Suppression d'une Source de Données dans un Système de Médiation: Cas d'absence d'une Source Equivalente

Latifa Baba-Hamed[1] and Farah Sedjelmaci[1],

[1] Département d'Informatique, Université d'Oran Es-sénia. B.P. 1524,
El M'Naouer, 31000 Oran, Algérie
{lbabahamed, fsedjelmaci}@yahoo.fr

Résumé. L'intégration de l'information fournie par de multiples sources de données hétérogènes est de plus en plus importante dans les systèmes d'information modernes. Dans ce contexte, les besoins des applications sont décrits par un schéma cible et la façon dont les instances du schéma cible sont dérivées à partir des sources de données est exprimée par des mappings. L'un des problèmes qui mérite d'être considéré est l'impact de l'évolution de schéma sur les mappings. Dans ce papier, nous nous intéressons à la suppression d'une source dans un système de médiation, et montrons comment mettre à jour les mappings affectés par cette opération dans le contexte de l'approche GAV (Global-as-view). Une source peut être supprimée parce qu'elle fournit toujours des informations obsolètes ou parce qu'elle est indisponible. Le modèle choisi, pour représenter les schémas sources ainsi que le schéma global est le modèle relationnel.

Mots-clés: Système de médiation, requêtes de médiation, source contributive, relation pertinente, graphe d'opération.

1 Introduction

De nos jours, les systèmes multi-sources sont de plus en plus développés. Ils sont définis comme l'intégration de plusieurs sources hétérogènes et distribuées. Les systèmes d'intégration consistent à fournir une vue uniforme des sources de données (appelée schéma global) et à spécifier un ensemble de requêtes appelées requêtes de médiation ou mappings opérationnels.

Plusieurs travaux concernant l'intégration de données ont été développés. Nous pouvons citer les travaux concernant le nettoyage des données (un état de l'art sur le nettoyage de données est donné dans [23]). Quelques approches ont été proposées pour générer les mappings d'une façon automatique ou semi-automatique [5, 16, 7, 29, 21, 13 et 10]. Quelques autres approches se sont occupées de l'évolution de schéma ainsi que de l'adaptation automatique des mappings [4, 12, 15, 25, 28 et 27]. Certains travaux ont considéré la qualité de données [20, 1] ou la qualité des schémas [18]. Enfin, d'autres recherches se sont intéressées à la définition des correspondances sémantiques entre deux schémas (aussi appelé matching de schéma) [9, 22].

Parmi ces systèmes d'intégrations, nous distinguons les entrepôts de données, les systèmes d'informations basés sur le web, ou encore les systèmes de médiation. Un système de médiation est un système qui permet d'interopérer sur un ensemble de sources hétérogènes et distribuées. Ses composants essentiels sont : le schéma global (appelé schéma de médiation), les mappings du schéma global avec les sources et les fonctions de transformation concernant l'hétérogénéité des données. Les mappings du schéma global avec les sources sont des requêtes, appelées requêtes de médiation.

La définition du schéma global, qui offre une vue uniforme des sources varie selon deux approches : une approche ascendante (Global As View ou GAV) où chaque objet du schéma est défini par une requête sur les sources (c'est l'approche utilisée dans TSIMMIS [6]), et une approche descendante (Local As View ou LAV) où chaque objet d'une source de données est défini par une requête sur le schéma global (c'est l'approche utilisée dans Information Manifold [11]).

Ce papier considère le problème de l'évolution de schéma dans le contexte de l'approche GAV. Il étudie plus précisément, la suppression d'une source de données dans un système de médiation suivant la méthodologie présentée dans [3] et améliorée dans [5] pour prendre en compte l'hétérogénéité des données. Il montre également, comment mettre à jour les mappings affectés par cette opération de suppression. Une source peut être supprimée parce qu'elle fournit toujours des informations obsolètes ou parce qu'elle est indisponible.

Ce papier est organisé comme suit : la section 2 est consacrée aux approches d'évolution de schéma. La section 3 présente brièvement, la méthodologie utilisée pour la génération des requêtes de médiation. Enfin, la section 4 montre l'impact de l'opération de suppression d'une source de données sur les requêtes de médiation, propose un algorithme qui propage ce changement au niveau médiation en suivant cette méthodologie, et décrit les métadonnées utilisées pour exécuter cette opération.

2 Méthodes d'Evolution de Schéma

L'évolution de schéma est un domaine de recherche assez vaste qui inclut les problèmes qui touchent aux changements opérés sur le schéma. Il a été étudié dans différents contextes et sous différentes conditions.

Dans les SGBD orientés objet, Banerjee et al [2] a donné une taxonomie des opérations de changement qui peuvent être détectées et a fourni une implémentation à chacune d'entre elles.

La maintenance incrémentale de vue [19] concerne les méthodes qui mettent à jour, efficacement, les vues matérialisées quand le schéma de la base est mis à jour. L'adaptation de vues [8, 17] est une variante de maintenance de vues qui utilise des méthodes pour garder les données des vues matérialisées mises à jour en réponse à des changements dans la définition de la vue elle-même.

Dans les systèmes d'intégration de données, plusieurs solutions ont été proposes pour l'adaptation automatique des mappings, nous les présentons dans ce qui suit.

Dans AutoMed [15], évolution de schéma et intégration sont combinées dans une plate-forme unique. Les schémas des sources sont intégrés dans un schéma global en leur appliquant une séquence de transformations primitives. Le même ensemble de

transformations primitives peut être utilisé pour spécifier l'évolution d'un schéma source vers un nouveau schéma. Les auteurs montrent comment adapter les mappings existant entre le schéma global et chacun des schémas source quand les schémas des sources évoluent. Cette approche repose sur le modèle de données HDM (Hypergraph Data Model). Le modèle HDM est un graphe constitué d'un ensemble d'entités ou de nœuds reliés par des arrêtes, ces dernières peuvent porter des contraintes. Un schéma S dans un modèle HDM est un triplet <N, A, C>, tel que : N est l'ensemble des nœuds, A l'ensemble des arrêtes et C l'ensemble des contraintes. Une requête q sur un schéma est une expression dont les variables appartiennent à l'ensemble $N \cup A$. La transformation d'un schéma source en un schéma global est une succession de transformations primitives élémentaires du genre *delEdge* (supprimer arrête), *addCons* (ajouter contrainte), *renNode* (renommer nœud), etc.

Bouzeghoub et al. [4] a considéré le problème de l'évolution dans le contexte de l'approche GAV. Les auteurs se sont basés sur la méthodologie définie dans [3, 5] pour la génération de requêtes de médiation. Etant donnés une relation de médiation, un ensemble de schémas source et un ensemble d'assertions linguistiques entre le schéma de médiation et les schémas des sources, les auteurs ont défini un algorithme qui découvre les requêtes de médiation définissant cette relation. Le processus d'évolution est une extension de cet algorithme. Leur solution se base sur le concept de relations pertinentes sur lesquelles des règles de propagation ont été définies. Chaque règle d'évolution est une règle E-C-A dans laquelle l'événement représente l'opération de changement et l'action est un ensemble de primitives de propagation à exécuter quand les conditions sont satisfaites. Les auteurs ont limité leur étude à quelques opérations de changement ; ils n'ont pas considéré l'ajout et la suppression d'une source de donnée dans un système de médiation. Loscios et Salgado [14] ont suivi la même démarche que l'approche [4] pour faire évoluer les mappings générés par Loscios [13]. Leur approche utilise le modèle XML pour représenter les schémas des sources et le schéma de médiation.

Xue [27] propose une approche incrémentale pour l'adaptation des mappings. Elle considère à la fois, la génération automatique des mappings et leur adaptation pour des schémas XML. Dans son approche, les mappings peuvent exprimer des jointures inter-sources. Elle ne suppose aucune homogénéité entre le schéma cible et les schémas des sources et génère des mappings dans un langage abstrait qu'elle traduit en XQuery. Elle peut également adapter des mappings exprimées en XQuery quand le schéma cible ou le schéma source évolue.

L'approche EVE (*Evolvable View Environment)* [12] constitue l'un des premiers travaux introduisant les opérations de changement dans les sources de données. Elle concerne le problème d'adaptation de définition de vues dans un environnement dynamique (appelé *problème de synchronisation de vues*). Pour résoudre ce problème, les auteurs proposent un langage de définition de vues étendu appelé *E-SQL*, qui est capable de définir des vues flexibles. Les attributs (A) dans la clause *SELECT*, les relations (R) dans la clause *FROM*, les clauses primitives (C) dans la clause *WHERE* sont les unités de base dans une vue ; elles sont appelées *composants de la vue*. Deux paramètres d'évolution sont attachés à chaque composant de la vue. Le *paramètre dispensable* est utilisé pour dire que le composant de la vue est exigé et, donc, doit être gardé dans la vue modifiée (quand la valeur est fausse). Le paramètre dispensable est noté *XD*, où *X* représente *A*, *R* ou *C*. Le *paramètre*

remplaçable spécifie si le composant de la vue peut être remplacé dans le processus de synchronisation de la vue (quand la valeur est vraie). Il est noté *XR*, où *X* représente *A*, *R* ou *C*. Les auteurs introduisent un modèle de description de sources d'information (MISD) qui permet à une grande classe de sources d'informations de participer dans leur système de façon dynamique, développent également, des stratégies de remplacement pour les composants affectés de la vue, et fournissent un ensemble d'algorithmes de synchronisation de vues basés sur ces stratégies.

Le projet Clio [21] a proposé une approche de génération de mappings entre un schéma source et un schéma cible. Ces schémas sont modélisés en relationnel ou en XML. L'approche présentée dans [25] complète le scénario ci-dessus. Velegrakis et al. prend les mappings générés par l'outil de mappings et les adapte quand les schémas évoluent, de façon à conserver leur cohérence. Les auteurs considèrent les changements non seulement au niveau de la structure des schémas source ou cible (ce qui peut rendre le mapping incorrect syntaxiquement) mais aussi au niveau de la sémantique des schémas (i.e. contraintes de schéma). Ils réalisent les changements non seulement sur des éléments atomiques, mais aussi sur des structures plus complexes incluant des tables relationnelles ou des structures XML imbriquées. Ils présentent, également, un algorithme d'adaptation de mapping qui détecte les mappings affectés par le changement et génère toutes les réécritures adéquates. Pour évaluer l'efficacité de leur approche, ils ont implémenté un prototype appelé *ToMAS*.

Yu et Popa [28] développent un outil pour adapter automatiquement des mappings générés par Clio'02. Considérons trois schémas S_1, S_2 et S_3, un mapping m_{12} entre S_1 et S_2 et un autre mapping m_{23} entre S_2 et S_3. Cette approche consiste à combiner m_{12} et m_{23} de façon à produire les mappings possibles entre S_1 et S_3. La composition des mappings m_{12} et m_{23} se fait en trois étapes : i) créer un ensemble de règles, à partir de m_{12}, pour montrer comment les éléments de S_2 sont exprimés en utilisant des éléments de S_1 ; ii) utiliser ces règles pour modifier m_{23} en transformant toutes les références à S_2 en des références à S_1 donnant comme résultat un ensemble de mappings M_{13} ; iii) vérifier la validité des mappings de M_{13}. Pour réduire le nombre de combinaisons, les auteurs présentent une méthode qui supprime tous les mappings originaux non affectés ainsi que les mappings redondants.

3 Principe de l'Approche de Génération de Mappings Utilisée

Pour étudier la suppression d'une source dans un système de médiation, nous avons choisi la méthode présentée dans [3, 5] pour la génération de requête de médiation dont nous rappelons le principe dans cette section.

Cette approche a été proposée dans le cadre des systèmes de médiation dans lesquels le schéma cible est appelé *schéma de médiation* et les mappings sont appelées *requêtes de médiation*. Elle considère que les schémas de médiation et des sources sont exprimés en relationnel et que les schémas de médiation sont définis par des experts du domaine indépendamment des sources. L'objectif de cette méthode est d'aider les utilisateurs à dériver les instances du schéma de médiation à partir des schémas des sources en générant un ensemble de requêtes de médiation candidates. Des requêtes de médiation sont générées pour chaque relation du schéma de

médiation. L'algorithme de la méthode peut être résumé en 3 étapes: (i) recherche des sources de données contributives; (ii) détermination des opérations candidates; (iii) définition de requêtes de médiation.

La première étape consiste à trouver toutes les relations source qui peuvent contribuer au calcul de la relation de médiation. Une relation source S_i est contributive si elle inclut quelques attributs de la relation de médiation. Dans ce cas, une *relation de mapping* est extraite ; la *relation de mapping* contient tous les attributs communs entre la relation de médiation et S_i. Les clés primaire et étrangère de S_i sont rajoutées à la relation de mapping. Considérons l'exemple suivant dans lequel il y a une seule relation de médiation *Rm(#K,A,B,C)* et quatre relations source *S1(#K,A,@X,Y)*, *S2(#X,B,Z)*, *S3(#B,C,W)* et *S4(#B,C,U)*. Les attributs clé primaire sont préfixés par # et les attributs clé étrangère sont préfixés par @. Dans cet exemple, quatre relations de mapping sont obtenues à partir de *S*, *S2*, *S3* et *S4*: *T1(#K,A,@X)*, *T2(#X,B)*, *T3(#B,C)* et *T4(#B,C)*.

La seconde étape recherche les jointures possibles entre les relations de mapping. L'opération de jointure est candidate dans deux cas : (i) les deux relations de mapping sont originaires d'une même source, dans ce cas nous considérons qu'une jointure est possible, s'il existe une contrainte référentielle explicite entre les deux relations sources ; (ii) les deux relations de mappings sont originaires de deux sources différentes, dans ce cas nous considérons qu'une jointure est possible si la clé primaire d'une relation a un attribut équivalent dans l'autre relation. La figure 1 montre un exemple des opérations possibles pour notre exemple. La jointure 1 est possible entre *T1* et *T2* parce qu'il y a une contrainte référentielle de *T1* à *T2* à travers l'attribut *X*. La jointure 2 est possible entre *T2* et *T3* parce que l'attribut *B* existe dans *T2* et dans *T3* et *B* est défini comme clé dans *T3*.

Dans le cas d'une opération de jointure, il se peut qu'il n'existe aucune contrainte référentielle implicite ou explicite entre deux relations sources contributives. Il serait possible de joindre deux relations sources *Si* et *Sj* à travers une troisième relation *Sk* qui n'est pas directement contributive au calcul de *Rm*. Cet algorithme inclut ces relations comme des *relations de transition* qui permettent la jointure entre les relations de mapping. Par exemple, soient les deux relations de mapping *T5(#D,E)* et *T6(#F,G)*. Il n'y a pas de jointure possible entre elles. Supposons l'existence de la relation source *S7(#F,@D,H)* et ni *F, D* et *H* n'est dans la relation de médiation, alors *S7* peut être utilisée pour joindre *T5* et *T6*: *T5* et *T7* à travers *D* ; *T6* et *T7* à travers *F*. Une relation de transition est générée à partir de *S7*: *T7(#F,@D)*, elle contient les clés primaire et étrangère uniquement. Les relations de mapping et les relations de transition sont appelées *relations pertinentes*. Disposant du graphe d'opérations défini sur les relations pertinentes, il devient facile de générer les requêtes de médiation à partir de chemins de calcul. Un *chemin de calcul* est un sous-graphe acyclique et connexe du graphe d'opérations qui enveloppe tous les attributs d'une relation de médiation. Définir des requêtes de médiation revient à énumérer tous les chemins de calcul du graphe d'opérations. Dans l'exemple de la figure 1, C1 = (1, 3) et C2 = (1, 2) sont deux chemins de calcul. Leurs requêtes de médiation correspondantes sont respectivement :

$E1 = \Pi_{K,A,B,C}[(\Pi_{K,A,X}S1) \bowtie (\Pi_{X,B}S2) \bowtie (\Pi_{B,C}S4)]$;

$E2 = \Pi_{K,A,B,C}[(\Pi_{K,A,X}S1) \bowtie (\Pi_{X,B}S2) \bowtie (\Pi_{B,C}S3)]$.

Les opérations basées sur les ensembles telles que l'union, la différence et l'intersection peuvent être utilisées sur les requêtes de médiations trouvées pour donner de nouvelles requêtes de médiation. Par exemple, E3 = E1U E2.

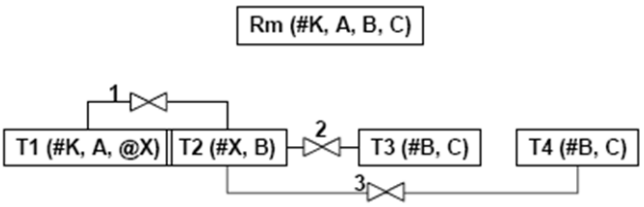

Fig. 1. Exemple de graphe d'opération.

4 Suppression d'une Source dans un Système de Médiation

L'évolution de schéma dans un système de médiation est un domaine de recherche d'actualité. Il s'agit de maintenir la cohérence du schéma global après un certain nombre d'opérations de changements effectués au niveau des sources de données. Ces changements peuvent affecter certaines requêtes de médiation, et par conséquent les réponses aux requêtes des utilisateurs peuvent être erronées. Une propagation, des modifications survenues dans les sources vers le schéma global, s'avère nécessaire si on veut garder la cohérence de notre système. Les changements considérés peuvent concerner l'ajout ou la suppression d'une relation, d'un attribut, d'une contrainte d'intégrité ou d'une source. Pour notre étude, nous nous sommes limitées à la suppression d'une source de données d'un système de médiation hétérogène. Une source de données peut être supprimée d'un système de médiation car elle fournit toujours des informations obsolètes ou parce qu'elle est indisponible. L'hétérogénéité peut être sémantique (l'utilisation d'une terminologie différente pour désigner deux concepts identiques par exemple *prix* et *prix-produit*), ou structurelle (comme par exemple, le format d'écriture d'une donnée ou bien encore son unité de mesure). L'approche choisie est une approche GAV pour définir les objets au niveau global [3], et le modèle choisi, pour représenter les schémas des sources ainsi que le schéma global, est le modèle relationnel.

Quand on retire une source, plusieurs cas peuvent se produire : (i) il existe une autre source équivalente (mais sans doute avec une moins bonne qualité), on régénère une requête de médiation avec cette autre source et on avertit l'utilisateur de la dégradation possible de la qualité ; (ii) il n'existe aucune autre source équivalente, on peut adopter deux attitudes : soit on supprime la relation de médiation qu'on ne peut plus calculer (nous avons adopté cette attitude dans ce papier), soit on génère des résultats partiels (étude en cours).

Dans cette section, nous décrivons d'abord les métadonnées sur lesquelles nous effectuons notre opération de changement, puis nous présentons l'algorithme général de la suppression d'une source.

4.1 Description des Métadonnées

Nous distinguons trois niveaux différents : local, intermédiaire et global. Chaque niveau contient un ensemble de tables définissant les métadonnées utilisées dans notre système.

Niveau Local. Ce niveau explicite cinq tables : *Source* (cette table inclut toutes les sources du système), *Source_relation* (cette table regroupe toutes les relations des sources), *Source_attribut* (cette table inclut tous les attributs de toutes les sources du système), *Attribut_étendu* (en plus de son nom et de son type de base un attribut est décrit par un ensemble de métadonnées représentant son type. Le type étendu d'un attribut A d'une relation R est défini comme un tableau associatif d'éléments à deux colonnes, où la première colonne décrit le nom de l'élément et la deuxième décrit sa valeur. Les éléments retenus dans notre application sont les suivants : *format*, *unité*, *échelle*, *précision* [5]), *Contrainte* (cette table inclut toutes les contraintes référentielles de toutes les sources du système). La table 1 décrit ces différentes tables.

Table 1. Description des relations du niveau local.

Nom de la table	Les attributs de la table	La description des attributs
Source	Id_source nom_source	Identificateur de la source Nom de la source
Source_relation	Id_rel_src Nom_rel_src Id_source	Identificateur de la relation source Nom de la relation source La source à laquelle appartient cette relation
Source_attribut	Id_att_src Nom_att_src Id_rel_src Type	Identificateur de l'attribut source Nom de l'attribut source La relation à laquelle appartient cet attribut Le type de base de cet attribut (Entier, Réel)
Attribut_étendu	Id_src_etend Elt_etend Val_elt_etend Id_att_src	Identificateur interne Le nom de l'élément type La valeur de cet élément type L'identifiant de l'attribut qu'on veut étendre
Contrainte	Id_contrainte Id_att_src1 Id_att_src2	Identifiant de la contrainte référentielle Identificateur du premier attribut Identificateur du deuxième attribut

Niveau Intermédiaire. Ce niveau explicite quatre relations : *Opération* (cette table inclut les informations concernant les graphes d'opérations), *Relation_pertinentes* (cette table regroupe toutes les relations de mappings et les relations de transition), *Correspond_S_S* (cette table inclut toutes les correspondances linguistiques entre les relations de sources différentes), *Correspond_S_M* (cette table inclut toutes les correspondances linguistiques entre les concepts des sources et les concepts du schéma de médiation (synonymie, abréviations, inclusions, et équivalences linguistiques des noms des attributs)). La table 2 décrit ces différentes tables.

Table 2. Description des relations du niveau intermédiaire.

Nom de la table	Les attributs de la table	La description des attributs
Opérations	Id_Op Type Rel1 Rel2 Arc Id_rel_med	Pour identifier chaque opération. Type de l'opération (jointure, union,…). Identifiant de la relation pertinente 1. Identifiant de la relation pertinente 2. Id de l'arc reliant les 2 relations pertinentes. La relation de médiation à laquelle appartient cette opération.
Relation_pertinentes	Id_rel_pert Id_rel_src Id_rel_med Type	Identificateur de la relation pertinente. Identifiant de la relation source à partir de laquelle on a dérivé cette relation pertinente. La relation de médiation pour laquelle on a dérivé cette relation pertinente. « mapping »ou « transition »
Correspond_S_S	Id_correp_s_s Id_att_src1 Id_att_src2	Ident. de la correspondance source_source. Identifiant du premier attribut Identifiant du second attribut
Correspond_S_M	Id_correp_s_m Id_att_src Id_att_med	Ident. de la correspondance source-médiation Identifiant de l'attribut source correspondant Identifiant de l'attribut de médiation

Table 3. Description des relations du niveau global.

Nom de la table	Les attributs de la table	La description des attributs
Relation_médiation	Id_rel_med Nom_rel_med	Identificateur de la relation de médiation Nom de la relation de médiation
Attribut _médiation	Id_att_med Nom_att_med Id_rel_med Type	Identificateur de l'attribut de médiation Nom de l'attribut de médiation La relation de médiation à laquelle appartient cet attribut. Le type de base de cet attribut
Attribut_étendu	Id_med_etend Elt_etend Val_elt_etend Id_att_med	Identificateur interne Le nom de l'élément type La valeur de cet élément type L'identifiant de l'attribut de médiation qu'on veut étendre.

Niveau Global. Ce niveau explicite trois relations : *Relation_médiation* (cette table regroupe toutes les relations du schéma de médiation), *Attribut_médiation* (cette table inclut tous les attributs du schéma de médiation), *Attribut_médiation_étendu* (Cette table répertorie les attributs étendus au niveau du schéma de médiation). La table 3 décrit ces différentes tables.

4.2 Algorithme de Suppression

L'algorithme *Remove-source* montre les modifications effectuées au niveau local et qui doivent être propagées au niveau intermédiaire. La suppression d'une source S_i consiste en la suppression de toutes ses relations source. La suppression d'une relation source S_{ij} conduit à la suppression de toutes ses contraintes et de tous ses attributs. La suppression d'un attribut implique la suppression de toutes les correspondances linguistiques qui lui sont associées. Pour refléter la suppression de la relation locale S_{ij}, la relation pertinente T_{ij} correspondante doit être supprimée du graphe d'opération, ainsi que toutes les opérations enveloppant T_{ij}.

Notre algorithme inclut quelques modules que nous décrivons dans ce qui suit. Le module *Update-corresp-s-s* supprime, de l'ensemble *corresp-s-s,* les correspondances linguistiques entre les deux sources auxquelles l'attribut *B* appartient. Il met à jour la table *Correspond-S-S*. Le rôle du module *Update-corresp-s-m* est de supprimer, de l'ensemble *corresp-s-m*, les correspondances linguistiques entre les concepts des sources et les concepts du schéma de médiation auxquelles l'attribut *B* appartient. Il met à jour la table *Correspond-S-M*. Le module *Update-ref-constraint* supprime, de l'ensemble des contraintes de référence *ref-cons,* les contraintes auxquelles l'attribut *B* appartient. Il nécessite la table *Contrainte*.

Le module *Update-relevant-rel* met à jour l'ensemble des relations pertinentes correspondant à la relation R_m dans le schéma de médiation et met à jour le graphe d'opérations G_{Rm}. Il utilise les tables *Relation_pertinentes* et *Opération*. En utilisant G_{Rm}, le module *Search-computation-path* recherche l'ensemble des chemins de calcul *CP* correspondant à R_m. Il se peut qu'aucun chemin de calcul ne soit trouvé après la propagation (c'est-à-dire $CP=\varnothing$); dans ce cas R_m devient non calculable et sera donc supprimée du schéma de médiation (cette suppression est effectuée par le module *Delete-RM*) puisque nous traitons, dans cette étude, la suppression d'une source en l'absence de sources équivalentes. Nous supposons qu'avant de lancer l'algorithme de suppression d'une source *Remove-source*(S_i, *S*), nous avons exécuté l'algorithme *Equivalence*(S_i, S_k, S_{nc}, *test)* dont le rôle est de tester l'équivalence entre la source (à supprimer) S_i et la source S_k, quelque soit S_k appartenant à l'ensemble des sources non contributives $S_{nc} = S - S_c$, et qu'il nous a retourné *test = faux* (qui veut dire qu'il n'existe pas de source équivalente à S_i).

Le module *Generate-query* génère l'ensemble des requêtes *Q* correspondant à l'ensemble *CP* uniquement quand $CP\neq\varnothing$.

Les suppressions de *B*, S_{ij} et S_i sont effectuées dans les tables *Source-attribut*, *Attribut-étendu*, *Source-relation* et *Source*.

```
Remove-source (S_i, S)
S_i: is the source to be removed
S: the set of the sources
S_c: the set of the contributive sources
S_ij: the schema of the relation source S_ij
 If S_i ∈ S_c then
  For each source relation S_ij
      For each mediation relation R_m
        For each attribute B of S_ij
          Update-ext-typ-src (EXT, B);
          Update-corresp-s-s (corresp-s-s, B);
          Update-corresp-s-m (corresp-s-m, B);
          Update-ref-constraint (ref-cons, B);
           S_ij = S_ij - {B}; // remove B from S_ij
        EndFor
        Update-relevant-rel (S_ij, M, OP);
      EndFor
     S_i = S_i - S_ij; // remove S_ij from the source S_i
  EndFor
  For each affected mediation relation R_m
    Search-computation-path (G_Rm, CP);
      // G_Rm is composed by M and OP
     // CP is the set of computation paths of R_m
    If CP ≠ ∅ then
       Generate-query (CP, Q);
       // Q is the set of queries to compute R_m
    Else
       Delete-RM (R_m, corresp-s-m, Relation-médiation,
        Attribut-médiation, Attribut-médiation-étendu)
    EndIf
  EndFor
  S= S-{S_i}; // remove the source S_i from S
 EndIf
End Remove-source

Update-relevant-rel (S_ij, M, OP)
M: the set of relevant relations in R_m
OP: the set of relational operations in G_Rm
S_ij: a relation source in the source S_i
  If T_ij ∈ M such that T_ij ⊆ S_ij
  Then M=M-{T_ij};
  EndIf
  For each operation op involving T_ij
  OP=OP-{op};
  EndFor
End Update-relevant-rel
```

Exemple. On considère deux relations de médiation R1 et R2, et l'ensemble des sources de données S= {S1, S2, S3, S4, S5, S6}. La table 4 résume les relations des six sources au niveau local et des deux relations de médiation au niveau global. Les figures 2 et 3 présentent les graphes d'opérations G_{R1} et G_{R2} correspondant aux relations de médiation R_1 et R_2 respectivement.

Table 4. Schémas des relations de médiation et des sources de l'exemple.

Niveau	Relations	Schéma de la relation
Médiation	R1 R2	R1 (#K, A, B, C) R2 (#K', D, E, F)
Source 1	S11 S12	S11 (#K, A,@X, R') S12 (#X,B,@Y,T)
Source 2	S21 S22	S21 (#Y, C, @W, U) S22 (#K', D, E, @P)
Source 3	S31 S32	S31 (#X, C, V) S32 (#R,V,O,@W)
Source 4	S41 S42 S43	S41 (#X, C, J) S42 (#Z, C, L) S43 (#K',D,E,L,@P)
Source 5	S51 S52	S51 (#W, Z, F) S52 (#P,N,@R)
Source 6	S61 S62	S61 (#K1,A1,B1,@D1) S62 (#D1, C1, E1)

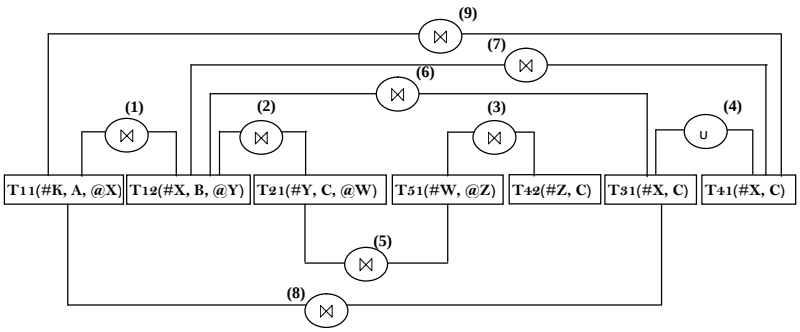

Fig. 2. Le graphe d'opérations G_{R1}.

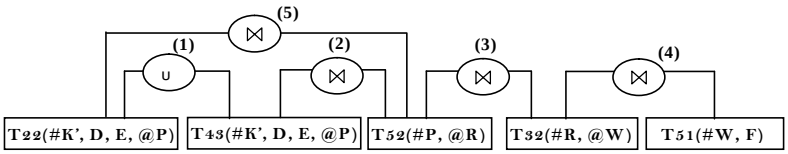

Fig. 3. Le graphe d'opérations G_{R2}.

La suppression de la source S_1 du système considéré, affecte uniquement le graphe d'opération G_{R1}, G_{R2} reste inchangé. Cette modification conduit à la suppression des relations pertinentes T_{11} et T_{12} de G_{R1} et toutes les opérations qui les enveloppent. La figure 4 montre le graphe G_{R1} après l'opération de suppression. Il en résulte qu'aucun chemin de calcul ne peut être trouvé après la propagation, ce qui implique que la relation de médiation R_1 devient non calculable.

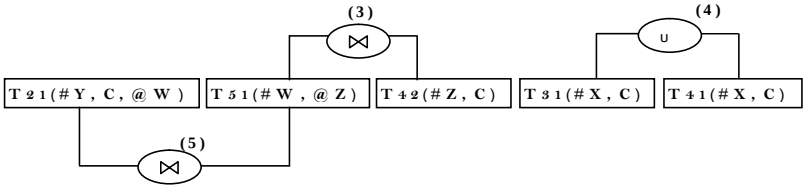

Fig. 4. Le graphe d'opérations G_{R1} après suppression de la source S_1.

5 Conclusion

Dans ce papier, nous avons présenté la suppression d'une source de données dans un système de médiation hétérogène dans le cas d'absence d'une source équivalente. Nous avons suivi une approche GAV à travers laquelle nous avons montré l'impact de cette opération sur le niveau médiation en ne considérant que les requêtes de médiation affectées par cette suppression. Nous avons traité un seul événement de suppression. Le traitement de plusieurs événements de suppression pose un problème de concurrence que nous pouvons résoudre en les sauvegardant dans une file d'attente, et de ne traiter qu'un seul à la fois comme nous l'avons montré dans ce papier. La mesure de la perte d'information ou de qualité après la suppression d'une source peut constituer une perspective à cette étude.

References

1. Akoka J., L. Berti-Equille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué-Thion, Z. Kedad, S. Nugier, V. Peralta, S. Sisaid-Cherfi, "A Framework for quality evaluation in data integration systems", 9th International Conference on Enterprise Information Systems (ICEIS'2007), Funchal, Portugal, June 2007.
2. Banerjee J., Kim W., Kim H., and Korth H., "Semantics and Implementation of Schema Evolution in Object-Oriented Databases," in SIGMOD, pp. 311-322, May 1987.
3. Bouzeghoub M. and Kedad Z., "Discovery View Expressions from a Multi-Source information System," Proceedings of the Fourth IFCIS International Conference on Cooperative Information Systems (COOPIS'99), Edinburgh, Scotland, pp.57-68, 1999.
4. Bouzeghoub M., Farias Lóscio B., Kedad Z., Ana Carolina Salgado A.S., "Managing the Evolution of Mediation Queries," Proc. Of the Int. Conf. on CoopIS'2003, pp. 22-37, 2003.
5. Bouzeghoub M., Kedad Z., Soukane A., "Improving Mediation Query Generation using Constraints and metadata," Bases de Données Avancées (BDA), Montpellier, pp. 385-405, 2004.

6. Chawathe S., Garcia-Molina H., Hammer J., Ireland K., Papakonstantinou Y., Ullman J., and Widom J., "TSIMMIS Project: Integration of Heterogeneous Information Sources," in Proc. of IPSI Conf., Tokyo, Japan, 1994.
7. Fletcher G.H.L. and Wyss C.M., "Data Mapping as Search", EDTB, pp. 95-111, 2006.
8. Gupta A., Mumick I. and Ross K., "Adapting Materialized Views after Redefinition," in SIGMOD, pp.211-222, 1995.
9. He B. and Chen-Chuan Chang K., "Automatic Complex Schema Matching across Web Query Interfaces: A correlation Mining Approach", Proc. of ACM Transactions on Database Systems, 2006.
10. Kedad Z. and Xue X., "Mapping generation for XML data sources: a general framework," WIRI (Web Information Retrieval and Integration), Tokyo, Japan, pp.164-172, 2005.
11. Kirk T., Levy A.Y., Sagiv Y. and Srivastava D., "The Information Manifold," in Proc. of AAAI 95 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, pp. 85-91, 1995.
12. Lee A.J., Nica A., and Rundensteiner E.A., "The EVE Approach: View Synchronization in Dynamic Distributed Environments," IEEE TKDE, vol.14, no. 5, pp. 931-954, 2002.
13. Loscios B.F., "Managing the Evolution of XML-based Mediation Queries", PHD thesis, Universidade Federal de Pernambuco (Brésil), April 2003.
14. Loscios B.F. and Salgado A.C., "Evolution of XML-Based Mediation Queries in a Data Integration System", in Proc. Of ER Workshops, Shanghai, China, pp.402-414, 2004.
15. McBien P. and Poulovassilis A., "Schema Evolution in Heterogeneous Database Architectures, a Schema Transformation Approach," in Proc. of CAiSE'02, Toronto, May, pp. 484-499, 2002.
16. Miller R.J., Hernandez M.A., Haas L.M., "Schema Mapping as Query Discovery," Proc. of the 26th Int. Conf. on VLDB'00, Cairo, Egypt, pp. 77-88, 2000.
17. Mohania M.K. and Dong G., "Algorithms for Adapting Materialized Views in Data Warehouses," in CODAS, pp.309-316, 1996.
18. Moraes Batista M.C. and Salgado A.C., "Minimality Quality Criterion Evaluation for Integrated Schemas", ICDM, 2007.
19. Mumick I., Quass D. and Mumick B., "Maintenance of Data Cubes and Summary Tables in a Warehouse," in SIGMOD, pp. 100-111, May 1997.
20. Peralta V., « Data Quality Evaluation in Data Integration Systems », thèse de Doctorat, Universidad de la República (Uruguay), Novembre 2006.
21. Popa L., Velegrakis Y., Miller R.J., Hernandez M.A., Fagin R. "Translating web data", Proc. of the 28th Int. Conf. on VLDB'02, Hong Kong, China, pp. 598-609, 2002.
22. Shvaiko P. and Euzenat J., "A Survey of Schema-based Matching Approaches", Proc. Of Journal Data Semantics IV, pp. 146-171, 2005.
23. Soukane A., "Génération automatique des requêtes de médiation dans un environnement hétérogène", thèse de Docteur de l'université de Versailles, 8 Décembre 2005.
24. Theodoratos D. and Sellis T.K., "Designing Data Warehouses", Data Knowledge Engineering, 31(3), pp. 279-301, 1999.
25. Velegrakis Y., Miller R.J., and Popa L., "Mapping Adaptation under Evolving Schemas," in Proc. Of the 29th VLDB Conf., Berlin, 2003.
26. Wiederhold G., "Mediators in the architecture of future information systems", IEEE Computer, Vol. 25(3), pp. 38-49, 1992.
27. Xue X., "Automatic Mapping Generation and Adaptation for XML Data Sources", thèse de Docteur de l'université de Versailles Saint-Quentin en Yvelines, 8 Décembre 2006.
28. Yu C. and Popa L., "Semantic Adaptation of Schema Mapping when Schemas Evolve", Proc. of the 31st Int. Conf. on Very Large Data Bases, Trondheim, Norway, 2005.
29. Zamboulis L., "XML Data Integration by Graph Restructuring", BNCOD, pp. 57-71, 2004.