données. En raison d'un grand nombre de sources de données, contenant éventuellement des informations redondantes et de qualité variée, il est également important d'adresser le problème d'adaptabilité du système de médiation aux besoins des utilisateurs, notamment en terme de qualité des données.

Les principales questions que l'on se pose sont : (1) Comment automatiser la génération des requêtes de médiation ? (2) Comment détecter et résoudre les problèmes liés à l'hétérogénéité ? (3) Comment évaluer la qualité du système de médiation ? (4) Comment donner la possibilité aux utilisateurs d'exprimer leurs préférences ? (5) Comment tenir compte des préférences de l'utilisateur dans la conception du système de médiation ?

En réponse à ces problèmes, nous proposons un système adaptatif d'aide à la génération de requêtes de médiation. Il a pour objectif d'une part de générer automatiquement des requêtes de médiation en tenant compte de l'hétérogénéité des données et d'autre part d'adapter les requêtes aux besoins des utilisateurs en termes de qualité. Notre démonstration présentera un prototype qui permet de générer des requêtes dans un contexte relationnel et XML, d'évaluer la qualité des données retournées aux utilisateurs et d'exprimer leurs préférences sous la forme de profils.

2. Génération des requêtes de médiation

Il est difficile d'envisager une approche manuelle pour la définition des requêtes de médiation en raison du grand nombre de sources qui peuvent être impliquées et du volume des méta-données les décrivant (description des schémas des sources et du schéma global, assertions des correspondances sémantiques etc.).

Le processus de génération de requêtes doit tenir compte de l'hétérogénéité des données sources. Les opérations qui composent une requête de médiation définie sur les sources ne sont valides que si les conflits sémantiques liés aux instances sont détectés et résolus. Par exemple : Une jointure de deux relations sources sur l'attribut prix peut retourner un résultat incorrect quand les prix sont exprimés dans des monnaies différentes (ex. euro et francs). La transformation des euro en francs ou inversement est une solution au problème.

Nous proposons un système qui permet de générer automatiquement des requêtes SQL dans le contexte relationnel et des requêtes XQuery dans le contexte XML. Il tient compte aussi de l'hétérogénéité des données. Les principales étapes de notre approche (Kedad et al., 1999) (Bouzeghoub et al., 2002) (Collet et al., 2004) sont : (1) Sélection d'un ensemble de sources pertinent pour le calcul du schéma global ; (2) Recherche des opérations candidates qui combinent cet ensemble pertinent, en fonction des assertions entre les schémas des sources et le schéma global, et des clés ; (3) Recherche des transformations dans une librairie de fonctions de transformation pour résoudre les problèmes liés à l'hétérogénéité ; (4) Génération de requêtes de médiation, intégrant des fonctions de transformation, à partir de l'ensemble pertinent et des opérations candidates.

Du point de vue fonctionnel, notre outil intègre un module de génération de requêtes qui permet de sélectionner un ensemble de sources pertinent pour le calcul du schéma global, détecter et résoudre les conflits sémantiques, identifier les opérations candidates qui combinent cet ensemble pertinent, générer les requêtes de médiation SQL dans le contexte relationnel intégrant des fonctions de transformations, et générer les requêtes de médiation XQuery dans le contexte XML.

3. Adaptabilité des requêtes aux besoins des utilisateurs

L'adaptation de l'information délivrée aux utilisateurs joue un rôle fondamental dans la conception et l'exploitation des applications de médiation. Notre outil permet d'une part d'exprimer les préférences des utilisateurs et d'autre part d'évaluer la qualité des donnés afin de délivrer des résultats adaptés à leurs préférences.

3.1. *Evaluation de la qualité*

La qualité des résultats dépend principalement de la qualité des données sources (cohérence, complétude, fraîcheur, etc.) et des propriétés des requêtes qui combinent ces données (coûts, retards, contraints, etc.). Dans notre approche, l'évaluation de la qualité se fait en exécutant des algorithmes d'évaluation, chacun spécialisé dans le calcul d'un facteur de qualité (temps de réponse, fraîcheur, etc.). Les algorithmes prennent en entrée les requêtes de médiation, les valeurs associées aux propriétés des requêtes et les valeurs de qualité des données sources, combinent ces valeurs, et génèrent en sortie des valeurs qui expriment la qualité des résultats des requêtes.

Parmi les différents facteurs de qualité, nous avons choisi la fraîcheur des données pour faire une première étude de notre approche (Bouzeghoub et al., 2004). Nous avons implémenté des algorithmes pour son évaluation selon différents scénarios dans un module d'évaluation de la qualité, intégré dans notre outil.

Du point de vue fonctionnel, le module d'évaluation de la qualité permet de choisir les propriétés les plus pertinentes pour une application donnée ; associer des propriétés aux requêtes de médiation ; incorporer dynamiquement de nouveaux algorithmes d'évaluation ; exécuter en parallèle différents algorithmes d'évaluation ; et décider si les valeurs attendues par les utilisateurs peuvent être satisfaites.

3.2. *Gestion des Profils de l'utilisateur*

La personnalisation de l'information s'exprime par un ensemble de critères et de préférences spécifiques à chaque utilisateur ou une communauté d'utilisateurs. Les données décrivant les préférences des utilisateurs sont souvent sauvegardées sous forme de profils. Le profil d'un utilisateur est composé d'un ensemble de catégories (dimensions) de préférences, par exemple, l'identité de l'utilisateur (nom, age, genre, etc), le domaine d'intérêt (mots clés ou requêtes), la qualité (facteurs de qualité), etc.

Nous proposons un méta-modèle de profil générique et extensible, qui regroupe un grand nombre de préférences proposées dans les approches existantes (Kostadinov, 2004). Un utilisateur donné peut ne pas avoir besoin de toutes les informations contenues dans le méta-modèle pour construire son profil. Par exemple un utilisateur peut s'intéresser à la qualité des données et être indifférent à la sécurité. Dans notre approche, l'utilisateur a la possibilité de choisir les composants de son profil à partir du méta-modèle ou de créer sa propre structure pour ensuite entrer les valeurs attendues des paramètres de personnalisation.

Notre outil intègre un gestionnaire de profils qui permet de construire le profil d'un utilisateur. La construction se fait en deux étapes : (1) choix de la structure du profil (catégories et attributs qui sont pertinents pour l'utilisateur) et (2) attribution des valeurs aux attributs.

4. Bibliographie

Bouzeghoub, M ., Kedad, Z., Soukane, A., « Génération de requêtes de médiation intégrant le nettoyage de données », *Revue du Ingénierie des Systèmes d'Information ISI'02*, 2002.

Bouzeghoub, M., Peralta, V., « A Framework for Analysis of Data Freshness », *Proc. of the Int. Workshop on Information Quality in Information Systems IQIS'2004*, 2004.

Collet C., Belhajjame K., Bernot G., Bruno G., Bobineau C., Finance B., Jouanot F., Kedad Z., Laurent D., Vargas-Solar G., Tahi F., Vu T. T., Xue X., « Towards a mediation system framework for transparent access to largely distributed sources », *Proc. of the Int. Conf. on Semantics of a Networked World Semantics for Grid Databases IC-SNW'2004*, 2004.

Kedad, Z, Bouzeghoub, M., « Discovering View Expressions from a Multi-Source Information System », *Proc. of the 4th. Int. Conf. on Cooperative Information Systems CoopIS'1999*, 1999.

Kostadinov, D., « Personnalisation de l'information et gestion des profiles utilisateurs », Rapport DEA. Université de Versailles, France, 2003.