Médiation sur la toile

Information Integration using Mediation

Auteur à contacter :
François GOASDOUÉ
Tél. : 01 69 15 58 46
Fax : 01 69 15 65 86
Mél. : fg@lri.fr

Auteurs :
François GOASDOUÉ
LRI (Université Paris Sud & CNRS), UR INRIA Futurs*
Bâtiment 490, Université Paris Sud
F-91405 Orsay Cedex

Marie-Christine ROUSSET
LRI (Université Paris Sud & CNRS), UR INRIA Futurs*
Bâtiment 490, Université Paris Sud
F-91405 Orsay Cedex

1er jet reçu le 16/09/03

* GEMO, Pôle Commun de Recherche en Informatique du plateau de Saclay, CNRS, École Polytechnique, INRIA et Université Paris Sud.

Résumé :

L'intégration d'information est une discipline récente et incontournable de l'informatique. Elle a pour but de faciliter aux utilisateurs de moyens informatiques l'accès aux informations disséminées sur les réseaux (Internet, intranets,…). Les applications d'intégration d'informations les plus connues du grand public sont certainement les moteurs de recherche (Google, Voila, Yahoo…). Une méthode d'intégration nommée *médiation* permet d'aller au-delà des services rendus par ces moteurs en concevant par exemple des portails de commerce électronique capables d'intégrer les données de plusieurs fournisseurs de contenu (Kelkoo,…). C'est cette méthode que nous abordons ici en présentant les travaux menés sur ce sujet dans l'équipe Intelligence Artificielle et Systèmes d'Inférence du Laboratoire de Recherche en Informatique de l'Université Paris-Sud.

Information integration is a recent key domain of computer science. It aims at facilitating to users the access of information distributed over networks (Internet, intranets). Well known information integration applications are search engines (Google, Voila, Yahoo,…).
An integration method called mediation *allows encompassing the services rendered by those engines by building for example web portals that integrate data from several content providers (Kelkoo,…). We introduce this method by presenting the work done on that topic in the Artificial Intelligence and Inference Systems group of the computer science laboratory of the Université Paris-Sud.*

Quotidiennement, nous faisons appel à différentes sources d'information (journaux, livres, personnes, télévisions,…) pour obtenir des éléments de réponse à des questions qui nous intéressent ou nous sont utiles. En général, nos recherches d'information aboutissent rapidement car nous savons à quelles sources nous adresser et comment interagir avec elles. De plus en plus de sources d'information sont stockées sur support informatique sous la forme de fichiers, bases de données ou pages Web, et sont accessibles via des réseaux tels des Intranets ou Internet. Ainsi, le Web donne un accès potentiel à un immense gisement d'informations qui sont disséminées mondialement et représentées/codées dans des formats de données variés. Interroger cette mine d'informations pour trouver des réponses pertinentes à des questions précises est une tâche difficile qui n'est pas résolue par les moteurs de recherche actuels tels que Altavista, Google, Voila ou Yahoo. En effet, ces moteurs de recherche sont de simples outils de localisation d'adresses URL des pages Web contenant des mots-clés fournis par l'utilisateur pour exprimer sa demande d'information (i.e., sa requête). Ils ne permettent pas de fournir une réponse précise à une question précise comme « quelles sont les adresses de restaurants gastronomiques dans le quartier latin ? ».

Un médiateur entre les utilisateurs et le réseau

L'intégration d'informations est un domaine de recherche en informatique qui a pour but de proposer des solutions pour la construction de systèmes de questions/réponses sur des informations distribuées et hétérogènes provenant de multiples sources de données pouvant être disparates. L'une de ces solutions est la *médiation*. Elle est née dans les années quatre-vingts lorsque le développement des réseaux donna accès à de nombreuses sources d'information et suscita l'envie de pouvoir les interroger de façon uniforme.

La médiation est fondée sur la construction de systèmes jouant le rôle de *médiateurs* entre les utilisateurs et le réseau en fournissant une interface d'interrogation centralisée et unique au-dessus de sources d'informations réparties et hétérogènes (voir Figure 1). Les utilisateurs

recherchant des informations sur un réseau n'ont donc qu'à savoir interroger le médiateur et comprendre le format de données de ses réponses. Le médiateur a à sa charge la recherche des réponses aux requêtes qui lui sont posées en identifiant les sources de données pertinentes, en les interrogeant une à une dans leur langage et leur vocabulaire de requêtes (qui varient d'une sources à l'autre), et éventuellement en combinant leurs données (dont le format peut varier d'une source à l'autre).

Les services rendus par la médiation la place à la base d'importants enjeux.
Elle permet une meilleure productivité au sein des entreprises par le déploiement rapide de systèmes d'information intégrant des bases de données existantes sans devoir changer ou reformater les données qui y sont stockées.
Pour certaines entreprises, la médiation est bien plus qu'une source de productivité, c'est une source de revenus. C'est le cas des entreprises déployant des portails de commerce électronique (Kelkoo, premier comparateur de prix européen, http://fr.kelkoo.com/).
Enfin, la médiation permet aussi de mettre en œuvre des choix politiques comme la simplification de l'accès aux administrations pour les citoyens au travers de portails internet (Administration électronique).

Fournir la bonne réponse

L'équipe Intelligence Artificielle et Systèmes d'Inférence du Laboratoire de Recherche en Informatique de l'Université Paris Sud est engagée dans la recherche sur la médiation depuis plus de cinq ans sur un thème central : la caractérisation de l'ensemble des réponses à une requête qui peuvent être calculées non pas directement à partir des données (restant stockées au niveau des sources d'information) mais seulement de leur description abstraite dont dispose le médiateur sous la forme de formules logiques appelées des *vues*. Cette

caractérisation permet de qualifier l'ensemble des réponses produites, et ainsi de garantir par exemple de ne produire *que* des réponses certaines, ou *toutes* les réponses certaines, selon les critères de qualité exigés ou désirés. Certaines applications ont en effet de fortes contraintes de qualité sur les réponses à fournir qui se traduisent par des propriétés dites de *correction* et de *complétude*. Par exemple, pour le commerce électronique qui est un des domaines d'application de nos travaux, un utilisateur attend que *tous* les produits qui lui sont proposés à la vente correspondent à la demande qu'il a formulée (correction). De plus, les fournisseurs de contenu souhaitent que *tous* les produits qui répondent aux souhaits d'un utilisateur lui soient proposés (complétude).

La difficulté du problème du calcul des réponses à une requête posée à un médiateur provient du fait que le médiateur n'a pas à sa disposition directe les données pour répondre à la requête mais seulement des vues décrivant par des formules logiques le contenu des différentes sources d'information qu'il peut interroger. Par exemple, il peut savoir que la source « Pariscope » fournit des adresses de restaurants à Paris et la source « Michelin » fournit l'adresse de tous les restaurants gastronomiques français ayant obtenu des étoiles au guide Michelin.

Pour répondre à une requête, un médiateur doit donc reformuler la requête initiale en un plan de requêtes directement exécutable sur les sources de données disponibles. Cette reformulation est obtenue par *réécriture de la requête en terme de vues.*

Une approche générique fondée sur la logique

La réécriture de requêtes en terme de vues est un problème d'inférence : il s'agit de produire des plans de requêtes qui impliquent logiquement la requête initiale afin de garantir que les réponses obtenues par exécution de ces plans sont bien des réponses certaines pour la requête

initiale. Nous avons étudié le problème de réécriture de requêtes en terme de vues, sa décidabilité, sa complexité, dans un cadre logique général et expressif généralisant la plupart des travaux existants sur les médiateurs. Ce cadre est fondé sur un langage de représentation de connaissances appelé *CARIN* (*C*oncepts And *R*ules *IN*tegrated) combinant deux paradigmes classiques en Intelligence Artificielle et en Bases de Données : les langages à base de classes et les langages à base de règles. Dans ce cadre, nous avons établi une condition *suffisante* générale pour que le calcul des réponses à une requête d'un médiateur soit correct, complet et efficace : *il suffit que la requête posée ait un nombre fini de réécritures maximales en termes des vues du médiateur dans le langage des requêtes conjonctives*.

Nous avons ensuite identifé des langages de requêtes et de vues intéressants d'un point de vue pratique et pour lesquels la condition présentée ci-dessus est respectée. Comprendre la portée pratique de ce travail théorique nous a permis de faire des choix pertinents qui ont été mis en œuvre dans les médiateurs PICSEL et MKBEEM développés en collaboration avec France Télécom.

Quel avenir pour la médiation ?

L'arrivée massive sur les réseaux de nouveaux utilisateurs et de leurs données fait apparaître de nouveaux besoins d'intégration d'informations. Le plus médiatique actuellement est l'échange de fichiers multimédia (musique, vidéos, etc.) entre internautes. Pour ces nouvelles applications, la médiation n'est pas adaptée car la centralisation de l'accès à un tel volume de données offre des performances trop faibles. Les médiateurs peuvent intégrer efficacement les données de quelques grandes entreprises, mais pas celles de millions d'utilisateurs.

Pour ces besoins nouveaux, une *médiation distribuée* est en passe d'émerger avec l'apparition des systèmes pair-à-pair (peer-to-peer). À titre d'exemple, on peut citer Napster ou Gnutella pour l'échange de fichiers multimédia.

Les systèmes pair-à-pair sont composés de pairs (serveurs autonomes). Chaque pair peut jouer indifféremment le rôle de serveur de données ou de médiateur avec les autres pairs, pour participer de manière distribuée et collaborative au calcul de réponses à des requêtes. Une requête est posée à un pair qui y répond à l'aide des données qui lui sont directement accessibles ou dont il sait qu'elles sont accessibles via un autre serveur.

Médiateurs et systèmes pair-à-pair sont des systèmes *incontournables et complémentaires*. Selon le volume de données à intégrer, on préfèrera une solution simple et rapide à mettre en œuvre grâce aux médiateurs pour des volumes de données raisonnables, ou des solutions plus complexes à base de systèmes pair-à-pair lorsque les données sont très nombreuses.

Nous travaillons actuellement à étendre nos résultats sur la médiation centralisée à la médiation distribuée. L'intérêt d'étudier le calcul des réponses à une requête dans les systèmes pair-à-pair est évident. Ceci va permettre de faire « passer à l'échelle » la taille des données pouvant être intégrées dans les applications contraintes qualitativement, aujourd'hui limitées à la capacité d'intégration des médiateurs (systèmes d'information, sites de e-commerce ou encore les portails internet/intranet).
C'est à cette étude que nous travaillons actuellement et nos premiers résultats sont encourageants.

Encadré

Les médiateurs PICSEL et MKBEEM

Nous avons mis en œuvre l'un de nos algorithmes de réécriture dans le projet PICSEL (Production d'Interfaces à bases de Connaissances pour des Services En Ligne, http://www.lri.fr/~picsel, voir Figure 2) réalisé pour France Télécom dans le cadre d'un CRE (Contrat de Recherche Externe) et en collaboration avec l'agence de voyage Dégriftour (http://www.degriftour.fr).
Ce projet a conduit au dépôt à l'APP (Agence pour la Protection des Programmes) de deux logiciels permettant de créer des médiateurs. L'avantage de ces médiateurs est de posséder des langages logiques puissants pour poser des requêtes très précises sur les informations à rechercher et pour décrire de façon abstraite le contenu des sources connectées au médiateur. L'inconvénient est que ces langages sont trop complexes pour être faciles à utiliser par un large public.

Nous avons élargi le public des médiateurs PICSEL dans un projet communautaire de l'IST (Information Society Technologies) sur le commerce électronique multilingue, MKBEEM (Multilingual Knowledge Based European Electronic Marketplace, http://www.mkbeem.com), rassemblant de nombreux centres de recherche scientifique (CNRS, Université de Clermont-Ferrand, Université de Montpellier, Université Polytechnique de Madrid, Université Paris Sud, Université Technique d'Athènes, VTT - Centre de recherche technique de Finlande) et industrielle (Ellos - La Redoute, France Télécom, Sema – Schlumberger, SNCF et Tradezone International Ldt).

L'un des buts de MKBEEM était de doter les médiateurs PICSEL d'interfaces en langue naturelle afin de le rendre « grand public ». Ainsi, les utilisateurs de médiateurs MKBEEM peuvent poser leurs requêtes dans leur propre langue.

Dans le prototype réalisé, les utilisateurs s'expriment en anglais, espagnol, finnois, français ou encore suédois pour acheter des produits de la SNCF Voyages (billets de train ou d'avion, location de voiture, réservation d'hôtel) et d'Ellos-La Redoute.

MKBEEM a été élu meilleure application des technologies du traitement de la langue par le réseau communautaire Euromap de transfert des technologies du traitement de la langue (E.work & E.commerce, 2001). Il a donné lieu, à ce jour, à plus de quarante articles, trois brevets et quatre transferts industriels. Un projet de transfert vers Orange et Wanadoo est en cours de montage par France Télécom R&D.

Glossaire

Format de données : représentation informatique d'informations tel XML.

Fournisseur de contenu : société mettant en vente des articles de son catalogue sur un site de commerce électronique.

Inférence : étape de raisonnement logique basé sur la déduction.

Requête : question posée à un système informatique dans son langage et dans son vocabulaire.

URL : adresse internet référençant un document. Par exemple, http://www.lri.fr est l'URL du Laboratoire de Recherche en Informatique de l'université Paris Sud.

Pour aller plus loin

François Goasdoué, Véronique Lattes and Marie-Christine Rousset, *The Use of CARIN Language and Algorithms for Information Integration: The PICSEL Project, International*

Journal of Cooperative Information Systems (IJCIS), World Scientific Publishing Company, Vol. 9, No. 4, p. 383-401, 2000.

Alain Bidault, Christine Froidevaux, Hélène Gagliardi, François Goasdoué, Chantal Reynaud, Marie-Christine Rousset et Brigitte Safar, *Construction de Médiateurs pour Intégrer des Sources d'Information Multiples et Hétérogènes : le Projet PICSEL*, Journal I3 : Information - Interaction - Intelligence, Vol. 2, Num. 1, 2002.

O. Corcho, A. Gomez-Perez, A. Léger, C. Rey and F. Toumani, *An Ontology-based mediation architecture for E-commerce applications*, Intelligent Information Systems 2003, 02-05 June 2003, Zakopane, Poland

François Goasdoué and Marie-Christine Rousset, *Querying Distributed Data through Distributed Ontologies: a Simple but Scalable Approach*, IEEE Intelligent Systems, septembre/octobre 2003.

François Goasdoué and Marie-Christine Rousset, *Answering Queries using Views: a KRDB Perspective for the Semantic Web*, ACM Journal - Transactions on Internet Technology (TOIT), Vol. 4, Num. 3, 2004.

Images et figures

Image de la page d'introduction à l'article

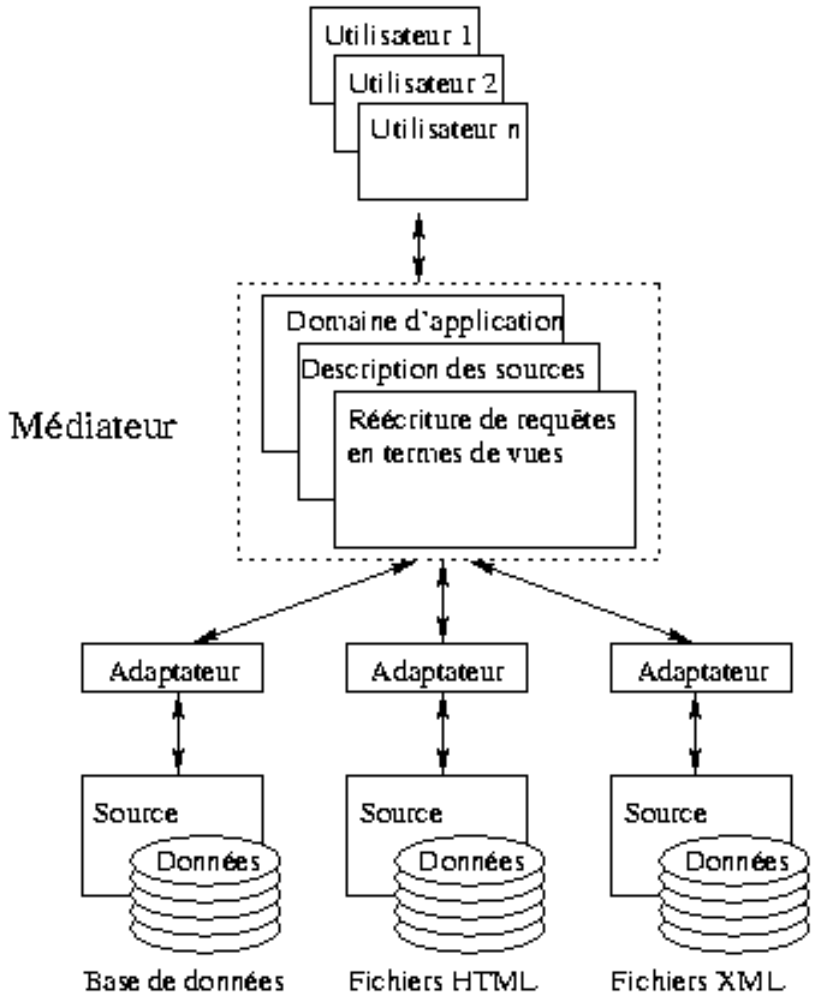

Figure 1 : Architecture d'un médiateur

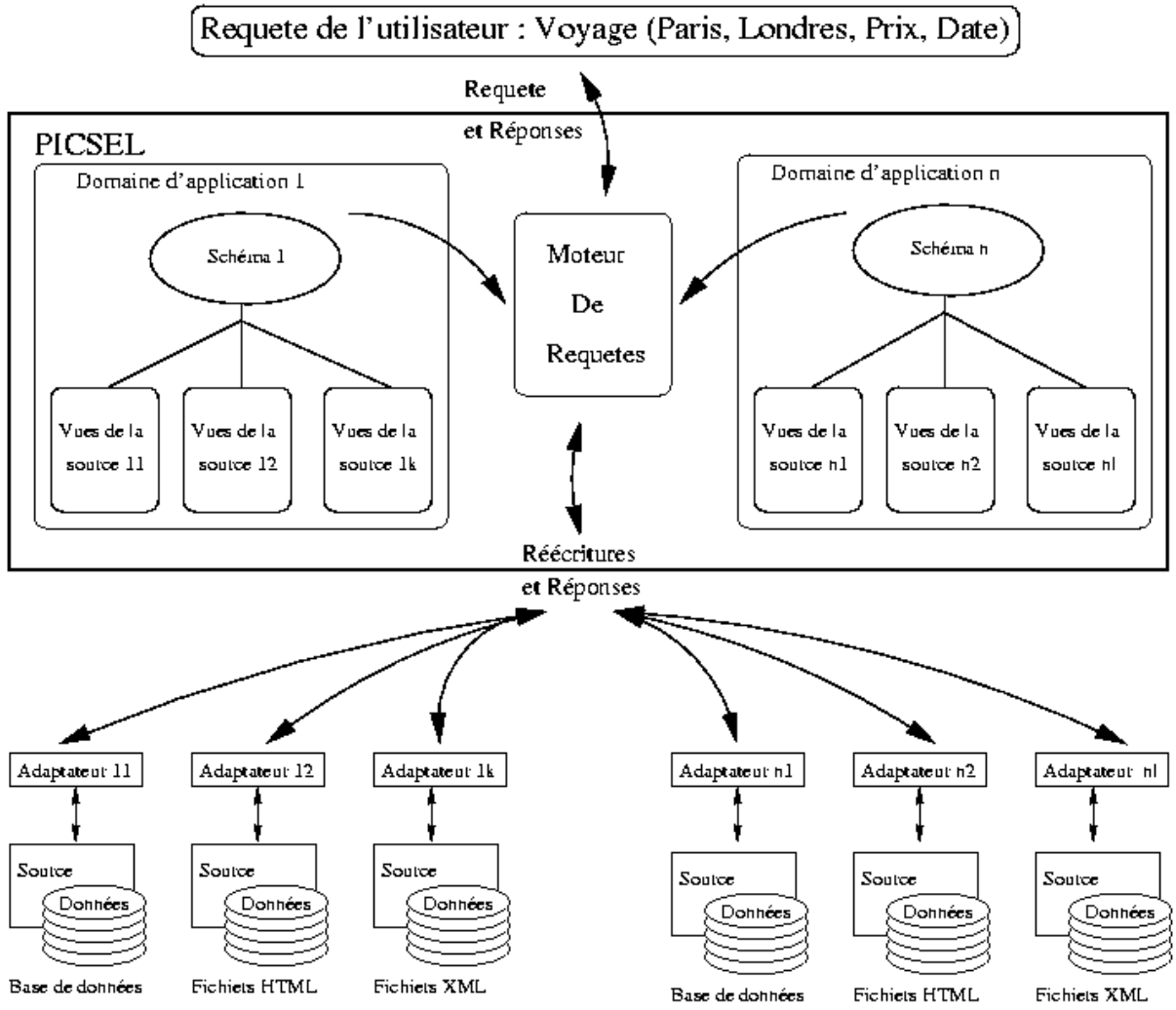

Figure 2 : Architecture d'un médiateur PICSEL