INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet CARAVEL

Systèmes de médiation d'information

Rocquencourt

THÈME 3A

Table des matières

Le projet Caravel est un nouveau projet issu de l'ancien projet Rodin. Pour plus d'information sur la transition de Rodin vers Caravel, consulter la page web du projet (`http://www-caravel.inria.fr/`).

1 Composition de l'équipe

Responsable scientifique

Eric Simon [DR Inria]

Responsable permanent

Patrick Valduriez [DR Inria]

Assistante de projet

Elisabeth Baqué [AI]

Personnel Inria

Daniela Florescu [CR]

Francois Llirbat [CR]

Conseiller scientifique

Georges Gardarin [professeur, université de Versailles]

Collaborateurs extérieurs

Luc Bouganim [MC, université de Versailles]

Mokrane Bouzeghoub [professeur, université de Versailles]

Françoise Fabret [chargée de travaux, CNAM, jusqu'au 1er mai]

Esther Pacitti [université de Rio de Janeiro]

Chercheurs invités

Chandra Mohan [IBM Almaden, USA, 6 mois]

Dennis Shasha [université de New York, USA, 6 mois]

Ingénieurs experts

Mokrane Amzal [depuis le 1er aout]

Francoise Fabret [depuis le 1er mai]

Florian Xhumari

Doctorants

Mokrane Amzal [boursier Inria, université de Versailles, jusqu'au 1er aout]

Helena Galhardas [université de Lisbonne]

Alberto Lerner [université de Rio de Janeiro, depuis le 1er septembre]

Olga Kapitskaia [AT&T Research, université Paris 6]

Ioana Manolescu [université de Versailles]

Maja Matulovic [boursière Inria, université de Versailles, jusqu'au 1er avril]

Hubert Naacke [boursier Inria, université de Versailles, jusqu'au 1er novembre]

Fabio Porto [université de Rio de Janeiro, depuis le 1er mai]

Joao Pereira [université de Lisbonne]

Khaled Yagoub [boursier MESR, université de Versailles]

Stagiaires

Abderrahim Taoudi [ENSIAS Rabat, Maroc]

Chafiq Ziazi [EMI Rabat, Maroc]

Spyridon Ligoudistianos [NTUA Athènes, Grèce]

Karthik Ranganathan [IIT Madras, Inde]

Jean-Pierre Matsumoto [Université Paris VII]

2 Présentation et objectifs généraux

L'objectif principal du projet CARAVEL est de concevoir et d'expérimenter des techniques qui permettent d'offrir une vue intégrée, cohérente, pertinente et actualisée de l'information disponible dans un réseau de sources d'information hétérogènes et autonomes – abrégé par *réseau d'informations* – ainsi que des moyens de navigation efficaces dans cette information pour différentes catégories d'utilisateurs. Cet objectif est mené au travers de quatre grands thèmes de recherche complémentaires visant à :

- *offrir un mode d'accès uniforme aux ressources d'un réseau d'informations.* Pour cela, nous développons des systèmes de médiation permettant soit d'interroger des données réparties via un langage de requêtes et de solliciter l'exécution de services sur ces données, soit de demander l'exécution de "workflows" qui effectuent des accès au réseau d'informations.
- *faciliter la construction et la maintenance d'entrepôts de données* résultant de la consolidation, de l'intégration et de la synthèse de données issues de sources multiples. Dans ce thème, nous développons des systèmes de médiation permettant l'intégration et la synthèse cohérente de données hétérogènes ainsi que le rafraîchissement automatique d'entrepôts de données.
- *faciliter la diffusion d'information.* Pour cela, nous développons des systèmes de "publication/souscription" servant d'intermédiaires entre des sources désirant publier des informations évoluant rapidement et des abonnés souhaitant être avertis (ou notifiés) des informations qui leur sont pertinentes.
- *offrir des moyens de navigation efficaces dans un réseau d'informations.* Pour cela, nous développons des systèmes qui d'une part facilitent la construction et la maintenance de sites Web garantissant de bonnes performances d'accès et d'autre part gèrent, via un modèle de données particulier, une multitude de façons pour un utilisateur de visualiser l'information et d'y accèder par raffinements successifs.

Les techniques développées dans ces thèmes de recherche prennent la forme de langages, de modèles de données ou d'algorithmes implementés dans des composants de type middleware, appelés *médiateurs*, qui s'interfacent entre des applications clientes et des serveurs d'information selon un modèle d'architecture à trois-tiers. Il existe plusieurs types de médiateurs qui sont précisés dans la suite de ce rapport d'activité. Un point fondamental est que les médiateurs développés dans le projet sont conçus de façon à être facilement assemblables entre eux, ce qui facilite leur utilisation combinée dans le déploiement d'applications et permet une grande synergie entre les différentes actions de recherche du projet.

3 Domaines d'applications

3.1 Systèmes d'information pour l'environnement

Résumé : *Le projet CARAVEL s'intéresse en priorité aux systèmes d'information pour l'environnement, domaine d'application riche en problèmes d'intégration*

d'information. Nous avons poursuivi l'analyse de la problématique de recherche liée aux systèmes d'information pour l'environnement à travers d'une part les projets européens Thetis (gestion de zones côtières) et Decair (prédiction de la qualité de l'air en milieu urbain), en collaboration avec le projet Air et d'autre part le projet franco-brésilien Ecobase. Plusieurs problèmes émergent de l'étude de ces applications : (1) faciliter la localisation des sources d'information (données ou services) pertinentes à une application et la manipulation de ces informations (par ex. utilisation d'un programme distant avec ses propres données locales), (2) faciliter l'intégration de données fortement hétérogènes et leur synthèse (par ex. construction de paramètres pour des modèles de prédiction de pollution), (3) organiser un "commerce électronique" de ces données sur des réseaux globaux, et enfin (4) aider à l'évaluation de la qualité des sources de données et à la comparaison des données synthétisées. Les recherches décrites dans la suite de ce rapport répondent aux trois premiers problèmes.

4 Résultats nouveaux

4.1 Recherche d'information dans un réseau d'informations structurées

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *Le problème général abordé dans ce thème est la fédération de sources d'information selon une approche de type "pull", ce qui signifie que les informations réparties sont accédées dynamiquement en réponse à des requêtes émises depuis des sites clients. Cette approche est appropriée, lorsque les informations stockées dans les sources changent fréquemment ou dans le cas où les sources d'information peuvent être interrogées sans que les informations puissent être copiées (par exemple pour des raisons juridiques). Cependant, cette approche soulève plusieurs problèmes : comment publier l'information et quel type d'information, comment intégrer l'information publiée, comment optimiser l'exécution de requêtes. L'étude de ces problèmes dans le contexte de domaines particuliers d'application nous a permis, d'une part de mettre en évidence de manière précise ces problèmes et d'autre part d'expérimenter nos solutions dans des applications réelles. Les recherches en cours dans cette action se développent dans trois directions. La première est le développement d'un système d'accès à un réseau d'informations. Ce système, appelé Le Select, est un médiateur qui permet de traiter des requêtes d'accès aux ressources d'un réseau d'informations. La seconde est la conception d'algorithmes d'optimisation statiques et dynamiques pour ce type de médiateur. La dernière est la conception d'un médiateur qui permet d'intégrer les services d'accès aux données offerts par un réseau d'information au travers d'un workflow. Dans ce cas, une requête au médiateur correspond à une demande d'exécution d'un workflow qui, génère éventuellement des accès aux ressources du réseau d'informations.*

4.1.1 Le Select

Participants : Florian Xhumari, Mokrane Amzal, Ioana Manolescu, Eric Simon.

L'étude, que nous avons menée dans le cadre du projet européen Thetis sur des applications de gestion intégrée de zones côtières en Méditerannée, nous a conduit à concevoir un nouveau système d'accès qui répond à deux objectifs : (1) faciliter la publication d'informations diverses (des données stockées, des schémas de base de données virtuelle, ou des programmes), et (2) supporter la manipulation de ces informations via un langage de requêtes. L'architecture de Le Select comprend deux types de composants : les adaptateurs et les médiateurs. Un site de publication se compose d'un médiateur et d'un ou de plusieurs adaptateurs. La publication d'information d'un certain type (e.g., des fichiers HTML, un programme C, une base de données, etc) nécessite la création d'un adaptateur qui lui est associé. Les adaptateurs de données ont un double rôle: ils uniformisent la représentation des données via un modèle de données relationnel étendu et ils exécutent des requêtes sur des sources de données locales. Les adaptateurs de programmes uniformisent la manière d'invoquer un programme et de spécifier ses données d'entrée au moyen de requêtes. Chaque adaptateur gère également les méta-données de la source qui lui est associée. Les médiateurs traduisent des requêtes globales pouvant porter sur plusieurs adaptateurs en un ensemble de requêtes locales (chacune pour un adaptateur) et une requête de composition du résultat. Ils fournissent aussi un système d'exécution pour intégrer les résultats des requêtes locales. Les médiateurs gèrent enfin les requêtes d'invocation de programmes. Les requêtes globales sont exprimées dans un langage proche de SQL, qui autorise l'invocation de fonctions sur des données, pourvu que ces fonctions soient préalablement publiées via des adaptateurs. Un site client se compose seulement d'un médiateur. Un site de publication peut également publier un schéma de base de données virtuelle, dont la spécification (à l'aide de requêtes) fait référence à des informations publiées par d'autres sites de publication. Dans ce cas, on dira que ce schéma correspond à une vue intégrée des informations publiées par les autres sites. Une première version du logiciel Le Select est opérationnelle et fait l'objet d'expérimentations dans le cadre du projet Thetis.

4.1.2 Modèle de coût générique

Participants : Hubert Naacke, Patrick Valduriez.

Pour traiter efficacement une requête, le médiateur doit produire un plan d'exécution optimisé, sous forme d'un arbre d'opérateurs algébriques, dont les sous-arbres sont envoyés aux adaptateurs concernés. Dans une base de données traditionnelle, l'optimisation de requêtes repose sur l'estimation de coût (en terme de temps de calcul) des différents plans d'exécution possibles. Dans un médiateur comme Le Select, ce type d'optimisation est délicat car les sources de données n'exportent pas d'information sur le coût des opérateurs supportés.

Notre approche à ce problème est nouvelle. Elle s'appuie sur un modèle de coût générique, qui peut être progressivement affiné par l'ajoût d'information de coûts spécifiques exportée par les adaptateurs. Ainsi, le développeur d'un adaptateur peut choisir de spécifier avec une interface standard, tout ou partie de l'information de coût. Par défaut, le médiateur supporte son modèle de coût générique et le corrige dynamiquement avec l'information importée des

adaptateurs.

Nous avons défini une méthode pour mesurer l'efficacité d'un modèle de coût. Puis, dans le contexte des sources Web documentaires, nous avons discerné les propriétés des sources, qui influencent grandement le temps de réponse des requêtes. Nous avons caractérisé des cas d'application en fonction du type des sources intégrées dans le système de médiation et des requêtes, que pose l'utilisateur.

Ensuite, nous avons construit un modèle de coût par défaut pour le médiateur et des modèles de coût spécialisés pour les adaptateurs. Grâce à la flexibilité du langage de description du modèle de coût, nous avons pu décrire précisément les aspects hétérogènes des méthodes d'accès aux données de chaque source. Puis, nous avons mesuré le gain apporté par le modèle de coût, pour ces cas d'application.

Pour l'optimisation logique des opérations de sélection dans les systèmes hétérogènes, nous obtenons une efficacité presque totale lorsque la stratégie de l'optimiseur vise, d'une part, à répartir les opérations entre le médiateur et les adaptateurs, et d'autre part, à employer au mieux les index et les vues matérialisées existant sur les sources. Du point de vue de l'optimisation physique, le modèle de coût nous permet de choisir l'algorithme le plus approprié pour traiter l'opération de jointure inter-site dans le médiateur.

4.1.3 Optimisation dynamique de requêtes

Participants : Luc Bouganim, Daniela Florescu, Francoise Fabret, Ioana Manolescu, Chandra Mohan, Patrick Valduriez.

Dans notre contexte réparti, les méta-données nécessaires à l'élaboration d'un plan d'exécution efficace peuvent être regroupées en quatre classes : les paramètres connus statiquement, avec plus ou moins d'exactitude (e.g., requête, statistiques, etc), les paramètres qui ne sont connus qu'au début de la phase d'exécution (e.g., ressources disponibles), les paramètres qui ne seront connus avec exactitude qu'en cours d'exécution (e.g., taille des résultats intermédiaires, etc), enfin les paramètres qui varient continuellement (e.g., débit du réseau, sources de données disponibles, taux d'arrivée des données, etc). Nous avons développé une méthode d'optimisation et d'exécution adaptée à ce contexte : nous proposons d'exploiter les méta-données dès qu'elles deviennent disponibles. Ainsi, un plan d'exécution initial (ou du moins, un ensemble de pré-calculs qui pourront faciliter l'élaboration de ce plan) est produit avec les méta-données connues statiquement, au début de la phase d'exécution. Ce plan fixe les grandes options pour l'exécution mais laisse certains degrés de liberté, ceux qui dépendent de paramètres inconnus à ce moment. Ensuite, l'exécution se déroule en plusieurs étapes entrecoupées par des phases de planification. Une phase de planification fournit un plan d'exécution conforme aux méta-données connues à ce moment précis. Elle est suivie par une phase d'exécution qui applique le plan et qui réagit conformément à ce qui a été prévu par le planificateur. Lorsqu'un événement susceptible de remettre en cause le plan se produit, la phase d'éxécution se termine et céde sa place à une nouvelle phase de planification. Cette méthode fournit un cadre générique pour l'optimisation dynamique. Son adaptation à un contexte donné demande de définir les heuristiques utilisées lors des phases de planification, les degrés de liberté par rapport au plan initial, ainsi que les événements remettant le plan en cause. Nous avons développé plusieurs al-

gorithmes, qui suivent cette méthode afin d'optimiser dynamiquement l'exécution des requêtes globales dans un médiateur d'accès aux données.

4.1.4 Médiateurs d'accès orienté workflows

Participants : François Llirbat, Eric Simon.

Certaines applications, comme le commerce électronique ou l'assistance électronique à la clientèle, exigent qu'un traitement spécifique soit appliqué à une requête en fonction de la valeur de ses paramètres d'appel (type de la demande, profil du client, etc). Une solution à la mise en oeuvre de telles applications consiste à spécifier un workflow qui, à partir des paramètres d'entrée de la requête, déduit de façon itérative et en fonction des réponses aux requêtes déjà posées, quelles sont les nouvelles requêtes à exécuter sur le réseau d'informations. Nous avons travaillé en collaboration avec l'équipe de Rick Hull, à Bell Labs (Etats-Unis), d'une part à la définition d'un modèle de workflow répondant à ces besoins et, d'autre part à la spécification d'un médiateur d'accès orienté workflow qui permet de piloter l'exécution d'un workflow en réponse à une requête. Un médiateur d'accès orienté workflow n'autorise l'accès aux ressources d'un réseau d'informations qu'à travers des requêtes prédéfinies par des workflows. Dans ce contexte, un problème intéressant est la mise au point de techniques d'optimisation tirant parti de la sémantique d'un workflow. Nous avons proposé un ensemble de techniques d'optimisations (algorithmes et heuristiques) basées sur une utilisation judicieuse du parallélisme et des possibilités d'exécution spéculative. Ces optimisations permettent de réduire le temps de réponse d'une requête tout en évitant une charge excessive des sources d'information sollicitées. Nous avons mis en oeuvre un médiateur d'accès orienté workflow basé sur ces techniques. En simulant l'environnement réparti, nous avons comparé et analysé l'efficacité de ces techniques pour des applications réelles ayant des caractéristiques différentes. Fort de cette expérience nous avons proposé des règles simples qui permettent de choisir les meilleures stratégies d'exécution en fonction des caractéristiques de l'application et des différentes charges sur le réseau d'information.

4.2 Recherche d'information dans un réseau d'informations semi-structurées

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *Le problème abordé dans ce thème est celui de l'interrogation et du stockage des données XML. Le formalisme XML émerge comme un format d'échange standard de données via Internet qui est bien adapté aux données dites "semi-structurées". La définition d'un langage d'interrogation des données XML est un premier problème à résoudre rapidement pour permettre le développement d'XML. Plusieurs langages ont été proposés, mais aucun standard n'est encore choisi. Un deuxième problème, tout aussi important est celui du stockage des données XML. En fait ces deux problèmes ne sont pas indépendants car la façon dont les données sont stockées se doit de tenir compte de la façon dont les données sont*

ensuite interrogées. Nos recherches nous ont conduits à proposer un langage d'interrogation structuré pour des données XML, puis à étendre ce langage pour lui permettre de mélanger recherche par mot clé avec recherche structurée. Enfin, nous avons mené une analyse approfondie sur l'organisation des données XML dans des tables relationnelles pour notre langage d'interrogation.

4.2.1 L'interrogation des données XML

Participants : Daniela Florescu, Ioana Manolescu.

En collaboration avec Mary Fernandez et Dan Suciu (ATT, USA) et Alon Levy (Université de Washington, USA), nous avons défini un langage d'interrogation de données XML, nommé XML-QL, qui supporte les fonctionnalités indispensables, requises dans les applications bases de données, telles l'extraction des données, la conversion, la transformation et l'intégration des données provenant de multiples sources hétérogènes. XML-QL est un langage déclaratif à la fois puissant et suffisamment simple pour pouvoir permettre une optimisation automatique efficace. Un avantage majeur de ce langage est la possibilité d'exprimer simplement des transformations de données XML d'un schéma à un autre ; cette caractéristique est d'autant plus intéressante que ce genre de traitement est très fréquent dans les bases de données. XML-QL a été le premier langage de requête pour XML proposé au Consortium W3C. Récemment, nous avons proposé une nouvelle extension de ce langage permettant de faire des recherches "floues", sans connaissance exacte de la structure des données, qui combinent des techniques d'interrogation structurée avec des recherches d'information à base de mots-clés.

4.2.2 Le stockage des données XML

Participant : Daniela Florescu.

Dans nos recherches, nous étudions également le problème du stockage des données XML dans un système de gestion de base de données (SGBD) relationnelle et celui de l'exécution des requêtes en utilisant le processeur natif de requêtes SQL de ces SGBDs. Nous examinons plusieurs façons d'organiser les données XML dans des tables relationnelles, et nous analysons la traduction des requêtes déclaratives XML-QL dans des requêtes SQL équivalentes, pour chacune des organisations de données proposées. Nous présentons une analyse comparative des performances obtenues. Cette analyse nous permet de mettre en évidence les compromis entre les différentes organisations possibles, en termes de taille de la base, de performances des requêtes et de performances des mises à jour. Malgré le fait que cette étude soit focalisée sur le langage de définition de données XML et le langage de requêtes XML-QL, les résultats obtenus relèvent plus généralement des modèles de données semi-structurées et d'une large classe de langages de requêtes pour des données semi-structurées.

4.3 Intégration et synthèse d'information

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *Ce thème aborde le problème de la fédération de sources d'information selon une approche de type "push", ce qui signifie qu'une source d'information, construite à partir de l'intégration d'autres sources, est maintenue automatiquement à jour en réponse à des changements signalés par ces sources. Cette approche est appropriée lorsque le processus d'intégration est particulièrement complexe ou dans le cas d'applications d'aide à la décision. Cependant, elle pose plusieurs problèmes : comment spécifier le processus d'intégration de plusieurs sources, comment et quand propager les changements des sources, comment spécifier le processus de rafraîchissement ? Nous étudions ces problèmes dans le contexte de domaines d'application précis. Pour l'instant, notre contexte est celui d'applications scientifiques, que nous abordons à travers le projet européen DECAIR et celui d'applications de télécommunications, fournies par Telecom Italia dans le cadre du projet européen DWQ. Ce thème comporte trois actions. Une première action est le développement d'un médiateur d'intégration, appelé AJAX, qui permet de spécifier un processus d'intégration et d'assister un utilisateur durant l'intégration proprement dite. Une autre action est la conception d'algorithmes de diffusion des changements des sources, qui garantissent une propriété de cohérence globale dans le réseau d'informations. Une troisième action est le développement d'un médiateur de rafraîchissement, qui permet de spécifier le processus de rafraîchissement d'une source à l'aide d'un workflow.*

4.3.1 Nettoyage de données

Participants : Daniela Florescu, Helena Galhardas, Dennis Shasha, Eric Simon.

De nombreuses applications, comme par exemple les applications d'aide à la décision, exigent que les données utilisées obéissent à certains critères de qualité. Un facteur important de détérioration de la qualité de données est le fait qu'un même objet réel peut être modélisé par plusieurs tuples, sans que l'utilisateur puisse se rendre compte que ces différents tuples représentent effectivement le même objet. Ce problème peut avoir plusieurs origines, parmi lesquelles les plus courantes sont les erreurs de saisie et l'absence de clé universelle - ce dernier point est dû, entre autres, au fait que les données proviennent de sources différentes ou d'erreurs de saisie.

Le traitement de ce problème est connu sous le nom de "nettoyage" de données. Nous avons mis au point une solution de nettoyage de données, qui se décline en trois points. Tout d'abord, nous proposons un environnement logiciel qui modélise un processus de nettoyage comme un assemblage de transformations appliquées aux données. Les transformations sont regroupées en quatre classes : "mapping", "matching", "clustering" et "merging". Deuxièmement, nous fournissons un langage bâti sur un sous-ensemble de SQL et sur des macro-opérateurs, qui permet de spécifier les transformations ainsi que leur enchaînement. Ce langage permet également de spécifier des points d'interaction avec l'utilisateur pendant le processus de nettoyage. Enfin, nous avons élaboré un ensemble de techniques d'optimisation adaptées aux applications de nettoyage de données, telles que l'évaluation mixte ou le "neighborhood hash join".

4.3.2 Maintien de la cohérence dans des bases de données répliquées

Participants : Esther Pacitti, Eric Simon.

Dans une base de données, où les données répliquées sont gérées selon un modèle asymétrique, une transaction peut valider, après mise à jour, une seule réplique (la copie primaire) sur le noeud maître. Les mises à jour sont ensuite propagées aux autres répliques (les copies secondaires) qui sont mises à jour dans des transactions de rafraîchissement séparées. La conception d'algorithmes chargés de maintenir la cohérence des répliques tout en minimisant la dégradation des performances due à la synchronisation des transactions de rafraîchissement est un problème crucial. Nous proposons un algorithme de rafraîchissement simple et général qui résout ce problème et nous prouvons sa correction. Le principe de l'algorithme est de faire attendre les transactions de rafraîchissement avant de les exécuter sur un noeud ayant des copies secondaires. Nous présentons ensuite deux optimisations majeures de cet algorithme. La première optimisation est basée sur des propriétés topologiques de la configuration des répliques. En particulier, nous caractérisons les noeuds pour lesquels il est inutile d'attendre. La deuxième optimisation améliore la fraîcheur des données en utilisant une stratégie de propagation immédiate des mises à jour. Notre évaluation de performances montre l'efficacité de cette optimisation.

4.3.3 Rafraîchissement d'entrepôts de données

Participants : Mokrane Bouzeghoub, Françoise Fabret, Maja Matulovic, Eric Simon.

La conception d'un entrepôt de données est une tâche complexe. Elle demande de choisir les vues à matérialiser dans l'entrepôt ainsi que le système de rafraîchissement de ces vues de façon à satisfaire, d'une part les exigences de l'utilisateur, d'autre part les restrictions imposées par les sources de données et enfin les contraintes techniques. Les exigences de l'utilisateur concernent essentiellement le temps de réponse aux requêtes et la qualité des données : fraîcheur, consistence, précision, disponibilité. Les sources, quant à elles, proposent une fenêtre de disponibilité pendant laquelle elles acceptent une surcharge de travail due au chargement ou au rafraîchissement de l'entrepôt. Si elles sont prêtes à accepter une forte surcharge lors du chargement initial de l'entrepôt, en revanche elles ne proposent généralement qu'une fenêtre restreinte de disponibilité lorsque l'entrepôt est devenu opérationnel. Nous avons montré que le rafraîchissement d'un entrepôt nécessite la planification de diverses tâches : stockage intermédiaire, nettoyage et intégration des données, mises à jour incrémentielles de l'entrepôt, etc. A partir de l'étude d'une variété d'applications, nous avons montré que ce processus se décrivait convenablement au niveau conceptuel à l'aide d'un workflow et nous avons proposé un workflow générique pour la spécification du rafraîchissement d'un entrepôt. Nous avons ensuite proposé un algorithme qui, étant donné d'une part un ensemble de vues matérialisées dans l'entrepôt, d'autre part des exigences de qualité en terme de fraîcheur et de consistence et enfin des limitations d'accès aux sources de données, trouve une spécification (si elle existe) de workflow de rafraîchissement. Si la solution n'existe pas, l'algorithme propose divers assouplissements des exigences, émanant de l'utilisateur ou des administrateurs des sources qui permettent de parvenir à un compromis. Enfin, nous avons montré qu'il était intéressant d'implanter un workflow de rafraîchissement

au moyen de règles actives. Ceci nous a conduit à concevoir et implanter un système d'aide à la génération de systèmes actifs offrant un modèle d'exécution et des algorithmes taillés sur mesure pour les besoins d'une application. Ce système se présente comme une boîte à outils logiquement décomposée en quatre unités fonctionnelles qui ne varient pas d'une application à l'autre. Cependant, les modes de communication entre ces unités ainsi que leurs structures de données partagées peuvent changer. Au niveau physique, chaque unité consiste en une hiérarchie de classes qui peuvent être réutilisées ou raffinées. Une première version de ce système écrit en Java nous a permis de construire un médiateur de rafraîchissement qui a été démontrée lors de la revue du projet Esprit LTR DWQ sur une application d'entrepôt de données pour les télécommunications.

4.4 Diffusion d'information par notification

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *Le problème général abordé dans ce thème est la fédération de sources d'information selon une approche de type "Publication-Souscription", ce qui signifie que certaines sources d'information peuvent publier et diffuser un ensemble d'informations, tandis que d'autres (parfois les mêmes) peuvent émettre des requêtes de souscription pour sélectionner, parmi l'ensemble de toutes les informations diffusées, celles qui leur sont pertinentes. Cette approche est appropriée lorsque les informations échangées sont dynamiques ou ont des durées de validité très courtes, ou lorsque les différentes sources d'information veulent protéger et contrôler l'accès à leurs données. Mettre en oeuvre une telle approche pose plusieurs problèmes : comment organiser la diffusion d'informations entre éditeurs et souscripteurs? dans quel langage les souscripteurs peuvent ils définir les informations qui les intéressent? comment garantir de bonnes performances en présence d'un grand nombre de souscripteurs, de publieurs et d'une grande quantité d'information échangée? Les travaux en cours autour de ce thème se développent actuellement sous la forme de deux actions de recherche. La première activité est le développement d'un système de diffusion d'information appelé Le Subscribe. Son architecture est basée sur l'utilisation de médiateurs de diffusion, qui permettent de diffuser l'information de publieurs vers des souscripteurs sous forme de notifications (ou évènements). La deuxième activité consiste à enrichir les qualités de détection de ces médiateurs, en proposant un mécanisme de détection d'évènements de seuils. Ceux-ci permettent à un souscripteur d'être notifié, quand une valeur associée à une certaine information dépasse un seuil donné. Dans ces deux activités, une attention particulière est portée sur la mise au point d'algorithmes garantissant de bonnes performances.*

4.4.1 Le Subscribe

Participants : Françoise Fabret, François Llirbat, João Pereira, Dennis Shasha.

De nombreuses applications exigent la mise en place de systèmes de médiation permettant à divers acteurs indépendants d'échanger des informations évoluant rapidement. Sur Internet, par exemple, les applications de commerce électronique (bourses d'échange, petites annonces, lancement de nouveaux produits, promotions ...) sont caractérisées par un grand nombre de participants éditeurs ou souscripteurs : les premiers désirant publier des informations (par exemple annoncer des promotions sur certains produits) et les seconds voulant être avertis de certaines de ces informations (par exemple des promotions de plus de 50composé d'un grand nombre d'entités autonomes ; Chacune d'entre elles est en charge d'une partie de l'activité de l'entreprise et gère ses propres données. Pour assurer la cohérence de l'activité globale, ces entités ont besoin de partager une partie de leurs données. Un moyen d'organiser ce partage, sans compromettre l'autonomie de chaque entité, consiste à utiliser un système de médiation permettant à chacune de publier des informations sur ses propres activités et d'être avertie des activités des autres. Pour l'ensemble de ces applications, le système de médiation doit avoir les qualités suivantes : (1) il doit pouvoir supporter un grand nombre de publieurs, de souscripteurs et une grande quantité d'information échangées, (2) il doit pouvoir supporter des changements fréquents dans la définition des participants et des informations échangées, (3) il doit donner au souscripteur un langage lui permettant de décrire précisément les informations qui l'intéressent, (4) il doit fournir des algorithmes de filtrage efficace de l'information publiée de façon à ce que chaque souscripteur ne soit averti que des informations qui lui sont pertinentes, (5) il doit garantir pour certaines applications plus temps réel (applications boursières) des délais de notification très courts. L'objectif de Le Subscribe est de fournir un système répondant à ces problèmes. Ce système propose à des utilisateurs-éditeurs de lui notifier leurs nouvelles informations. Il est ensuite capable de diffuser ces notifications - après les avoir filtrées - auprès des utilisateurs-souscripteurs intéressés.

4.4.2 Détection d'évènements de seuil

Participants : Françoise Fabret, Dennis Shasha.

Le mécanisme des fonctions de seuil est utilisé dans le cadre d'applications réagissant à l'évolution de données dynamiques. Parmi ces applications, nous citerons la gestion de risques dans le contexte de l'environnement (inondation, incendie, ...), les activités boursières (achat et vente d'actions en bourse), la gestion d'objets mobiles (par exemple la gestion des taxis), ou encore la surveillance des réseaux. Pour toutes ces applications, des données, évoluant continuellement, sont produites par des sources autonomes (par exemple des capteurs). Une fonction dite de seuil calcule, à partir de ces données, un indicateur (par exemple un "facteur de risque"). Une alerte (ou évènement de seuil) est déclenchée, lorsque la valeur de cet indicateur franchit un seuil donné. Nous nous intéressons au problème de la détection efficace des événements de seuil et proposons une architecture de type publieur/souscripteur. Les sources sont des publieurs et le détecteur de seuil est un souscripteur/publieur : il s'abonne aux informations publiées par les sources et il notifie les événements de seuil aux applications utilisatrices. Notre recherche se concentre sur les problèmes algorithmiques liés à l'efficacité de la détection. Dans une approche naïve, le détecteur demande d'être averti de toutes les informations produites par les sources (par exemple de chacune des mesures produites par chaque capteur) dès que

ces informations sont disponibles. Nous avons montré que, pour une large classe de fonctions de seuil, il n'est pas indispensable que toutes ces informations soient transmises au détecteur pour que la détection puisse avoir lieu. Nous avons caractérisé cette classe de fonctions et nous proposons des algorithmes de filtrage de l'information utile. Dans cette approche, le détecteur souscrit auprès de chaque source pour n'être averti que lorsqu'un seuil local est franchi et, à tout moment, il peut modifier une (ou plusieurs) de ses souscriptions, ou interroger une source sur l'évolution de la donnée, dont elle est responsable. Nous avons établi plusieurs métriques permettant de mesurer l'efficacité de la détection et avons mis en évidence que, dans tous les cas, l'efficacité de la détection dépend essentiellement du choix des seuils locaux. Actuellement nous évaluons diverses heuristiques utilisées pour choisir ces seuils locaux. En particulier, nous évaluons dans quelle mesure l'utilisation de connaissances sémantiques sur les données produites par les sources peut améliorer la qualité de ces heuristiques.

4.5 Navigation dans un réseau d'informations

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *Dans ce thème nous abordons le problème de la présentation de données à des utilisateurs "naïfs", ce qui sous-entend que les utilisateurs ne sont pas capables d'exprimer des requêtes dans un langage de bases de données tel que SQL, OQL ou XML-QL. Les sites Web sont des instruments appropriés pour cela, car ils proposent un mode conversationnel très simple, basé sur la navigation. Mais l'accès navigationnel à des bases de données par le web pose des problèmes de performances. Le thème comporte deux actions. La première action vise à développer un système qui facilite la gestion du contenu et de la structure de sites Web tout en garantissant de bonnes performances d'accès grâce à l'utilisation de techniques d'anté-mémorisation. La deuxième action vise à développer un système qui offre une alternative à la présentation hiérarchique d'informations – telle qu'on la rencontre généralement dans les catalogues électroniques sur le Web.*

4.5.1 Gestion de sites Web à partir de bases de données

Participants : Daniela Florescu, Khaled Yagoub.

Récemment, l'idée de définir la structure et le contenu d'un site Web de façon déclarative, au moyen d'un programme logique, plutôt que par des programmes impératifs difficiles à analyser et optimiser, s'est imposée. Cela revient à définir le site comme une vue des données, souvent distribuées et hétérogènes, à partir desquelles le site est construit. Un problème essentiel de cette approche, d'un point de vue performances, est de déterminer le niveau de matérialisation à choisir pour la vue incarnée par le site Web. Une solution est de matérialiser les pages web statiquement, avant que les utilisateurs ne commencent à naviguer dans le site. Une autre approche est de matérialiser les pages Web dynamiquement, lorsqu'elles sont demandées par les utilisateurs. Nous proposons une approche hybride, dans laquelle certains précalculs sont effectués et stockés. Le problème est de choisir une configuration optimale à matérialiser, c'est-à-dire un ensemble de vues et de fonctions que l'on doit anté-mémoriser. Comme le choix

automatique d'une configuration optimale parmi toutes les configurations possibles est trop complexe, nous proposons des heuristiques efficaces, qui permettent de faire le compromis entre la taille des données précalculées, la contrainte sur la fraîcheur des données et le temps nécessaire pour générer les pages Web.

Un autre problème que nous adressons est la vérification des contraintes d'intégrité d'un site Web. Cela représente une tâche extrêmement importante dans le processus de création des sites Web et il est hautement souhaitable de pouvoir l'automatiser. C'est un des grands avantages de l'utilisation d'une définition déclarative que de pouvoir analyser automatiquement cette définition, dans le but de vérification de contraintes d'intégrité. Dans ce cadre, le problème peut être formulé comme suit : "Etant donnés (1) la définition déclarative d'un site, (2) les contraintes d'intégrité caracterisant les données de base et (3) un ensemble de contraintes d'intégrité que la structure et le contenu du site doivent satisfaire, il faut tester automatiquement si le site Web satisfait bien ces contraintes d'integrité, en raisonnant seulement sur la définition intensionnelle du site (c.-a-d. sans avoir besoin de construire de façon extensionnelle le site)." La solution proposée à ce problème passe par : (1) l'utilisation d'un modèle déclaratif de site Web, (2) la définition d'un langage de contraintes permettant de définir des contraintes tant sur la base de données que sur la structure et le contenu du site, (3) des algorithmes de vérification automatique permettant de tester si oui ou non un site donné vérifie les contraintes requises. De plus, notre technique fournit des outils d'aide à la création de site en analysant les raisons d'échec lors de la vérification des contraintes et en proposant des solutions alternatives.

4.5.2 Navigation adaptative dans des collections de données

Participants : Eric Simon, Alberto Lerner.

Dans la plupart des sites Web servant d'accès à de grandes bases de données (par exemple, les sites de commerce électronique), les données sont présentées aux utilisateurs de manière hiérarchique. Typiquement, l'utilisateur accède à un niveau d'une hiérarchie de données (par exemple, une catégorie de produits dans un catalogue) puis navigue dans cette hiérarchie en suivant des liens. Ce principe est efficace lorsque les utilisateurs recherchent leurs données en suivant la même logique de classification hiérarchique que celle qui est utilisée par les données. Dans le cas contraire, il est possible de definir une hiérarchie multiple qui prend en compte différents modes de navigation dans les données. Cependant, ceci peut rapidement augmenter la complexité de la structure du site Web jusqu'à rendre la spécification de cette structure extrèmement laborieuse et difficile à maintenir.

Pour repondre à ce problème, nous avons développé en collaboration avec Dennis Shasha (NYU, USA) un système qui offre un mode de navigation non hiérarchique dans des collections de données. Ce système, appelé Attman, utilise un modèle de données original composé de tables relationnelles et de cinq dépendances qui expriment des liens sémantiques entre les données. Chaque dépendance définit une relation de pertinence entre les données : les données d'une table sont pertinentes si elles satisfont les relations de pertinence auxquelles elles participent. L'utilisateur qui se connecte au système voit une liste de tables dont le contenu est consultable. Puis l'utilisateur peut sélectionner des lignes pertinentes dans une table. Le système calcule alors toutes les données qui demeurent pertinentes au regard de cette sélec-

tion en utilisant les dépendances définies par le concepteur de l'application. La liste des tables pertinentes restante est ensuite présentée à l'utilisateur. Chacune de ces tables ne contient que des lignes pertinentes. L'utilisateur peut alors continuer sa navigation. En dehors du modèle de données, nous avons développé des algorithmes efficaces qui permettent d'effectuer le calcul des données pertinentes. Le système possède une architecture à trois tiers qui permet à une application cliente de se connecter à un serveur Attman qui puise ses données dans un serveur de données. Le système est opérationnel et fait l'objet de diverses expérimentations pour des applications de gestion de catalogues électroniques.

5 Actions régionales, nationales et internationales

5.1 Actions régionales

A l'INRIA, nous entretenons une collaboration étroite avec le projet VERSO, notamment sur les problèmes d'accès à des bases de données hétérogènes. Nous coopérons aussi avec le projet AIR dans le domaine des systèmes d'information pour l'environnement, notamment pour les contrats européens THETIS et DECAIR. Enfin, nous collaborons avec le laboratoire PrIsm de l'Université de Versailles sur les problèmes d'accès à l'information distribuée et le raffraîchissement des entrepôts de données.

5.2 Actions européennes

5.2.1 Esprit R&D Miro-Web

L'équipe participe au projet Esprit Miro-Web qui a commencé en novembre pour 2 ans. Les partenaires sont Bull, Ibermatica, Osis et le GMD. L'Inria est partenaire associé de Bull. L'objectif est de construire un système d'accès à des sources de données hétérogènes sur le Web ainsi que les outils associés et d'expérimenter le système sur deux applications pilotes : un système d'information hospitalier (pays Basque) et une application du tourisme (Allemagne). Dans ce projet, l'INRIA et Bull adaptent les technologies de médiateur et d'adaptateur (projet Disco) et participent à l'expérimentation.

5.2.2 Esprit LTR DWQ

L'équipe participe au projet Esprit Long Term Research DWQ (Foundations of Data Warehouse Quality) lancé en octobre 1996 pour une durée de trois ans. Ce projet est coordonné par l'université d'Athènes (Y. Vassiliou) avec pour partenaires les universités de Rome (M. Lenzerini), Aachen (M. Jarke), Saarbruck (W. Nutt) et l'IRST de Trento (E. Franconi). L'objectif général du projet est de développer des techniques et des outils permettant une conception et une mise en œuvre rigoureuse d'entrepôts de données, basées sur des facteurs de qualité de données bien définis. Les outils seront prototypés et validés en coopération avec de grands utilisateurs européens. La contribution de l'équipe CARAVEL est d'apporter des solutions à deux problèmes clés des entrepôts de données : le maintien à jour efficace et fiable des données d'un entrepôt en fonction des changements des sources de données et la conception du schéma de la base de données de l'entrepôt, de façon à optimiser à la fois l'exécution des requêtes

décisionnelles et le maintien à jour de l'entrepôt. Ce travail s'appuie sur l'expérience que nous avons déjà acquise sur les bases de données actives et les stratégies de matérialisation de vues.

5.2.3 Telematics THETIS

L'équipe participe au projet Telematics THETIS qui a commencé en Avril 1998 pour une durée de 2 ans et demi. L'objectif est de construire un système d'accès à des sources de données environnementales hétérogènes sur le Web, afin de faciliter la gestion des zones côtières de la mer Méditerrannée, pour des utilisateurs comme l'IFREMER en France, HR Wallingford en Angleterre ou l'IMBC en Crète. L'originalité du système est de combiner des techniques d'indexation spécialement conçues pour des données océanographiques de type image ou numérique avec des techniques de médiation de requêtes bases de données. L'indexation permet une recherche très large mais grossière de sources d'information, tandis que les requêtes bases de données permettent l'interrogation fine de sources de données à partir de la connaissance de leur structure. Dans ce projet à forte composante utilisateur, l'INRIA développe un nouveau système de recherche d'informations distribuées hétérogènes (Le Select).

5.2.4 Environnement et climat DECAIR

L' objectif du projet DECAIR est de fournir des données de meilleure qualité aux organismes en charge de la prévision de la pollution urbaine. En particulier le projet se concentre sur la qualité des données fournies comme données d'entrée aux modèles de pollution de l'air. Ces données sont de différents types: données géograhiques, données d'occupation des sols, données météorologiques, données d'émission de polluants, etc. Pour atteindre cet objectif des efforts de recherche sont prodigués dans deux directions complémentaires : D'abord le projet explore la possibilité d'utiliser des données satellites pour améliorer la précision et la fraîcheur des données d'entrées. L'objectif est ici de fournir des méthodes et des algorithmes de traitement d'images satellites qui sont adaptés au problème de la pollution de l'air. De plus le projet étudie la mise au point d'un système d'information adapté capable d'accéder, traiter, transformer et intégrer des données provenant de plusieurs sources distantes comme les satellites, les stations aux sols, des bases de données. Ce système a en charge de maintenir automatiquement la fraîcheur et la qualité des données utilisées par les modèles.

Pour valider cette approche, nous construisons un prototype appelé " démonstrateur DECAIR" capable de gérer l'exécution de la chaîne de traitement, de l'acquisition des images satellitaires jusqu'à la présentation des paramètres d'entrée aux modèles de qualité de l'air. Ce prototype sera testé avec deux modèles de qualité de l'air, l'un mesurant la qualité de l'air sur Madrid, l'autre sur Berlin. L'architecture du prototype devra être suffisament flexible pour permettre, dans des developpements futurs, d'élargir l'ensemble des données d'entrée qui peuvent être accédées automatiquement, d' intégrer et utiliser facilement de nouveaux modèles, de faciliter l'application de ces modéles à de nouveaux sites, de détecter et prendre en compte les changements météorologiques rapides en cours de l'exécution des modèles.

Le projet DECAIR est un projet européen et pluri-disciplinaire. Il est coordonné par l'ERCIM. Il implique des équipes de recherches specialisées dans la modélisation de la qualité de l'air: GMD de Berlin, et UPM à Madrid, des équipes specialisées dans les systèmes d'infor-

mation pour l'environnement: CLRC-RAL en Angleterre, FORTH-ICS en grèce et le projet CARAVEL de l'INRIA, des équipes spécialistes de l'analyse d'image :Le projet AIR de l'INRIA et des partenaires industriels: BULL en France et SICE en Espagne.

5.3 Actions internationales

5.3.1 Europe

- université NTUA d'Athènes, Grèce (Timos Sellis, Yannis Vassiliou), avec qui nous collaborons dans le projet LTR DWQ.
- université d'Aachen, Allemagne (Matthias Jarke), avec qui nous collaborons dans le projet LTR DWQ.
- université de Rome (Maurizio Lenzerini), avec qui nous collaborons dans le projet LTR DWQ.
- FORTH (Christos Nikolaou), avec qui nous collaborons dans le projet Thetis.

5.3.2 Amérique du Nord

- université de Maryland (Michael Franklin et Louiqa Raschid).
- IBM, Almaden, Californie (Mike Carey, Chandra Mohan).
- AT&T Research, New Jersey (Dan Suciu).
- Bell Labs, New Jersey (Narain Gehani, Rick Hull); F. Llirbat a rendu visite pendant 1 mois et demi à cette équipe durant l'été et E. Simon a rendu visite à cette équipe en octobre.
- NYU, New York (Dennis Shasha) ; E. Simon a travaillé avec cette équipe durant 1 mois.
- université de Washington (Alon Levy).
- université d'Alberta, Edmonton, Canada (Tamer Özsu).

5.3.3 Amérique du Sud et Amérique Centrale

- universités de Rio de Janeiro (PUC, UFRJ et UNI-Rio), avec lesquelles nous avons un projet de coopération CNPQ-Inria sur les systèmes d'information pour l'environnement. Nous avons organisé un workshop sur ce thème à l'Inria en novembre 1999.

6 Diffusion de résultats

6.1 Animation de la Communauté scientifique

Daniela Florescu a co-organisé le Workshop on Query Processing for Semistructured Data, Jerusalem, Israel, janvier 1999 et le Séminaire à Dagsthul sur le thème "Data Models for Data

Integration", juillet 1999. Elle représenete également l'INRIA dans le groupe de travail W3C sur la définition du langage de requêtes standard pour XML. Patrick Valduriez a présidé le comité de programme européen de la conférence internationale VLDB'99. Eric Simon a co-présidé le comité de programme industriel de la conférence ACM SIGMOD'99. Mokrane Bouzeghoub a co-présidé le comité de programme de la conféerence internationale ER'99.

L'équipe a participé aux comités de programme des colloques suivants:

- Int. Conf. of the Eighth World Wide Web Conference (WWW'8): D. Florescu
- Int. Conf. on Very Large Databases (VLDB): D. Florescu, P. Valduriez
- Int. ACM SIGMOD Conf: E. Simon
- Int. Conf. on Data Base Theory (ICDT): E. Simon
- Int. Conf. on Data Engineering (ICDE): D. Florescu, E. Simon
- Conf. Nationale de Bases de Données Avancées (BDA): D. Florescu
- Int. Conf. Cooperative Information Systems (COOPIS): D. Florescu
- Workshop on Query Processing for Semistructured Data: D. Florescu
- Workshop on Intelligent Information Integration: D. Florescu
- Workshop on Web and Databases (WebDb): D. Florescu

L'équipe contribue aussi à des comités de lecture et associations :

- Int. Journal on Intelligent and Cooperative Database Systems, World Scientific (P. Valduriez).
- Int. Journal on Distributed and Parallel Database Systems, Kluwer Academic Publishers (E. Simon, P. Valduriez).
- VLDB Journal (P. Valduriez).
- VLDB Endowment (P. Valduriez).
- Journal of Data and Knowledge Engineering, North Holland (G. Gardarin).
- Network and Information Systems Journal, Hermes (M. Bouzeghoub, rédacteur en chef ; G. Gardarin, E. Simon, P. Valduriez).

6.2 Enseignement

- Bases de données réparties, magistère de l'ESSEC (P. Valduriez). Bases de données actives, ENST Bretagne (3ème année), 12 heures (E. Simon).
- Bases de données, algorithmique et programmation, CNAM de Versailles, 3ème année (F. Fabret).
- Bases de données à objets, Université Paris Sud (MIAGE 2ème année), 6 heures (D. Florescu, I. Manolescu).

7 Bibliographie

Livres et monographies

[1] M. Bouzeghoub, F. Fabret, H. Galhardas, M. Matulovic, J. a. Pereira, E. Simon, *Fundamentals of Data Warehousing*, Springer-Verlag, 1999.

[2] J. A. Marques, J. a. Pereira, A. R. Silva, *Pattern Languages of Program Design 4*, Addison-Wesley, 1999.

Articles et chapitres de livre

[3] L. Bouganim, D. Florescu, P. Valduriez, «Parallel Query Execution in NUMA Multiprocessors», *DAPD 7*, 1, 1999.

[4] D. Florescu, A. Deutsch, M. Fernandez, A. Levy, D. Maier, D. Suciu, «Querying XML data», *DEBull 22*, 3, 1999, p. 27–34.

[5] D. Florescu, D. Kossmann, «Storing and Querying XML Data Using an RDBMS», *DEBull 22*, 3, 1999.

[6] E. Pacitti, E. Simon, «Update Propagation Strategies to Improve Data Freshness in Lazy Master Schemes», *VLDB Journal*, 1999.

Communications à des congrès, colloques, etc.

[7] M. Bouzeghoub, F. Fabret, M. Matulovic, «Modeling the Data Warehouse Refreshment Process as a Workflow Application.», *in : DMDW*, June 1999.

[8] M. Bouzeghoub, C. Quix, P. Vassiliadis, «Towards quality-oriented data warehouse usage and evolution», *in : CAiSE*, June 1999.

[9] M. Bouzeghoub, D. Theodoratos, «Data Currency Quality Factors in Data Warehouse Design.», *in : DMDW*, June 1999.

[10] S. Carvalho, A. Lerner, S. Lifschitz, «An Object-Oriented Framework for the Parallel Join Operation», *in : DEXA-WS*, IEEE Computer Society, p. 34–38, September 1999.

[11] S. Carvalho, M. Vianna e Silva, R. Melo, F. Porto, «Persistent Object Synchronization with Active Relational Databases», *in : TOOLSUSA*, IEEE, August 1999.

[12] D. Florescu, A. Deutsch, M. Fernandez, A. Levy, D. Suciu, «XML-QL: A Query Language for XML», *in : WWW*, 1999.

[13] D. Florescu, M. Fernandez, A. Levy, D. Suciu, «Verifying Integrity Constraints on Web Sites», *in : IJCAI*, 1999.

[14] D. Florescu, M. Friedman, I. Zachary, A. Levy, D. Weld, «An Adaptive Query Execution Engine for Data Integration», *in : SIGMOD*, 1999.

[15] D. Florescu, A. Levy, D. Suciu, K. Yagoub, «Optimization of Run-Time Management of Data Intensive Web Sites», *in : VLDB*, 1999.

[16] D. Florescu, A. Levy, D. Suciu, K. Yagoub, «Run-Time Management of Data Intensive Web Sites», *in : WebDB99*, 1999.

[17] D. Florescu, I. Manolescu, A. Levy, D. Suciu, «Query Optimization in the Presence of Limited Access Patterns», *in : SIGMOD*, 1999.

[18] R. Hull, F. Llirbat, E. Simon, J. Su, B. Kumar, G. Dong, G. Zhou, «Declarative Workflows that Support Easy Modification and Dynamic Browsing», p. 69–78, 1999.

[19] R. Hull, F. Llirbat, J. Su, G. Dong, B. Kumar, G. Zhou, «Efficient Support for Decisions Flows in E-commerce Applications», *in : International Conference on Telecommunications and Electronic Commerce , Nashville*, November 1999.

[20] E. Pacitti, P. Minet, E. Simon, «Fast Algorithms for Maintaining Replica Consistency in Lazy Master Replicated Databases», *in : VLDB*, September 1999. A long version of this paper is available as an INRIA technical report.

Rapports de recherche et publications internes

[21] M.-J. Blin, F. Fabret, «A Coordination Mechanism to Prevent the Violation of Distributed Constraints», *rapport de recherche*, RR, June 1999.

[22] L. Bouganim, F. Fabret, C. Mohan, P. Valduriez, «Dynamic Scheduling of Complex Distributed Queries», *rapport de recherche n° 3677*, RR, April 1999.

[23] D. Florescu, H. Galhardas, D. Shasha, E. Simon, «An Extensible Framework for Data Cleaning (extended version)», *rapport de recherche n° 3742*, RR, 1999.

[24] D. Florescu, D. Kossmann, «A Performance Evaluation of Alternative Mapping Schemes for Storing XML Data in a Relational Database», *rapport de recherche n° 3680*, RR, Mai 1999.

[25] D. Florescu, A. Levy, D. Suciu, K. Yagoub, «Run-time Management of Data Intensive Web Sites», *rapport de recherche n° 3684*, RR, March 1999.

[26] D. Florescu, I. Manolescu, A. Levy, D. Suciu, «Query Optimization in the Presence of Limited Access Patterns», *rapport de recherche*, RR, 1999.

[27] O. Kapitskaia, «Traitement de requêtes dans les systèmes d'intégration de sources de données distribuées», *rapport de recherche*, Thèse de doctorat, Novembre 1999.

[28] M. Matulovic-Broqué, «Aide à la conception et à l'implémentation d'un mécanisme d'exécution des règles actives», *rapport de recherche*, Thèse de doctorat, Mai 1999.

[29] H. Naacke, «Modèles de coût pour médiateurs de bases de données», *rapport de recherche*, Thèse de doctorat, Novembre 1999.

[30] E. Pacitti, P. Minet, E. Simon, «Fast Algorithms for Maintaining Replica Consistency in Lazy Master Replicated Databases», *rapport de recherche n° 3654*, RR, April 1999.