INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet CARAVEL

Systèmes de médiation d'information

Rocquencourt

THÈME 3A

Table des matières

1 Composition de l'équipe

Responsable scientifique

Eric Simon [DR Inria]

Responsable permanent

Francois Llirbat [CR]

Assistante de projet

Elisabeth Baqué [AI]

Personnel Inria

Daniela Florescu [CR, jusqu'au 1er novembre]

Patrick Valduriez [DR, jusqu'au 1er septembre]

Collaborateurs extérieurs

Luc Bouganim [MC, université de Versailles]

Mokrane Bouzeghoub [professeur, université de Versailles]

Thierry Fuhs [CR, CEMAGREF]

Chercheur invité

Kenneth Ross [Columbia University, New York, USA, 5 mois]

Chercheur post-doctorant

Arno Jacobsen

Ingénieurs experts

Mokrane Amzal

Francoise Fabret

Florian Xhumari [jusqu'au 1er avril]

Doctorants

Maria-Claudia Cavalcanti [Université de Rio de Janeiro, 3 mois]

Helena Galhardas [université de Lisbonne]

Alberto Lerner [université de Rio de Janeiro]

Ioana Manolescu [université de Versailles]

Fabio Porto [université de Rio de Janeiro, jusqu'au 1er octobre]

Joao Pereira [université de Lisbonne]

Khaled Yagoub [boursier MESR, université de Versailles]

Stagiaires

Jean-Pierre Matsumoto [Université Paris VI]

Cezar-Cristian Andrei [Ecole Polytechnique de Bucarest]

Aurelian Lavric [Ecole Polytechnique de Bucarest]

Dan-Alexandru Olteanu [Ecole Polytechnique de Bucarest]

Radu Preotiuc [Ecole Polytechnique de Bucarest]

Cristian-Augustin Saita [Ecole Polytechnique de Bucarest]

2 Présentation et objectifs généraux

L'énorme quantité d'informations aujourd'hui disponible sur le Web et la très grande disparité de ces informations aussi bien dans leur contenu que dans leur mode d'accès, rendent bien souvent laborieuse la recherche de données précises. Chacun aimerait avoir accès à une vue intégrée et à jour de ces informations, ce qui suppose à la fois une structuration uniforme et cohérente des données et des modes d'interrogation adaptés aux besoins de l'utilisateur. Le projet Caravel répond à ce problème fondamental d'intégration de sources d'informations au travers de trois grands thèmes de recherche complémentaires:

- Thème 1: il s'agit de faciliter la publication de ressources dans un réseau ainsi que l'accès à ces ressources au moyen de langages de haut niveau. Les ressources peuvent être des données (structurées ou non) ou des services (librairies, programmes scientifiques, sites Web, etc), l'ensemble formant un *système d'information global*. Deux difficultés majeures se posent: réduire considérablement l'effort de développement nécessaire à la publication de ressources et mettre au point des méthodes d'optimisation pour les langages de haut niveau proposés.

- Thème 2: il s'agit de faciliter la production de données élaborées à partir de données et de services publiés dans le système d'information global. Les principales difficultés à résoudre sont d'intégrer des données hétérogènes de façon cohérente et correcte, d'assembler judicieusement des programmes disparates dans une chaîne de traitement de données et enfin d'exécuter efficacement de telles chaînes de traitement.

- Thème 3: il s'agit de faciliter la création et la maintenance de sites Web utilisant les ressources du système d'information global. Les difficultés sont d'administrer la structure logique d'un site tout en garantissant de bonnes performances de consultation et d'offrir des méthodes de navigation adaptatives en fonction de l'intérêt de l'utilisateur.

Plusieurs actions de recherche sont menées dans chacun de ces thèmes. Deux grandes actions structurent le premier thème. La première s'inscrit dans une approche de type "pull" pour l'accès aux ressources du système d'information global, tandis que la seconde se situe dans une approche de type "push". Le second thème distingue également deux actions de recherche qui visent à aider d'une part à la génération de programmes efficaces de nettoyage de données hétérogènes ("data cleaning", en anglais) et d'autre part à l'assemblage et l'exécution de workflow scientifiques distribués. Enfin, le dernier thème recouvre deux actions qui visent à faciliter l'administration de sites Web performants et à offrir des moyens de navigation adaptatifs à l'utilisateur.

Les techniques conçues dans ces actions de recherche prennent la forme de langages de bases de données, de modèles de données ou d'algorithmes. Ces techniques sont implantées dans des composants logiciels modulaires qui s'interfacent entre des applications clientes et des serveurs d'information selon un modèle d'architecture à trois-tiers. D'un point de vue stratégique, nous concevons des composants logiciels 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. De plus, nous nous efforçons d'expérimenter nos composants dans le cadre d'applications réelles en collaboration avec des partenaires utilisateur via des contrats industriels.

3 Fondements scientifiques

L'histoire de la recherche en bases de données est exceptionnelle par sa productivité, son transfert industriel et son impact économique. Reconnue depuis un peu plus de 20 ans comme une discipline de recherche de base par les Etats-Unis suivis par la plupart des pays industrialisés, la recherche en bases de données a été conduite d'abord dans les laboratoires des grands groupes industriels pour être généralisée ensuite aux laboratoires publics et universités. Les Systèmes de Gestion de Bases de Données (SGBD) sont aujourd'hui des logiciels de base essentiels dans tout système d'information. Intuitivement, un SGBD permet à des utilisateurs de poser avec une certaine souplesse des requêtes pour manipuler (rechercher et modifier) une grande masse de données persistantes. Il doit contrôler la concurrence de ces accès tout en garantissant la cohérence, l'intégrité, la confidentialité et la sécurité des données.

Depuis l'apparition vers la fin des années 60 des premiers SGBD hérités des systèmes de gestion de fichiers, d'importants résultats théoriques et pratiques ont ponctué l'histoire des

bases de données. L'invention du *modèle relationnel* en 1970 est l'événement le plus marquant (il a valu à son auteur, Tedd Codd, le prix Turing de l'ACM en 1982). Le modèle relationnel a permis d'établir les fondements mathématiques qui manquaient au domaine et a ouvert de grandes perspectives de recherche, notamment en conception de schémas normalisés et en langages de requêtes déclaratifs. Les premières retombées de ces recherches ont été de faciliter l'administration et la manipulation de bases de données et d'accroître la productivité des utilisateurs.

Cependant, la puissance des langages relationnels qui permettent d'exprimer des requêtes complexes a longtemps posé des problèmes de performances. Ceux-ci ont été progressivement résolus par des efforts continus, durant plus de quinze ans, en recherche et développement, avec en particulier des algorithmes efficaces pour traiter les opérateurs relationnels, des techniques d'optimisation de requêtes, le support intégré efficace des transactions et l'exploitation du parallélisme pour exécuter les opérateurs relationnels sur calculateur multiprocesseur. Ces deux derniers points ont valu a Jim Gray le prix Turing de l'ACM en 1998.

Depuis 1981, les projets Sabre puis Rodin ont participé activement à ce mouvement de la recherche en concevant et en expérimentant des techniques afin d'améliorer les fonctionnalités et les performances des SGBD. Ces techniques ont pris la forme de langages et de modèles à base de règles et d'objets qui étendent la puissance d'expression des modèles de bases de données existant, d'algorithmes d'optimisation de langages de bases de données ainsi que d'algorithmes et de structures de données pour l'exécution d'opérations coûteuses de bases de données et pour l'exécution concurrente de transactions. Diverses collaborations avec des industriels (surtout via des contrats européens) nous ont permis d'évaluer nos solutions dans des systèmes complets impossibles à développer dans le contexte d'un projet Inria (e.g., évaluation de nos algorithmes d'optimisation de requêtes pour bases de données parallèles sur le système DBS3 de Bull sur machine KSR, ou évaluation d'un protocole de contrôle de concurrence dans le système Validity developpé par la société NCM). Enfin, ces collaborations ont donné lieu à des transferts industriels de logiciels (e.g., Omnis, Disco) ou de solutions integrées à des produits (e.g., O2 Engine, Java Universal Binding).

Avec la création du projet Caravel, nous avons redéfini notre problématique de recherche autour de l'intégration d'information dans un réseau composé de sources d'information hétérogènes et autonomes. Deux raisons principales fondent nos décisions. La première est l'évolution des applications de base de données résultant des progrès technologiques, de l'explosion du Web et de l'internet, ainsi que de l'importance croissante des applications d'aide à la décision. La deuxième raison est le degré de maturité auquel sont parvenus les SGBD commercialisés. Ce dernier point a deux conséquences: d'une part certains problèmes sont maintenant considérés comme résolus et d'autre part, les industriels sont souvent les mieux placés pour continuer à améliorer les performances et les fonctionalités des noyaux de SGBD. Nos recherches actuelles s'appuient considérablement sur notre expérience en conception de langages et de modèles de bases de données ainsi qu'en algorithmes d'exécution distribuée d'opérations de bases de données et d'optimisation de langages.

4 Domaines d'applications

La stratégie du projet Caravel repose sur des collaborations avec des partenaires utilisateurs dans des contrats de recherche à finalité applicative. Cette stratégie nous permet de mieux comprendre les besoins d'applications complexes dans le domaine de l'intégration d'information et d'identifier des problèmes de recherche nouveaux (e.g., modèle de workflow scientifiques, modèle et algorithmes pour le nettoyage de données, middleware pour les applications scientifiques). De plus, la collaboration avec des utilisateurs nous offre les moyens d'expérimenter nos solutions dans des contextes d'utilisation réelle.

Jusqu'à présent, le projet s'est surtout intéressé aux systèmes d'information pour l'environnement car c'est un domaine d'application très riche en problèmes d'intégration d'information: les problèmes d'intégration se posent à grande échelle, les sources d'information ont une forte autonomie due à la pluridisciplinarité du domaine et l'intégration d'information est nécessaire aux nombreuses applications d'aide à la décision. Le choix d'un domaine d'application nous a permis d'accumuler depuis cinq ans une expertise reconnue, ce qui facilite le développement de nouvelles collaborations et l'approfondissement des problèmes de recherche qui nous concernent. Les applications principales sur lesquelles nous travaillons sont la gestion de ressources naturelles en zones côtières (projet européen Thetis), la prédiction de la qualité de l'air en milieu urbain (projet européen Decair) et l'analyse des phénomènes de bio-corrosion sur les plates-formes pétrolières (projet Ecobase).

D'autres applications environnementales sont en cours d'exploration comme la gestion de risques liés à des phénomènes naturels (par exemple, inondations) ou à des accidents (par exemple, marées noires). Mais depuis cette année, nous examinons aussi des applications dans le domaine de la santé qui présentent des caractéristiques semblables aux applications que nous avons déjà étudiées: dossier électronique du patient, base de données génétiques universelle (avec le Centre National de Génotypage), gestion de données en neuroimagerie (projet d'action de recherche coopérative).

5 Logiciels

Cette année, un effort très important a été investi sur le développement de composants logiciels qui intègrent des solutions élaborées au cours des années précédentes. En résultat, cinq prototypes de composants logiciels ont été démontrés cette année au cours des deux plus importantes conférences internationales en bases de données, SIGMOD et VLDB (ces démonstrations sont sélectionnées par un comité d'évaluation). Un point important est la mise au point de méthodes de développement et l'utilisation d'outils de génie logiciel destinés à améliorer la robustesse et la pérénité de nos logiciels. Enfin, deux logiciels sont actuellement utilisés en dehors du projet: Le Select et Weave.

La figure ci-dessous donne un synopsis des composants logiciels développés dans le projet Caravel et de leurs interactions potentielles. Les logiciels Le Select/Agora et Le Subscribe sont deux réponses possibles au problème abordé dans le thème 1 du projet. Chacun de ces logiciels offre une vue uniforme et intégrée des informations disponibles dans un système global, mais à travers des modes d'accès différents: Le Select suit une approche de type "pull" (i.e., requête/réponses) et Le Subscribe suit une approche de type "push" (i.e., abonne-

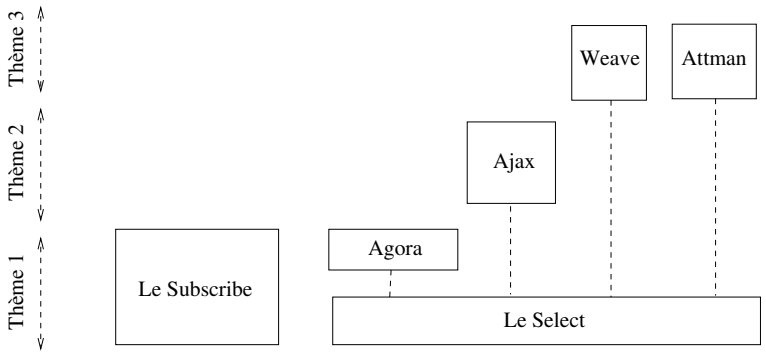

FIG. 1 – *Vue d'ensemble des logiciels*

ment/notification). Le logiciel Agora donne une représentation uniforme en XML des données tandis que Le Select en donne une vue relationnelle étendue. Le logiciel Ajax qui s'inscrit dans le thème 2 du projet, aide à la génération de programmes efficaces de nettoyage de données extraites par exemple du système d'information global via Le Select. Les données produites par Ajax peuvent à leur tour être publiées via Le Select. Les logiciels Attman et Weave s'inscrivent dans le thème 3 du projet. Ils permettent de construire des sites Web à partir de données qui seraient accédées via Le Select.

5.1 Le Select

Participants : Mokrane Amzal, Ioana Manolescu, Eric Simon [correspondant], Florian Xhumari, Aurelian Lavric.

Le Select est un nouveau système middleware développé depuis 1998 dans le cadre du projet européen Thetis pour répondre aux besoins des applications scientifiques de partager des données et des programmes. Le Select est un successeur du système Disco (développé dans le projet Rodin de 1995 à 1998 dans le cadre de l'action Dyade Médiation, puis transféré en 1999 à la société LibertyMarket qui commercialise le portail Kelkoo.com). Le Select possède une architecture distribuée "peer to peer" de type médiateur/adaptateur. L'objectif général de Le Select est de permettre à des auteurs de ressources, c'est à dire de données ou de services, de facilement publier ces ressources vers une communauté d'utilisateurs en leur donnant une vue uniforme et intégrée et enfin de permettre à des utilisateurs de manipuler cette vue uniforme à travers un langage de haut niveau. Les données ont une représentation uniforme exprimée dans le modèle de données relationnel étendu à des types de données définis par l'utilisateur. La première version de ce logiciel est diffusé depuis le mois d'Octobre. Le Select est actuellement utilisé par plusieurs universités (UNIRIO, UFRJ, IME et PUC-Rio au Brésil), centres de recherche (CNR en Italie, CEMAGREF en France, ICS-FORTH et IMBC en Grèce) et sociétés (Alcatel Industries en France, HR-Wallingford en Angleterre) pour le développement d'applications environnementales. Deux applications ont déja été démontrées dans le cadre du projet Thetis.

5.2 Agora

Participants : Daniela Florescu, Ioana Manolescu [correspondant], Florian Xhumari, Dan Olteanu.

Le système Agora offre une vue uniforme en XML des sources de données publiées à l'aide du logiciel Le Select. L'utilisateur peut manipuler cette vue uniforme des données au moyen de requêtes exprimées dans le langage Quilt. Les requêtes Quilt sont traduites en requêtes SQL exprimées sur un schéma relationnel générique. Toute donnée en format relationnel ou XML peut se décrire comme une vue (au sens base de données) exprimée sur ce schéma générique. Une étape de réecriture transforme la requête SQL exprimée sur le schéma générique en une requête exécutable par Le Select sur les sources de données concernées. Les données résultant de cette exécution sont ensuite traduites en format XML et retournées à l'utilisateur. L'intérêt majeur d'Agora est de permettre l'interrogation efficace de données relationnelles et XML dans une même requête Quilt. Ce logiciel a été démontré à la conférence VLDB'2000 avec Le Select mais ne fait pour l'instant pas l'objet de diffusion extérieure au projet.

5.3 Ajax

Participants : Daniela Florescu, Helena Galhardas [correspondant], Eric Simon, Cristian Saita.

Le logiciel AJAX est un outil d'aide à la génération de programmes efficaces pour le nettoyage de données. AJAX offre un langage de haut niveau pour la spécification de programmes de nettoyage de données. Un programme dans ce langage décrit un graphe à flôts de données dont les noeuds sont des opérations de nettoyage. AJAX propose un ensemble de cinq opérateurs logiques paramétrables qui peuvent exprimer toutes les opérations de transformation de données nécessaires. AJAX offre également un environnement de mise au point sophistiqué qui permet d'inspecter le déroulement d'un programme et d'intervenir manuellement sur le résultat des transformations, de solliciter explicitement l'assistance de l'utilisateur depuis le programme de nettoyage via la génération d'exceptions et d'assister l'utilisateur dans le deboguage d'un programme via un mécanisme d'explication des exceptions générés. AJAX genère du code Java qui optimise l'exécution des opérations logiques de transformation. Le prototype a été présenté lors de la conférence SIGMOD'2000. Il est actuellement utilisé afin de nettoyer les 2 millions de références bibliographiques en informatique du site Web Citeseer qui sont collectées automatiquement sur le Web. Les premières expériences réalisées pour 100.000 références ont montré la puissance du langage de spécification d'AJAX et l'intérêt des techniques d'optimisation mises en oeuvre.

5.4 Le Subscribe

Participants : Françoise Fabret, François Llirbat [correspondant], Joao Pereira, Arno Jacobsen, Radu Preotiuc, Ken Ross.

Le Subscribe est un système de publication/souscription ("publish/subscribe", en anglais) dont le développement a débuté en 1999. Ce système est dédié à la diffusion en mode "push"

d'informations ayant la forme d'ensembles de couples "attribut-valeur" appelés événements. Le langage de souscription supporté par le système est simple: chaque souscription consiste en une conjonction de prédicats sur les valeurs des attributs. L'objectif principal de ce système est de supporter un très grand nombre de souscriptions (plusieurs millions) et un haut débit d'événements (plusieurs centaines par seconde). Les souscripteurs et les éditeurs peuvent communiquer avec le système en utilisant le protocole Java RMI ou HTTP. Le Subscribe est composé de plusieurs composants logiciels, chacun étant responsable d'une fonctionnalité du système: filtrage des événements, notification des événements auprès des souscripteurs, etc, .. Ces composants peuvent être répartis sur plusieurs machines ou résider sur une même machine. Le système offre différents modes de notification : par e-mail, de façon immédiate en utilisant le protocole UDP ou Java RMI, ou sur demande des souscripteurs. Le point fort du système réside dans son module de filtrage qui implémente des algorithmes de pattern matching très performants. Le Subscribe a été démontré au Caire lors de la conférence VLDB'2000. Il va aussi être démontré en Janvier 2001 à New-York dans le cadre de la conférence et de l'exposition Linuxworld.

5.5 Weave

Participants : Daniela Florescu, Khaled Yagoub [correspondant], Cezar Andrei.

Le logiciel Weave est un système de gestion de sites Web à usage intensif de données. Il permet de construire des sites de façon déclarative ce qui facilite leur conception et leur mise en oeuvre et réduit le coût de leur maintenance. Weave possède une architecture configurable de caches à plusieurs niveaux permettant de cacher des données extraites d'une base de données sous forme de tables relationnelles, de fragments XML ou de pages HTML. La possibilité de cacher des fragments de données XML assure un contrôle sémantique très fin des informations cachées ce qui est très important dans des sites manipulant des données avec des droits d'utilisation limités (e.g., oeuvres d'art). Weave offre un langage de spécification déclaratif (appelé WeaveL) qui permet de définir le schéma d'un site (c'est-à dire de sa structure en pages et en hyper-liens), et la spécification de différentes stratégies de caching. Weave offre également un environnment complet de suivi des performances d'un site Web via la génération de statistiques sur l'utilisation du site. Weave a été démonstré dans plusieurs conférences internationales (EDBT'00, WWW'00 et VLDB'00). Il est utilisé pour la construction et la gestion des sites Web des projets Caravel à l'Inria-Rocquencourt (`http://www-caravel.inria.fr`) et Aida à l'IRISA (`http://www.irisa.fr/aida/aida-new`).

5.6 Attman

Participants : Alberto Lerner, Eric Simon [correspondant].

Le logiciel Attman, développé en collaboration étroite avec Dennis Shasha (NYU, USA), offre un mode de navigation non hiérarchique dans des collections de données. Ce système utilise un modèle de données original composé de tables relationnelles et de cinq types de 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 sélection en utilisant les dépendances définies par le concepteur de l'application. La liste des tables pertinentes restantes 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 d'une expérimentation afin de construire un site Web adaptatif pour JavaDoc. Une autre expérimentation est en cours d'élaboration avec le Centre National de Génotypage afin de construire un site Web adaptatif pour l'ensemble des données génétiques.

6 Résultats nouveaux

La présentation des résultats de recherche est organisée selon les trois thèmes de recherche du projet présentés en Section 2. Globalement, les contributions ont surtout porté cette année sur la conception d'algorithmes et de méthodes d'optimisation et leur validation au travers de nos composants logiciels. On ne présente que les actions de recherche ayant donné lieu à des publications au cours de l'année 2000.

6.1 Accès à des ressources distribuées

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *le problème général abordé dans ce thème est celui de la publication de ressources dans un réseau ainsi que l'accès à ces ressources au moyen de langages de haut niveau. Les ressources peuvent être des données (structurées ou non) ou des services (librairies, programmes scientifiques, sites Web, etc), l'ensemble formant un* système d'information global. *Le système de médiation a la charge de mettre en relation de façon transparente les éditeurs des ressources avec les clients qui veulent utiliser ces ressources. Les actions de recherche qui structurent ce thème se distinguent selon l'approche, push ou pull, utilisée. Pour l'approche pull nous nous sommes concentrés sur deux questions essentielles. La première question est quelle vue uniforme des données faut-il présenter à l'utilisateur et quel est le langage de requêtes associé. Jusqu'à présent nous avons centré nos recherche sur le modèle relationnel et le langage SQL. Cette année, nous avons mené des recherches sur l'utilisation de XML comme formalisme de représentation uniforme des données, et sur la conception d'un langage de requêtes pour XML. La deuxième question est quelles techniques d'exécution efficaces peut-on élaborer pour les langages de requêtes supportés par le système de médiation. Dans l'approche push les clients sont intéressés par des informations hautement volatiles et dynamiques (des événe-*

ments); ils souscrivent des abonnements auprès du médiateur indiquant les informations qui les intéressent, les éditeurs font parvenir leurs informations (événements) au médiateur. Ce dernier se charge d'alerter les souscripteurs chaque fois qu'une information intéressante est émise par un éditeur. Dans la mesure où l'utilisation de médiateurs push dans les applications web telles les applications de commerce électronique "B2C" (bourse d'échange, billeterie, informations sur le trafic,...) est conditionnée par la faculté du système à supporter un grand nombre de clients et un flux élevé d'événements, la question est quelles sont les techniques de filtrage efficaces dans un tel contexte.

6.1.1 Optimisation de requêtes incluant des traitements coûteux

Participants : Luc Bouganim, Francoise Fabret, Fabio Porto, Patrick Valduriez.

Dans le cadre des applications environnementales que nous envisageons avec Le Select, les scientifiques peuvent typiquement avoir à poser des requêtes impliquant des données et des programmes (par exemple, un programme d'extraction de motifs dans une image) publiés en divers points du réseau. Même s'il est possible d'exprimer ces requêtes en SQL, les techniques classiques d'optimisation et d'exécution de requêtes sont insuffisantes pour deux raisons. Tout d'abord les optimiseurs de requêtes SQL considérant que le facteur prédominant est le coût des opérations de jointure, cherchent à minimiser ce coût en jouant sur l'ordre dans lequel sont exécutées les jointures. A contrario, dans notre contexte, le coût prédominant est celui de l'exécution des programmes et du transport des données volumineuses (telles que des images) depuis le site publiant ces données vers les lieux de traitement. D'où la nécessité d'établir des techniques d'optimisation spécifiques qui minimisent le nombre d'exécutions de programmes et la quantitié de données volumineuses transférées. Nous proposons deux techniques qui exploitent les possibilités d'optimisation inhérentes à l'architecture distribuée de systèmes comme Le Select. La première consiste à exécuter les traitements chers (appel des fonctions et transfert des données volumineuses) le plus tard possible: l'optimiseur planifie donc d'exécuter d'abord les opérations standard (sélections, jointures, etc). La seconde consiste à utiliser du caching pour éviter des traitements redondants. Le coût total en temps d'exécution est minimisé en parallélisant l'exécution des traitements: paralléllisme entre le transfert des données et l'exécution des programmes, exécution concurrente de plusieurs programmes. Le problème pour l'optimiseur est de décider de l'ordonnancement des programmes entre eux, et de décider la forme de parallélisme inter-programme: parallélisme pipeline, ou parallélisme indépendant. Nous proposons des algorithmes de planification dynamique permettant de s'adapter au flot de données constaté en cours d'exécution. Nous montrons que l'approche dynamique peut apporter des gains considérables en termes de temps de réponse. En effet, une mauvaise appréciation du flot de données peut avoir des conséquences désastreuses sur le temps de réponse si la planification est faite statiquement sans possibilité d'adaptation à l'exécution. Or, nous montrons que, dans de nombreux cas, il est très difficile d'avoir une connaissance du flot de données de façon statique. En effet, d'une part les statistiques sont peu fiables dans le contexte qui nous intéresse, et d'autre part même avec des statistiques très précises, il est la plupart du temps impossible d'avoir des informations sur la distribution des données (cette dernière n'est connue

qu'à l'exécution), or cette information est déterminante pour planifier l'exécution de plusieurs fonctions en parallélisme pipeline.

6.1.2 L'interrogation des données XML

Participant : Daniela Florescu.

Depuis quelques années, le langage XML (eXtensible Markup Language) est de plus en plus souvent utilisé comme langage de description de données, en raison de sa flexibilité et de son grand pouvoir d'expression. Ainsi, pour échanger des données (dont les formats et schemas peuvent être differents) entre plusieurs organisations coopérantes, une solution consiste à choisir XML comme le modèle commun, et de présenter toutes les collections de données comme des documents XML. D'autres contextes d'application justifient la mise en place des vraies bases de données/documents XML (e.g. documents administratifs, fonds bibliographiques etc.). De tous ces contextes applicatifs surgit le besoin d'un langage d'interrogation de haut niveau pour ce nouveau format de données. L'élaboration d'un standard pour le langage de requêtes pour XML est en cours, elle s'effectue dans le cadre d'un groupe de travail du World Wide Web Consortium (W3C); l'INRIA a été activement representée dans ce groupe de travail. En collaboration avec Jonathan Robie (Software AG) et Donald Chamberlin (IBM), nous avons élaboré un nouveau langage de requêtes pour XML nommé Quilt. Ce langage combine des éléments utiles de plusieurs langages existants et une syntaxe structurée similaire aux requêtes SQL. Quilt est un langage fonctionnel, centré sur la notion d'expression. Un type important d'expression est constitué par les expressions de chemin qui, à partir de la racine d'un document, sélectionnent un ensemble d'attributs ou d'éléments dans le document. Pour spécifier les expressions de chemin dans des documents XML, Quilt utilise la version courte de XPath, le standard du W3C. Par sa structure, Quilt permet aussi une interrogation très facile des données en format relationnel (présentées sous forme XML). Quilt a été proposé en novembre 2000 au groupe de travail du W3C sur les langages de requêtes pour XML et vient d'être adopté par ce groupe de travail.

6.1.3 Index plein-texte et requêtes XML

Participants : Daniela Florescu, Ioana Manolescu.

Dans le cadre de l'interrogation des données XML, la complexité de la structure des données, ou l'absence d'information sur la structure sont une source de difficulté majeure pour l'utilisateur. En effet, la structure d'un document XML peut être décrite par une grammaire (bien plus compliquée qu'un schéma relationnel), mais cette description n'est pas toujours présente. Pour explorer le contenu d'une source de données, l'utilisateur a donc besoin d'une mécanisme pour poser des requêtes qui ne nécessitent pas une connaissance préalable de la structure. De telles requêtes sont proches des recherches documentaires sur mot-clé, habituelles lors de la recherche d'information dans des documents HTML. Nous avons proposé d'ajouter à un langage d'interrogation de données XML l'expression de critère de sélection à base de mots-clé. Cette extension est rendue possible par l'utilisation d'un index plein-texte sur le contenu des documents XML. Cet index peut être matérialisé sur le même site que celui qui contient les

données ou, dans le cas d'une architecture à médiateur, il peut être stocké dans le médiateur si les propriétaires des données ne désirent pas héberger un tel index sur leur site. Au niveau du langage, cet index est rendu visible par la présence d'une fonction spéciale qui permet de tester si un certain élément contient ou non un mot donné. Nous avons montré l'interêt qu'il y a à ce que l'index retienne non pas seulement l'information sur les mots contenus directement dans un élément (sous forme de texte), mais aussi sur les mots contenus dans les descendants de cet élément, jusqu'à une profondeur fixée à l'avance.

6.1.4 Intégration de données XML et relationelles

Participants : Daniela Florescu, Ioana Manolescu.

Dans certains types d'applications d'intégration de données très structurées (tables relationelles) et de documents (qui peuvent être modélisées en XML), il est nécessaire d'être capable de répondre à des requêtes portant sur les deux types de données en même temps. Un exemple de ce genre d'application est la gestion des données médicales: certaines informations font référence aux données personnelles des malades et sont plutot structurées, tandis que le dossier médical cummule les annotations faites par le personnel médical au cours d'une longue période de temps (parfois plusieurs années) et est plus proche par structure d'un document XML. Le médiateur Le Select, réalise déjà l'intégration de données relationelles; nous l'avons donc étendu avec la capacité d'interroger, en même temps que des données relationelles, des documents XML. Pour ce faire, nous avons construit un adapteur (wrapper) spécial pour des documents XML, qui se base sur l'interface DOM, et qui présente le document à l'unité de traitement de requêtes sous la forme d'une collection de tables virtuelles. En posant une requête sur ces tables, on extrait de ce document les informations désirées. Par la suite, les tuples fournis par ce wrapper sont traités dans un cadre uniforme, avec les tuples de données relationnelles, ce qui permet la construction du résultat à partir des deux sources d'information. Nous avons réalisé le système Agora, qui ajoute à LeSelect une interface de type XML : toutes les données publiées via Le Select sont vues dans un format XML sur lequel les utilisateurs posent des requêtes en Quilt. Ces requêtes des utilisateur sont traduites vers des requêtes SQL sur un schéma canonique virtuel. Une étape supplémentaire de réecriture transforme la requête posée sur le schéma virtuel en une requête exécutable par LeSelect. Les tuples résultant de cette exécution sont ensuite fournis à un module qui en compose les éléments XML demandés par la requête en Quilt.

6.1.5 Algorithmes de pattern matching

Participants : Françoise Fabret, François Llirbat, João Pereira, Ken Ross, Arno Jacobsen.

"Publier et souscrire" (publish/subscribe en anglais) est un paradigme dans lequel des utilisateurs expriment leurs sujets d'intérêt (des "souscriptions") et des agents externes (pouvant être eux-même des utilisateurs) "publient" des événements (par exemple des offres). Le rôle des logiciels pour "publier et souscrire" est d'envoyer les événements aux propriétaires des souscriptions satisfaites par ces événements. Par exemple, la souscription émise par un sous-

cripteur peut exprimer l'intérêt de ce dernier pour toute proposition de voyage en avion ayant certaines caractéristiques et dont le prix n'excède pas une certaine somme. Parallèlement, un événement publié peut consister en une offre de voyage en avion dans laquelle sont précisées certaines propriétés du voyage et son prix. En fait, une souscription ressemble à un déclencheur (trigger) en ceci qu'elle exprime une requête continuelle (plus exactement une condition) assortie d'une action (consistant en général à avertir le souscripteur) qui est déclenchée chaque fois que la condition est satisfaite. Cependant le concept de souscription est moins général que celui de déclencheur. De ce fait, de nouvelles structures de données et des implémentations doivent permettre de concevoir des systèmes pour la publication/souscription qui supportent de façon efficace un très grand nombre de souscriptions et un haut débit d'événements. Ceci conduit à la conception de nouveaux algorithmes et à de nouvelles techniques d'implémentation. En combinant des structures de données, avec des politiques de caching spécifiques et une technique particulière d'exécution des requêtes, nous avons obtenus un module de filtrage des événements capable de traiter 600 événements par seconde alors que 6 milions de souscriptions consistant en des conjonctions de triplets (attribut, comparateur, valeur) sont présentes dans le système. Le rôle d'un algorithme de filtrage est de calculer pour un événement arrivant dans le système, quelles sont parmi les souscriptions actuellement présentes dans le système celles qui sont satisfaites par cet événement. Nos algorithmes de filtrage sont basés sur deux principes. Tout d'abord, ils sont conçus pour s'exécuter exclusivement en mémoire centrale, d'où la nécessité d'optimiser la localité temporelle et spatiale pour prendre en compte le comportement des processeurs en terme de caching (cache processor). Ensuite, ils ont pour but de réduire au maximum le nombre de souscriptions à tester grace à l'utilisation de techniques d'indexage et de groupement des souscriptions. Les algorithmes ont été testés dans le module de filtrage du prototype Le Subscribe.

6.2 Production de données dérivées

Mots clés : hétérogénéité, optimisation, répartition.

Résumé : *Lorsqu'on recherche une information précise dans un système d'information global, cette information n'existe pas toujours à l'état brut dans une source d'information. Dans ce cas, l'information doit être construite sur mesure (c'est souvent le cas dans les applications d'aide à la décision). Deux cas de figure assez différents se produisent. Dans le premier cas, la donnée recherchée peut être obtenue par intégration, consolidation, ou restructuration de données existantes. Le nettoyage de données ("data cleaning", en anglais) qui intervient dans la construction d'entrepôts de données ("data warehouse", en anglais) est un cas typique de cette situation. Dans l'autre cas, les données recherchées peuvent s'obtenir au moyen d'un programme publié dans le système d'information global. Par exemple, on recherche une prédiction de l'évolution d'une nappe de pétrole lors d'une marée noire; cette information peut être calculée par un programme scientifique de modélisation de l'évolution de nappes. Mais l'exécution de ce programme peut à son tour nécessiter des données d'entrée qui n'existent pas de façon brute et qui doivent être elles*

aussi calculées. On obtient ainsi une chaîne de traitement dont les étapes de traitement correspondent à l'exécution de programmes publiés. Dans ces deux cas de figure, toute la difficulté est de simplifier la mise en oeuvre du calcul de ces données recherchées et de garantir que les données produites ont la qualité désirée. Les deux actions qui composent ce thème ciblent ce problème avec des approches adaptées aux deux cas de figure cités.

6.2.1 Modèles, langages et algorithmes pour le nettoyage de données

Participants : Daniela Florescu, Helena Galhardas, Eric Simon.

Le problème du nettoyage de données est bien connu dans le domaine des systèmes d'aide à la décision et des entrepôts de données où il constitue l'un des problèmes les plus difficiles à résoudre. De nombreux outils, appelés ETL ou "data cleansing", ont été développés pour répondre à cette difficulté. Cependant, dans le cas d'applications telles que la migration de données très faiblement structurées vers des données structurées ou l'intégration de données scientifiques dans des domaines pluri-disciplinaires (e.g., la santé ou l'environnement), les outils existant destinés à l'écriture de programmes de nettoyage de données sont très insuffisants. Le problème principal rencontré est la conception d'un programme qui modélise un graphe à flots de données capable de transformer les données d'origine en données correctes et cohérentes et qui s'exécute efficacement sur de gros volumes de données. Ce problème résulte de deux lacunes dans les systèmes existant: (i) le manque de séparation claire entre la spécification logique des opérations de transformation de données nécessaires et leur implantation physique, et (ii) le manque de fonctionalités permettant d'assister l'utilisateur dans la mise au point de son programme de nettoyage. Les recherches effectuées dans cette action ont répondu à ces lacunes par la proposition d'un langage déclaratif de spécification de programmes de nettoyage, d'un modèle d'exécution logique pour les opérations de nettoyage de données et des algorithmes efficaces de mise en oeuvre de ces opérations. Le modèle d'exécution intègre quatre opérateurs spécifiques (mapping, matching, clustering, merging) qui permettent de décomposer un flot de données en plusieurs flots, de recomposer plusieurs flots par comparaison de la similitude de leurs données, de partitioner un flot en groupes, ou de fusionner un groupe de données en une donnée unique. Tous ces opérateurs sont paramétrables par des fonctions fournies par l'utilisateur (par exemple, des fonctions de calcul de similitude entre données). Le langage utilise une syntaxe proche de SQL et permet d'exprimer les opérations du modèle d'exécution de façon déclarative. Il permet également de spécifier les conditions dans lesquelles des exceptions doivent être générées et l'utilisateur doit être sollicité pour une intervention manuelle dans le processus de nettoyage. Un mécanisme sophistiqué de gestion d'exceptions et d'explication assiste l'utilisateur à mettre au point son programme. Enfin, le choix d'algorithmes efficaces pour exécuter les opérations est effectué par un optimiseur. Toutes ces propositions ont été implantées dans le logiciel AJAX et validées sur une application de nettoyage des références bibliographiques utilisées par le cite Web Citeseer avec un échantillon de 100.000 références. Cette validation a permis de vérifier l'utilité du langage et des fonctionnalités de gestion d'exceptions dans la mise au point de programmes ainsi que la performance des algorithmes proposés.

6.2.2 Workflow scientifiques

Participants : François Llirbat, Eric Simon.

Les workflows scientifiques, dérivés des workflows de gestion, on été introduits comme un moyen pratique pour spécifier des expériences scientifiques. En effet, ils fournissent un modèle permettant une description formelle des expériences, facilitant ainsi leur exécution automatisée. Ils permettent aussi l'archivage et l'interrogation des expériences passées. Enfin, en imposant une notation commune, ils permettent aux différents laboratoires de comparer plus facilement leurs expériences. Cependant les solutions actuelles imposent une gestion centralisée des expériences autour d'une même base de données. Cette approche n'est pas adaptée à un contexte d'information global dans lequel les scientifiques peuvent partager facilement leurs données, leurs programmes et leurs expériences au travers d'internet tout en preservant le principe d'autonomie: chaque donnée et chaque programme produit par les scientifiques d'un laboratoire restent la propriété du laboratoire et sont maintenus par leurs créateurs. Dans un tel contexte, un premier problème est la publication de ces ressources de façon à permettre leur utilisation par d'autres scientifiques. Ce problème de publication est un problème difficile car il faut fournir un environnement de publication dans lequel les scientifiques puissent spécifier les conditions (parfois complexes) dans lesquelles leur données ou programmes peuvent être utilisés. Une fois ces informations publiées, une deuxième difficulté est la mise au point d'expériences distribuées conbinant l'exécution de plusieurs modèles sur différentes données distribuées sur le réseau. Pour mettre au point de telles expériences il faut trouver les bons modèles, sélectionner les données les plus pertinentes et choisir les bons traitements sur ces données et finalement organiser leur exécution. Les recherches que nous avons effectuées dans ce domaine ont conduit à la mise au point d'un modèle formel pour les workflows scientifiques avec une sémantique déclarative bien définie. Ce modèle distingue trois niveaux d'information. Tout d'abord l'information brute comprenant le code des programmes et le contenu des données directement utilisées par les programmes, deuxièmement l'information contextuelle qui décrit chaque donnée et chaque programme ainsi que les contraintes opérationnelles associées, et enfin le schema du workflow qui décrit le flot de données entre les programmes et leur synchronisation. La mise au point d'expériences avec ce modèle consiste en trois phases. D'abord, l'expérience est spécifiée en definissant un schema de workflow à partir des schema contextuels décrivant les différent types de données et de programmes publiés au travers du système. La deuxième phase est une phase d'instanciation qui permet de choisir les instances de programmes et les données les mieux adaptées à l'expérience; cette phase peut être automatique ou semi-automatique. La dernière phase consiste en l'exécution distribuée des ces programmes sur le réseau. Cette dernière phase est basée sur l'utilisation du middleware LeSelect.

6.3 Construction de sites Web

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 par exemple dans les catalogues électroniques sur le Web.

6.3.1 Techniques de caching pour les sites web

Participants : Daniela Florescu, Khaled Yagoub.

Un site web à usage intensif de données est un site web qui gère et qui met à la disposition des utilisateurs un grand volume de données, qu'il permet de consulter et de modifier. Ces dernières années, ce type de site a évolué depuis des sites simples servant des pages statiques à des sites qui servent un grand nombre de pages web générées dynamiquement à partir de grandes bases de données. Dans ce contexte, la demande d'une page par un client peut nécessiter une interaction coûteuse avec le système de gestion de base de données pour la connexion au système et l'exécution des requêtes nécessaires à la récupération des données, risquant ainsi d'augmenter considérablement le temps d'attente du client. De plus ce type de site utilise généralement des applications ou des scripts écrits dans des langages tels que Perl, C, ou C++ pour construire dynamiquement des pages Web à partir de la base de données. Ces applications Web sont codées en dur ce qui réduit les possibilités d'optimisation automatique. Pour répondre à ce problème, nous avons développé une architecture configurable de caches à plusieurs niveaux qui permet de cacher des données extraites d'une base de données sous forme de tables relationnelles, de fragments XML et/ou de pages HTML. Notre solution s'appuie sur une spécification déclarative de la structure d'un site Web. Les données à publier dans le site sont stockées et gérées dans une base de données relationnelle, le contenu et la structure du site sont décrits à l'aide d'un modèle logique (dit graphe xml du site) qui est, généralement basé sur la notion de graphe et la présentation graphique, quant à elle, est spécifiée en utilisant des feuilles de style XSL. La correspondance entre le modèle logique et les données dans leur modèle de base (c'est-à-dire relationnel) est spécifiée à l'aide d'un langage déclaratif appelé WeaveL. Depuis ce langage, il est possible de spécifier et de personnaliser différentes stratégies de caching. Nous avons évalué, expérimentalement, les performances du système en variant les stratégies de gestion de cache en utilisant une plate-forme de test spécifique (WeaveBench). Les résultats obtenus montrent clairement qu'une stratégie de cache mixte, lorsqu'elle est bien choisie, est généralement la meilleure.

7 Actions régionales, nationales et internationales

7.1 Actions régionales

A l'INRIA, nous entretenons depuis de nombreuses années une collaboration étroite avec le projet VERSO. Cette année, la collaboration a été marquée par la participation de François

Llirbat au projet Xylème. 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 avons collaboré avec le laboratoire PrIsm de l'Université de Versailles sur les problèmes d'accès à l'information distribuée et les workflow scientifiques.

7.2 Actions européennes

7.2.1 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 a développé le système Le Select.

7.2.2 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 géographiques, données d'occupation des sols, données météorologiques, données d'émission de polluants. 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 la maintenance automatique de la fraîcheur et de 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 doit 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 d'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. Les partenaires de ce projet sont: le GMD à Berlin, l'UPM

à Madrid, le CLRC-RAL en Angleterre, le FORTH-ICS en Grèce, BULL en France et le SICE en Espagne.

7.3 Actions internationales

7.3.1 Europe

- Ecole Polytechnique de Bucarest avec qui nous avons signé un protocole d'accord. Dans ce cadre, nous avons accueilli cette année cinq étudiants roumains pour un stage de 6 mois. Deux d'entre eux se sont inscrit en DEA à Paris et effectueront leur stage au sein de l'équipe. Trois nouveaux stagiaires seront accueillis l'année prochaine.

- Yannis Ioannidis (Université d'Athènes) et Timos Sellis (NTUA, Athènes) avec qui nous travaillons sur les workflow scientifiques.

- Donald Kossman (Université de Munich) avec qui nous travaillons sur des techniques d'optimisation des langages de requêtes pour XML.

- Université technique de Lisbonne avec laquelle nous avons 2 contrats de coopération financés par la "Coopération Technique et Technologique Ambassade de France-ICCTI".

7.3.2 Amérique du Nord

- IBM, Almaden, Californie. Nous travaillons avec Chandra Mohan sur l'optimisation dynamique de langages de requêtes et avec Don Chamberlin sur la conception du langage Quilt.

- Bell Labs, New Jersey (Narain Gehani, Rick Hull). Cette année, les résultats du travail de recherche mené par François Llirbat et Eric Simon avec les chercheurs de l'équipe Vortex dirigée par Rick Hull ont abouti à un transfert industriel important au sein de Lucent Technologies.

- NYU, New York. Dennis Shasha, avec qui nous développons de fortes collaborations sur les projets Ajax, Attman et Le Subscribe, a séjourné dans notre équipe pendant une semaine.

- Université d'Alberta, Edmonton, Canada (Tamer Özsu).

- Université de Toronto. Nous avons proposé un projet de recherche franco-canadien en collaboration avec Arno Jacobsen qui prend un poste d'assistant professeur après avoir effectué un séjour de post-doc ERCIM dans notre équipe.

7.3.3 Amérique du Sud et Amérique Centrale

- universités de Rio de Janeiro (PUC, UFRJ IME et UNI-Rio), avec lesquelles nous avons un projet de coopération CNPQ-Inria (ECOBASE) sur les systèmes d'information pour l'environnement. Dans ce cadre, nous avons accueilli un thèsard pendant an et demi et nous accueillons une autre thèsarde pendant 3 mois. Françoise Fabret et Eric Simon ont

séjourné à Rio pendant 10 jours. Par ailleurs, nous organisons conjointement un workshop international sur l'intégration d'information en avril 2001.

8 Diffusion de résultats

8.1 Animation de la Communauté scientifique

Daniela Florescu représente l'INRIA dans le groupe de travail W3C sur la définition du langage de requêtes standard pour XML. Patrick Valduriez a participé à l'expertise du laboratoire CERMICS (ENPC-INRIA). Il est aussi membre du comité d'experts du RNTL.

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): E. Simon, P. Valduriez
- Int. ACM SIGMOD Conf: D. Florescu
- Conf. Nationale BDA: F. Llirbat
- Int. Conf. on Data Engineering (ICDE): D. Florescu, E. Simon
- Conf. Int. CARI: P. Valduriez

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).
- Network and Information Systems Journal, Hermes (M. Bouzeghoub, rédacteur en chef, E. Simon, P. Valduriez).

8.2 Enseignement

Eric Simon occupe un poste de Directeur-Professeur à temps partiel dans le département de Génie Informatique de l'Ecole Supérieure d'Ingénierie Léonard de Vinci depuis le 1er Avril 2000. Patrick Valduriez occupe dorénavant un poste de professeur à l'Université de Paris 6.

- entrepôts de données, PULV, 40 heures: E. Simon
- Bases de données réparties, mastère SIR de l'ENST-ESSEC, 30 heures et ISTY de l'UVSQ, 18 heures: P. Valduriez.

- Cours d'algorithmique, PULV, 10 heures: F. Llirbat
- Bases de données, PULV, 20 heures: F. Llirbat
- Bases de données à objets, Université Paris Sud (MIAGE 2ème année), 6 heures (D. Florescu, I. Manolescu).

8.3 Brevets

Trois brevets ont été élaboré lors des séjours de Francois Llirbat et d'Eric Simon a Bell Labs et déposé par Lucent technologies. Ils sont en cours de validation.

- Brevet "Declarative Workflow System Supporting Side-Effects" par R. Hull, F. Llirbat, E. Simon, G. Zhou, J. Su, G. Dong.
- Brevet "Eager Evaluation of Tasks in a Workflow System" par R. Hull, B. Kumar, F. Llirbat, G. Zhou.
- Brevet "Data Item Evaluation Based On the Combination Of Multiple Factors" par R. Hull, F. Llirbat, E. Simon, G. Zhou.

9 Bibliographie

Articles et chapitres de livre

[1] L. Bouganim, F. Fabret, P. Valduriez, C. Mohan, « A Dynamic Query Processing Architecture», *IEEE Data Engineering Bulletin - Special issue on adaptive query processing*, 2000.

[2] H.-A. Jacobsen, O. Günther, G. Riessen, « Component Leasing on the World Wide Web», *Special issue of Netnomics Journal "Information and Communication Middleware", Baltzer Science Publisher*, 2000.

[3] H.-A. Jacobsen, B. Krämer, « Design Patterns for Synchronization Adaptors of CORBA Objects », *Special issue of L'OBJET Journal on "Object Orientation and Formal Methods", Hermes Publisher*, 2000.

[4] F. Llirbat, E. Simon, J.-P. Berroir, «Specifying Scientific Experiments by Means of Declarative Workflow», *Systems Analysis Modelling Simulations*, 2000, à paraître.

[5] E. Pacitti, E. Simon, « Update Propagation Strategies to Improve Freshness in Lazy Master Replicated Databases», *The VLDB Journal*, février 2000.

[6] P. Vassiliadis, M. Bouzeghoub, C. Quix, «Towards Quality-Oriented Data Warehouse Usage and Evolution», *Information Systems*, 2000.

Communications à des congrès, colloques, etc.

[7] C. Bobineau, L. Bouganim, P. Pucheral, P. Valduriez, «PicoDBMS: Scaling down Database Techniques for the Smartcard», *in : VLDB*, Cairo, Egypt, 2000.

[8] L. Bouganim, F. Fabret, P. Valduriez, C. Mohan, «Dynamic Query Scheduling in Data Integration Systems», *in : Int. Conf. on Data Engineering (ICDE)*, San Diego, California, 2000.

[9] D. Florescu, D. Kossmann, I. Manolescu, «Integrating keyword search into XML query processing», *in : BDA*, Blois, 2000.

[10] D. Florescu, D. Kossmann, I. Manolescu, «Integrating keyword search into XML query processing», *in : WWW*, p. 119–135, Amsterdam, Holland, 2000.

[11] H. Galhardas, D. Florescu, D. Shasha, E. Simon, «AJAX: An Extensible Data Cleaning Tool», *in : SIGMOD (demonstration paper)*, 2000.

[12] H. Galhardas, D. Florescu, D. Shasha, E. Simon, «An Extensible Framework for Data Cleaning», *in : ICDE (poster paper)*, 2000.

[13] H. Galhardas, D. Florescu, D. Shasha, E. Simon, «Declaratively cleaning your data using AJAX», *in : BDA*, Blois, France, October 2000.

[14] H.-A. Jacobsen, B. Krämer, «Modeling Interface Definition Language Extensions», *in : 37th International Conference on Technology of Object-Oriented Languages and Systems (TOOLS-37)*, Sydney, Australia, 20-23 November 2000.

[15] F. Llirbat, R. Hull, B. Kumar, G. Zhou, J. Su, G. Dong, *in : Int. Conf. on Data Engineering (ICDE)*, San Diego, California, 2000.

[16] I. Manolescu, D. Florescu, D. Kossmann, D. Olteanu, F. Xhumari, «Agora: Living with XML and relational», *in : VLDB*, Cairo, Egypt, 2000.

[17] N. Marchand, H.-A. Jacobsen, «An Economic Model to Study Dependencies Between Software Application Vendors and Application Service Providers», *in : 3rd Berlin Internet Economics Workshop*, May 26th 2000.

[18] J. Pereira, F. Fabret, F. Llirbat, R. Preotiuc-Pietro, K. A. Ross, D. Shasha, «Publish/Subscribe on the Web at Extreme Speed», *in : Proccedings of the 26th VLDB Conference*, 2000.

[19] J. Pereira, F. Fabret, F. Llirbat, D. Shasha, «Efficient matching for web-based publish/subscribe systems», *in : Proc. of the Int. Conf. on Cooperative Information Systems (COOPIS)*, Eilat, Israel, 2000.

[20] K. Yagoub, D. Florescu, C. C. Andrei, V. Issarny, «Building and Customizing Data-intensive Web Site using Weave», *in : Proc. of the Int. Conf. on Very Large Data Bases (VLDB)-software demonstration*, 10-14 september 2000.

[21] K. Yagoub, D. Florescu, P. Valduriez, V. Issarny, «Caching Strategies for Data-Intensive Web Sites», *in : Proc. of the Int. WWW Conf. (poster)*, May 15-19 2000.

[22] K. Yagoub, D. Florescu, P. Valduriez, V. Issarny, «Caching Strategies for Data-Intensive Web Sites», *in : Proc. of the Int. Conf. on Very Large Data Bases (VLDB) and in (BDA)*, 10-14 september 2000.

[23] K. YAGOUB, D. FLORESCU, P. VALDURIEZ, V. ISSARNY, « WEAVE: A Data-Intensive Web Site Management System», *in : Proc. of the Conf. on Extending Database Technology (EDBT)-software demonstration*, March 27-31 2000.

Rapports de recherche et publications internes

[24] F. FABRET, F. LLIRBAT, A. JACOBSEN, J. PEREIRA, K. A. ROSS, D. SHASHA, « Efficient Matching Algorithms for Publish/Subscribe systems», *rapport de recherche*, INRIA, December 2000.

[25] H.-A. JACOBSEN, « Load Balancing and Performance Monitoring RFP DRAFT », *rapport de recherche*, Object Management Group, August 14th 2000, (Request for Proposal Draft, document number orbos/00-08-14).

[26] M-J.BLIN, F.FABRET, « A cooperative work framework integrating collaboration and cooperation», *rapport de recherche*, Université de Dauphine, Paris, December 2000.