INTÉGRATION DE DONNÉES: CONTEXTES ET APPROCHES

Gilles Nachouki et Rafic Hage Chehade[1]
Labo. LINA – équipe GDD, faculté des sciences et des techniques, université de Nantes, France
[1] IUT Saida, université libanaise, Liban
rhagechehade@ul.edu.lb

RÉSUMÉ

Le besoin de partager et de connecter entre elles des sources de données hétérogènes augmente de plus en plus, en particulier avec la croissance du nombre de sources de données en provenance du Web. Depuis des décennies, de nombreux travaux de recherche dans le domaine des systèmes d'information se sont focalisés sur le problème de partage de données dans les deux contextes : fédéré et médiateur et plus récemment dans les deux contextes: pair-à-pair et le Web de données (ou linked data sources). Dans cet article, nous passerons en revue ces différentes approches en nous focalisant sur le partage de données dans le contexte du Web.

Mots-clés: conflits de données, approche fédérée, médiateur, pair-à-pair, web de données

ABSTRACT

The need to share and connect between the heterogeneous data sources is increasing more and more, especially with the growing number of data sources from the Web. For decades, research in the field of information systems have focused on the problem of data sharing in sevaral contexts: federated, mediator and more recently in peer-to-peer and Web data (or linked data sources). In this article, we review these approaches by focusing on data integration in the context of the Web.

Keywords: data conflicts, federated approach, mediator, peer-to-peer, data web

INTRODUCTION

Avec l'augmentation du nombre de sources de données, le besoin de partager et d'accéder à plusieurs sources pour les interroger conjointement s'accroit fortement. Le partage d'informations devient crucial avec les mutations technologiques de ces dernières années, tout particulièrement dans le domaine du Web. Le partage d'informations est essentiel aussi bien pour améliorer le temps d'accès que pour la gestion des applications complexes nécessitant des accès à des gros volumes de données souvent stockées sur des sites distants. Ce problème est connu depuis des décennies, sous le nom d'interopérabilité ou d'intégration de données. L'interopérabilité désigne la capacité d'échange de données et de services entre des systèmes en dépit de leurs différences naturelles. L'interopérabilité implique une distribution d'informations, une communication et une compréhension mutuelle entre différents domaines d'exécution (Boulanger & Dubois, 1997).

Les problèmes issus de l'interopérabilité se situent au cœur des préoccupations de recherches actuelles. Nous pouvons distinguer deux types d'interopérabilité: l'interopérabilité physique et l'interopérabilité sémantique. L'interopérabilité physique consiste à échanger des données et des services entre systèmes hétérogènes au niveau bas de l'inter-connectivité physique: ce niveau fait appel aux systèmes de communication et aux protocoles de transport de données; l'interopérabilité sémantique consiste à échanger des données et des services au dessus de l'inter-connectivité. Il s'agit ici de construire des ponts sémantiques permettant une compréhension mutuelle entre chaque paire de sources de données (Johannesson & Jamil, 1994; Pitoura *et al*., 1995). Aujourd'hui avec le développement des réseaux et en particulier le Web il est possible d'accéder à des bases de données distantes, structurées ou semi structurées et souvent hétérogènes. Ces sources de données sont souvent modifiées (par exemple, des pages web) et leur intégration nécessite de ce fait des techniques nouvelles.

Le problème d'intégration a fait l'objet de nombreuses investigations par des chercheurs du monde entier afin de proposer des solutions d'intégration de données essentiellement dans les deux contextes: fédéré et médiateur mais aussi, plus récemment, dans les contextes pair-à-pair et le Web de données. Dans cet article, nous introduisons un cadre théorique et méthodologique à la problématique d'interopérabilité sémantique des sources de données (bases de données relationnelles ou objets, XML, *etc*…) hétérogènes afin de présenter les principales approches proposées dans la littérature.

CADRE THÉORIQUE ET APPROCHES

Les bases de données ont pour objectif la modélisation d'un univers à l'aide d'un ensemble intégré de données. Face à l'accroissement des quantités de données à stocker et à gérer, deux démarches de conception sont apparues: la première est la création au niveau de l'entreprise d'une vaste base de données centralisée. Toutefois la création de cette base est techniquement impossible en raison de la lourde tâche de gestion qu'elle nécessite mais également en raison du problème de performance. La deuxième démarche réside dans la création de plusieurs bases de données. Ces bases sont souvent plus petites et en général conçues plus ou moins indépendamment les unes des autres au niveau de chaque service ou département au sein d'une même entreprise. Deux principales méthodes ont été proposées pour la conception de bases de données réparties (Batini *et al*., 1986; Breitbart *et al*., 1986) (Ceri *et al*., 1987): la première suit une approche descendante et la deuxième suit une approche ascendante. L'approche descendante consiste à répartir le schéma d'une base de données sur plusieurs sites (services ou départements) alors que dans l'approche ascendante, partant d'un ensemble de schémas de bases de données à intégrer, l'objectif est de proposer un schéma unifié global modélisant l'ensemble de ces données. Nous nous intéressons ici à l'approche ascendante. Parmi les principaux problèmes posés dans le cadre de cette approche, nous distinguons la recherche des correspondances sémantiques inter-schémas *schéma matching* (Rahm & Bernstein, 2001) de bases de données et la spécification des règles d'intégration associant les schémas des bases de données locaux au schéma global *schéma mapping* (Miller *et al*., 2000). Deux éléments (dans deux schémas différents) sont dits en correspondance sémantique (Parent & Spaccapietra, 1996) s'ils décrivent le même élément du monde réel (objet, lien ou propriété). La définition intentionnelle d'une correspondance sémantique est appelée assertion. Un point important dans le processus d'intégration est la cohérence de ces correspondances. Dans ce processus, il est souvent nécessaire de demander à l'administrateur de confirmer ou infirmer des correspondances. Une fois que les correspondances sont établies, l'intégration proprement dite peut commencer. Ainsi, chaque

assertion est analysée pour déterminer quelle est la représentation des éléments en correspondance qui doit être incluse dans le schéma global et ensuite définir des règles réunissant le schéma global aux schémas initiaux des bases de données.

Dans la section suivante, nous commençons par recenser les principaux types de conflits existant entre des sources de données hétérogènes puis nous présentons des techniques de recherche (semi-automatique) des correspondances inter-schémas (schéma matching) ainsi que d'autres techniques utilisées pour associer des schémas locaux au schéma médiateur (schéma mapping). Ces deux types de techniques sont indispensables au processus d'intégration de données. Nous proposons ensuite un aperçu de quelques approches d'intégration de sources de données dans les contextes fédéré, médiation, pair-à-pair et Web de données.

CONFLITS DE DONNÉES

Différents types de conflits entre des schémas de sources de données sont distingués dans la littérature (Parent & Spaccapietra, 1996): structurels, données/méta-données, classification, descriptifs et modèles de données.

***Structurels*:** Les conflits structurels apparaissent lorsqu'il y a des différences entre diverses modélisations possibles dans deux schémas. Ainsi, le même concept peut être modélisé avec deux constructeurs différents. Par exemple, le nom d'un attribut dans un schéma correspond à une relation dans un autre schéma.

***Données/méta-données*:** Il existe des conflits de type données/méta-données lorsqu'une donnée est en relation avec une méta-donnée, par exemple, *conférence* est une valeur associée à l'attribut *type-publication* dans un schéma alors que *conférence* est modélisée comme étant un attribut dans un autre schéma.

***Classification*:** Les conflits de classification apparaissent lorsqu'il existe des différences dans le contenu de deux attributs dans deux schémas. Par exemple, l'attribut *article* dans un schéma décrit des articles de conférences et dans un autre schéma cet attribut décrit des articles de conférences et de journaux. Généralement, ce type de conflits se traduit à travers des assertions par l'utilisation d'une relation ensembliste telle que $\subseteq$ et $\cap$ *etc…* (autre que l'équivalence).

***Descriptifs*:** Les conflits descriptifs apparaissent entre concepts possédant des caractéristiques différentes bien qu'ils soient proches sémantiquement. Inversement, ce type de correspondances apparait entre concepts possédant des caractéristiques communes bien qu'ils soient différents sémantiquement. Nous trouvons plusieurs cas :

- *Nommage*. Un même nom peut être utilisé pour dénoter deux concepts différents (problème d'homonymie) ou un même concept est décrit par des noms différents (problème de synonymie).
- *Identité*. Un même concept peut être identifié dans deux schémas différents avec deux identificateurs différents. Par exemple, le même concept employé est identifié dans le schéma Sh_1 par un numéro *matricule* et dans le schéma Sh_2 par un *nss (numéro de sécurité sociale).*
- *Échelle*. Ce type de conflits décrit un même objet qui est représenté par des unités

différentes. On peut trouver également des conflits proches tel que le conflit de précision ou des conflits concernant deux attributs sémantiquement équivalents mais dont le domaine de définition sur lequel est défini chaque attribut est différent.

***Hétérogénéité de modèles*:** Lorsque les schémas à intégrer sont exprimés dans des modèles de données différents, on parle alors de correspondance d'hétérogénéité des modèles de données. La traduction de ces schémas dans un modèle pivot permet de résoudre ce type de conflits. Le paragraphe suivant présente des techniques de recherche des correspondances sémantiques inter-schémas.

RECHERCHE DES CORRESPONDANCES INTER-SCHÉMAS

La découverte de correspondances inter-schémas *schéma matching* est une étape importante dans le processus d'intégration de données où les schémas sources de données représentent un degré d'hétérogénéité assez élevé. Il est évident que la recherche des correspondances inter-sources est le seul moyen pour permettre le partage de données. Les correspondances sélectionnées servent plus tard pour établir des règles d'intégration *schéma mapping* afin de migrer des données d'un schéma source (ou plusieurs schémas sources) vers un schéma cible. Il est clair que la recherche de ces correspondances ne peut se faire manuellement surtout lorsqu'un grand nombre de sources sont impliquées ou lorsqu'une source comporte un grand nombre d'éléments. Les résultats des évaluations montrent que le processus de recherche de correspondances est loin d'être un processus complètement automatisé (Euzenat, 2009). Il s'avère que la recherche automatique des correspondances inter-sources nécessite souvent le recours à un expert pour atteindre un niveau de précision plus important: l'utilisateur doit déterminer les correspondances qui sont correctes et éliminer celles qui sont Fausses-Positives (ou correspondances trouvées par une application alors qu'elles ne sont pas correctes réellement). Pour cela de plus en plus des logiciels proposent un processus semi-automatique et itératif pour la découverte des correspondances inter-schémas. La recherche des correspondances inter-schémas a fait l'objet de plusieurs travaux de recherche ces dix dernières années, notamment dans les systèmes d'intégration de données. On peut trouver dans (Rahm, & Bernstein, 2001; Bellahsene *et al*., 2011) un résumé exhaustif des approches d'intégration de schémas, ainsi que des comparaisons entre plusieurs méthodologies. Des systèmes permettant la recherche des correspondances inter-schémas de façon (semi-)automatique, nous citons: *Falcon* (Hu & Cheng, 2008), *Cupid* (Madhavan *et al*., 2001) et *Spicy* (Bonifati *et al*., 2008).

MÉDIATION INTER-SCHÉMAS

L'intégration des schémas constitue une problématique importante. Ainsi, par exemple il est nécessaire d'établir un schéma global d'une base de données à partir des différents schémas/vues des utilisateurs (ou groupes d'utilisateurs). Un autre exemple est celui de l'hétérogénéité de données: les schémas sources de données sont souvent conçus séparément et présentent des hétérogénéités diverses au niveau des modèles de données utilisés ainsi qu'au niveau du nommage des éléments des schémas ou au niveau structurel. Dans le cadre de l'intégration de données, le schéma global ne peut souvent pas contenir les données de toutes les sources de données à intégrer en raison du grand volume que représentent ces données et de la mise à jour permanente de ces données. Le but de la médiation inter-schémas *schéma mapping* est de permettre de venir à bout de l'hétérogénéité

sémantique des schémas. Cette médiation associe les schémas sources à intégrer au schéma global décrivant l'ensemble de données. Les liens reliant des schémas sources au schéma global seront utilisés plus tard pour la réécriture des requêtes exprimées sur le schéma *global* en d'autres requêtes exprimées cette fois sur les schémas *sources*. Nous allons maintenant décrire succinctement ces techniques d'intégration.

APPROCHES D'INTÉGRATION

Nous présentons dans ce paragraphe des techniques d'intégration de données dans chacun des contextes suivants : fédéré, médiateur, pair-à-pair et Web de données.

***Contexte fédéré*:** Dans un contexte *fédéré*, après avoir recensé toutes les correspondances inter-schémas (décrites sous forme d'assertions), chaque assertion est analysée pour déterminer quelle est la représentation des éléments en correspondance qui doit être incluse dans le schéma global et proposer ensuite des règles réunissant le schéma global aux schémas initiaux des bases de données. Dans ce contexte, la fusion des schémas exports est fortement inspirée des travaux effectués sur l'intégration des vues (Batini *et al*., 1986; Parent & Spaccapietra, 1994).

***Contexte médiateur*:** Dans un contexte *médiateur*, plusieurs stratégies ont été proposées (Rahm et al., 2001) pour définir des correspondances entre un schéma global (ou médiateur) et les schémas sources (bases de données, XML *etc*…). Les stratégies pour la définition des correspondances sémantiques entre le schéma médiateur et les schémas locaux ont été appelées : *Global-As-View* (GAV), *Local-As-View* (LAV), *Global-Local-As-View* (GLAV) (Lenzerini, 2002), *Both As View* (BAV) (Brien & Poulovassilis, 2003) et *BYU Global Local As View* (BGLAV) (Xu, & Embley, 2004). Dans Miller *et al*. (2000), les auteurs traitent le problème de médiation comme un processus de découverte de requêtes (vues) permettant ainsi de transformer les schémas sources de données vers le schéma médiateur (prédéfini) à partir des correspondances entre éléments. La génération de ces requêtes est basée sur *Clio* (Haas *et al*., 1999; Hernandez *et al*., 2001), un prototype permettant de générer (de manière semi-automatique) des requêtes SQL réalisant des associations entre les schémas sources et le schéma global.

***Contexte pair-à-pair*:** Dans un contexte *Pair-à-Pair* (P2P), la définition d'un schéma global unique est pratiquement non conseillée ou impraticable. Pour cette raison la plupart des systèmes P2P évitent la maintenance d'un schéma unifié. À l'opposé, leurs approches se passent de la structure centrale et peuvent être classées en trois catégories : *médiation deux-à-deux*, *médiation entre petits groupes*, *médiation avec super-pairs*.

***Contexte Web de données*:** Le processus de traitement de requêtes sur le web s'est extrêmement complexifié allant du traitement de requêtes sur une source de données unique à plusieurs sources de données distribuées et maintenant à l'intégralité du Web surtout avec l'apparition du Linked Data Sources ou le Web de données. Dans un contexte *Web de données*, les sources de données sont décrites par des ontologies différentes connues par la communauté du web sémantique. Dans ce contexte les données des sources sont liées entre elles contenant des données appartenant à un ou plusieurs domaines (*e.g*. DrugBank, dbpedia, LinkedMdb, *etc*…) et une des problématiques de recherche est celle de la recherche des données. En effet, il est difficile à un utilisateur de connaître le contenu de toutes les sources de données qui sont accessibles sur le Web de données et par conséquent il a du mal à formuler ses requêtes. Ce problème vient du fait que les sources de données interrogées ne

possèdent pas de schéma unique (comme dans une base de données traditionnelle) et qu'un utilisateur souhaitant interroger le Web doit connaître les différentes sources de données (vocabulaires, liens inter-sources, *etc*…). Il est donc important de proposer une nouvelle approche offrant à l'utilisateur la possibilité de formuler ses requêtes sur un schéma global modélisant l'ensemble de données qu'il souhaite obtenir sans se soucier des contraintes physiques propres à chacune des sources. Le point commun entre la plupart des systèmes développés selon cette approche est qu'ils exigent la connaissance des sources de données à traiter lors de l'envoi de la requête. Un des défis à résoudre, dans ce contexte, consiste à rechercher sur le Web de données les sources de qui sont capables de répondre aux requêtes formulées par l'utilisateur et le traitement de ces requêtes.

CONCLUSION

Dans cet article, nous nous sommes intéressés au domaine de l'intégration de données dans des contextes différents: dans un contexte fédéré, un des problèmes à résoudre est celui de la traduction de schémas entre modèles. Ce problème est lié fortement à la nature du modèle commun choisi. Celui-ci doit être sélectionné en fonction de sa richesse sémantique qui doit être supérieure (ou égale) à celle des modèles conceptuels locaux (utilisés dans la conception des schémas locaux de bases de données). Cela permet d'enrichir le schéma fédéré par des informations supplémentaires initialement inexistantes dans les schémas locaux. Le deuxième problème concerne l'intégration de schémas locaux dans un seul schéma global. Dans un contexte médiateur et dans un environnement dynamique tel que le web, l'objectif est de fournir une intégration *flexible* des sources de données. Cette flexibilité conduit à une nouvelle approche dans laquelle n'existe pas une véritable intégration de schémas de sources de données dans un seul schéma global. Dans un contexte pair-à-pair, les principales approches proposées dans la littérature pour l'intégration de données ne sont pas adaptées car elles imposent un schéma global qui en plus d'être un frein à l'évolution de schémas, complique aussi le partage de données. Or, l'objectif de ces réseaux est de faciliter le partage et de permettre un accès rapide aux données. Dans un contexte Web de données, l'objectif est de proposer une nouvelle approche afin de donner l'impression à l'utilisateur de manipuler une seule source de données. Ainsi un utilisateur peut soumettre des requêtes sur un schéma global sans se soucier des vocabulaires ou des contraintes physiques telles que la façon d'accéder à ces sources ou le nombre de liens qui les relient.

Enfin, quelque soit le contexte, l'un des problèmes posés dans ce domaine est celui de la recherche des correspondances sémantiques entre les schémas à intégrer. Plusieurs techniques et des prototypes de recherche ont été proposés dans la littérature pour la mise en correspondance des éléments des schémas des sources de données à intégrer.

RÉFÉRENCES

Batini, C., Lenzerini, M. and Navathe, S. 1986. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys (CSUR)*, 18: 323-364.

Bellahsene, Z., Bonifati, A. and Rahm, E. 2011. *Schema matching and mapping*. Springer.

Bonifati, A., Mecca, G., Pappalardo, A., Raunich, S. and Summa, G. 2008. The spicy system: towards a notion of mapping quality. *Proceedings of International Conference on Management of Data (SIGMOD)*, p. 1289-1294.

Boulanger, D. and Dubois, G. 1997. Objets et coopération de systèmes d'information. *In: Ingénierie Objet: Concepts et Techniques*, éd. C. Oussalah, InterEditions, p. 339-

376.

Breitbart, Y., Olson, P.L. and Thompson, G.R. 1986. Database integration in a distribution heterogeneous database system. *Proceedings* of *International Conference of Data Engineering (ICDE)*, pp. 301-310.

Brien M.C.P. and Poulovassilis, A. 2003. Data integration by bi-directional schema transformation rules. *Proceedings* of *International Conference of Data Engineering (ICDE)*, pp. 227-238.

Ceri, S., Pernici, B. and Widerhold, G. 1987. Distributed database design methodologies. *Proceedings of the IEEE*, 75: 533 – 546.

Euzenat, J. 2009. Results of the ontology alignment evaluation. *Proceedings of the 4th International Workshop on Ontology Matching*, p. 73-126.

Haas, L.M., Miller, R.J., Niswonger, B., Tork M., Schwaez, P.M. and Wimmers, E.L. 1999. Transforming heterogeneous data with database middleware: beyond integration. *IEEE Data Engineering Bulletin*, 22: 31-36.

Hernandez, M.A., Miller, R.J. and Haas, L.M. 2001. Clio : A semi-automatic tool for schema mapping. *Proceedings of the International Conference on Management of data (SIGMOD)*, p. 607.

Hu, W. and Cheng, G. 2008. Matching large ontologies: A divide-and-conquer-approach. *Data Knowledge Engineering (DKE)*, 67: 140–160.

Johannesson, P. et Jamil, H. 1994. Semantic interoperability : Context, issues, and research directions. *Proceedings of the Int. Conf. on Cooperating Information Systems (Coopis)*, p. 180-191.

Lenzerini, M. 2002. Data integration: a theoretical perspective. *Proceedings of the Twenty-First Symposium on Principles of Database Systems*, p. 233-246.

Madhavan, J., Bernstein, P.A. and Rahm, E. 2001. Generic schema matching with cupid. *Proceedings of the International Conference on Very Large DataBases (VLDB)*, p. 49-58.

Miller, R.J., Haas, L.M. and Hernandez, M.A. 2000. Schema mapping as query discovery. *Very Large DataBase (VLDB)*, p. 77-88.

Parent, C. and Spaccapietra, S. 1996. Integration de bases de données: panorama des problèmes et des approches. *Ingénierie des Systèmes d'Information (ISI)*, 4: 1-18.

Parent, C. and Spaccapietra, S. 1994. View integration: a step forward in solving structural conflicts. *In IEEE Trans Data Knowl Data Eng (TKDE)*, 6: 258-274.

Pitoura, E., Bukhres, O. and Elmagarmid, A. 1995. Object orientation in multidatabase systems. *In ACM Computing Surveys*, 27: 141-195.

Rahm, E. and Bernstein, P.A. 2001. A survey of approaches to automatic schema matching. *Very Large Data Bases Journal*, 10: 334-350.

Xu, L. and Embley, D.W. 2004. Combining the best of global-as-view and local-as-view for data integration. *Information Systems Technology and its Applications (ISTA)*, pp. 123-136.