INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet VERSO

Bases de Données

Rocquencourt

THÈME 3A

Table des matières

1 Composition de l'équipe

Responsables scientifiques

Serge Abiteboul [DR]

Sophie Cluet [DR]

Assistante de projet

Danièle Moreau [AJA, en commun avec le projet MEVAL]

Personnel INRIA

Stéphane Grumbach [DR]

Anne-Marie Vercoustre [DR, Janv-Août]

Luc Segoufin [CR]

Conseillers scientifiques

Claude Delobel [Professeur, Univ. Paris 11]

Michel Scholl [Professeur, CNAM]

Collaborateur extérieur

Bernd Amann [Maître de Conférence, CNAM]

Chercheurs invités

Catriel Beeri [Professeur, Univ. Jersusalem, 3 mois]

Léonid Libkin [Chercheur, Bells Labs-USA, Jan-Sept]

Victor Vianu [Professeur, U.C. San Diego, 4 mois]

Ingénieur expert

Guy Ferran [IE INRIA]

Chercheurs doctorants

Vincent Aguiléra [Ingénieur Ministère Equipement, Ecole des Ponts]

Federico Arambarri [Boursier EGIDE, UNLP-Argentine, mars-nov]

Grégory Cobéna [XTélécom, Oct-Déc]

Irini Fundulaki [Boursière INRIA, CNAM]

Amélie Marian [Boursière MENRT, Paris-Dauphine, Jan-Août]

Laurent Mignet [Boursier MENRT, CNAM]

Benjamin Nguyen [Boursier MENRT, Orsay]

Pierangelo Veltri [Boursier INRIA, U. Paris 11]

Fanny Wattez [Boursière INRIA, Paris 1, Jan-Nov]

Chercheur post-doctorant

Till Westmann [Univ. Mannheim, Mars-Oct]

Stagiaires

Grégory Corona [DESS Paris XI, Mars-Août]

Philippe Couspeyre [DESS Paris VI, Mai-Sept]

John Freddy Duitama [DEA Paris I, Mi-avril-Mi-août]

Antoine Galland [DEA Paris VI, Mars-Août]

Jérémy Jouglet [Polytechnique, Avril-Juil]

Sandrine Lafois [Paris VI, Avril-Sept]

David Le-Niniven [DESS Paris XI, Mars-Août]

Carlos Lobato [Maîtrîse Univ.Versailles, Mars-Août]

Benjamin Nguyen [DEA Paris XI ENS Cachan]

Mihai Preda [DEA Paris XI ENS Cachan]

Maya Ramanath [Indian Institute Bangalore, Mai-Août]

2 Présentation et objectifs généraux

Les données sont de plus en plus complexes, distribuées, hétérogènes, répliquées, multiformes, changeantes. L'objectif du projet est l'étude des problèmes fondamentaux posés aux systèmes de gestion de bases de données existants et le développement de solutions novatrices appropriées. Notre but est d'obtenir des systèmes plus ouverts à des données plus riches (documents, cartes), ouverts vers le réseau. Suivant une tradition bien établie, nous étudions les problèmes sur le plan théorique et sur le plan pratique, dans un effort permanent d'adapter nos résultats fondamentaux aux réalités du monde industriel avec lequel nous avons des partenariats.

L'émergence de XML comme format d'échange standard pour Internet ouvre des perspectives particulièrement intéressantes pour l'utilisation des données du WWW. En effet, à l'opposé de HTML, XML apporte de l'information sur la structure des documents et, par là même, sur leur sémantique. Il nous paraît primordial de mettre tout en œuvre pour exploiter cette nouvelle propriété des données du Web. Nous voulons notamment permettre une recherche plus intelligente sur le Web et une meilleure utilisabilité des informations trouvées. Les techniques que nous avons développées ces dernières années autour des données hétérogènes et semistructurées s'appliquent parfaitement à cette nouvelle préoccupation du projet. Nous poursuivons également nos travaux sur les fondements théoriques des bases de données et nos activités sur les bases de données spatiales et le commerce électronique.

Notre action de recherche s'articule cette année autour de trois thèmes : (i) bases de données et le Web (ii) gestion de données multidimensionnelles, et (iii) systèmes de règles actives pour le commerce électronique.

Nous avons des relation industrielles avec la société MatchVision dans le cadre du projet RNRT GAEL et la société Xyleme SA créée en Septembre 2000 en grande partie à partir de travaux du projet Verso. Au niveau européen, nous participons aux projets C-Web, MesMuses et TMR Chorochronos. Nous collaborons aussi avec l'Université de Tel Aviv sur l'"usine du Futur" dans le cadre d'un contrat AFIRST.

3 Fondements scientifiques

3.1 Bases de données et le Web

Mots clés : XML, DTD, stockage, indexation, architecture, ontologie, versions, requête temporelle, notification, moteur de recherche, intégration, World Wide Web, schéma, document, hétérogénéité.

Malgré le formidable développement du Web et son utilité pour de nombreuses communautés, la recherche et l'exploitation d'informations s'y révèlent encore très difficiles. Les outils actuels de recherche sur le Web ne permettent que l'expression de requêtes vagues auxquelles sont données des réponses peu précises qu'il faut péniblement analyser pour obtenir l'information réellement recherchée. Les limitations actuelles sont principalement dûes à l'utilisation du langage HTML pour écrire les pages Web. Ce langage est très pauvre : il n'apporte aucune information structurelle ou sémantique.

Pour pallier cette déficience, le *World Wide Web* consortium propose de remplacer HTML par le langage XML. Le projet Verso mise sur cette révolution et a développé cette année un système, nommé Xyleme, qui permet de collecter, analyser et intégrer tous les documents XML du Web et leur évolution. Ce système offre ainsi une vision simple et structurée de la masse d'information que représente le Web, permettant par la même, une interrogation fine et précise. Cette action s'inscrit dans la continuité des nombreux travaux du projet sur ce thème.

Cette action ambitieuse aborde de nombreux aspects, dont certains ont été traités par les équipes amies de l'université de Manheim (groupe bases de données du Professeur Moerkotte) et du LRI (équipe Intelligence Artificielle et Systèmes d'Inférence).

- **Stockage**

  Une collaboration avec l'université de Manheim a permis de mettre au point un système de stockage efficace des données et méta-données XML. Par ailleurs, nous avons étendu un logiciel d'indexation plein-texte pour prendre en compte les éléments structurels contenus dans les documents.

- **Langage de requêtes**

  Partant de notre expérience sur les langages OQL, POQL et Lorel et des travaux issus du W3C, nous avons défini un langage de requêtes complet dont l'algèbre inclut de manière efficace toutes les opérations nécessaires à la sélection d'arbres XML, leur filtrage et la construction de résultats dans le même modèle. Ce langage a été implanté partiellement.

- **Evolutions temporelles**

  Nous avons conçu et implanté des techniques permettant la gestion de versions de documents basées sur une notion de *delta*. Nous avons également implanté un système de souscription de requêtes basé sur le monitoring et les requêtes continues.

- **Intégration sémantique**

  La nature hétérogène des données implique un traitement sémantique de celles-ci afin de les intégrer ou de les classer. L'extraction des connaissances sémantiques reposent sur des techniques proposées par l'équipe IASI. Le stockage et l'utilisation de ces connaissances pour l'interrogation conviviale des documents ont été pris en charge par Verso.

- **Acquisition de données**

  La technologie des moteurs de recherche a été adaptée pour rechercher des données XML sur le Web. Notamment, des techniques ont été mises en œuvre pour comprendre quand rafraîchir une page XML ou HTML afin de suivre en temps presque réel les évolutions du Web. Il s'agit de prendre en compte l'importance des documents, leur taux de changement estimé, ainsi que l'intérêt que chaque document peut représenter pour des utilisateurs du système.

- **Architecture**

  On a aussi étudié et proposé des solutions aux problèmes spécifiques liés à la taille de l'entrepôt (de l'ordre du téra-octet) : regroupement, partitionnement, réseau, parallélisme.

3.2 Portails pour Communautés Web

Mots clés : communauté Web, World Wide Web, intégration sémantique, méta-données, XML, RDF.

Le World Wide Web n'est pas seulement une incroyable source d'information qui change chaque seconde, mais également un labyrinthe où l'utilisateur a souvent du mal à trouver l'information qui l'intéresse. Les moteurs de recherche traditionnels fondés sur une indexation plein-texte de documents (HTML, PDF, Ascii, ...) ont rapidement montré la limite de la recherche par contenu simple. Ainsi, la plupart proposent des systèmes de classification (Web directories) qui permettent d'organiser les URLs par rapport à différentes catégories/sous-catégories comme Culture, Culture/Musique, Culture/Musique/Opera etc... Le moteur de recherche devient un *portail* permettant d'organiser des pages en accord avec une structure sémantique prédéfinie (ontologie, thésaurus) et indépendante du contenu propre de ces pages.

Actuellement, un grand nombre d'entreprises vendent avec succès des portails destinés à des communautés d'utilisateurs qui veulent partager et échanger des documents (texte, audio, image, vidéo) concernant un certain *domaine d'intérêt*. Les notions de partage et d'échange d'information concernent plusieurs niveaux. Le premier niveau est implanté par Internet et permet un accès physique aux données dans un environnement hétérogène et distribué. Le deuxième niveau permet un traitement syntaxique de l'information et le nouveau standard XML répond très bien à la nécessité de représenter la structure de documents et de données en général (voir 3.1). Le troisième niveau concerne la sémantique des informations. Ce niveau est essentiel pour un vrai partage de l'information non seulement au niveau de la structure mais également au niveau de la compréhension par l'utilisateur. Une solution est la création d'informations descriptives ou méta-données concernant les documents sources. Ces descriptions permettent le partage de ressources Web entre plusieurs utilisateurs.

L'objectif du projet CWeb, qui s'est terminé fin Septembre 2000, était l'évaluation et la définition d'une plate-forme pour la génération de tels portails sémantiques. Les problèmes étudiés concernaient notamment le modèle de description de ressources et l'accès aux ressources en fonction de leurs descriptions sans connaître leur localisation. Le projet Européen MESMUSES (Metaphor for Science Museums), qui a démarré en Octobre 2000, confronte la plate-forme CWeb à des besoins de deux musées scientifiques (Cité des Sciences et de l'Industrie, Paris et Istituto E Museo Di Storia Della Scienza, Florence).

3.3 Données complexes

Mots clés : multidimensionnel, contrainte, topologie, statistique, données semi-structurées.

Les données multidimensionnelles et semi-structurées, posent des problèmes fondamentaux aux systèmes de gestion de bases de données, non seulement à cause de la taille considérable et en croissance constante de ces données, mais aussi à cause de la complexité des relations qu'entretiennent ces données entre elles.

Nous nous sommes tout d'abord intéressés à la définition de modèles de données et de langages de requêtes de haut niveau d'abstraction pour le temps, l'espace et plus généralement

les données multidimensionnelles. Pour cela, nous travaillons depuis plusieurs années sur un modèle, basé sur l'approche par contraintes, dans laquelle un objet géométrique est représenté par une formule de la logique du premier ordre. Ce modèle permet de fournir à l'utilisateur une vue des données indépendante de leur représentation physique et d'unifier le mode de représentation des données spatiales et des autres données alphanumériques. Dans les systèmes d'information spatiale existants, ces deux caractéristiques font défaut.

Dans ce cadre, un effort important a été fait dans l'étude et la mise au point de langages de requêtes de haut niveau avec un fort pouvoir d'expression tout en conservant une complexité d'évaluation raisonnable. Nous avons aussi proposé des tentatives de solution au problème de l'optimisation de ces requêtes qui sont exprimées indépendamment de leur stratégie de calcul. On propose, pour cela, une méthode basée sur des évaluations en deux temps, l'une approximative, l'autre exacte, ansi qu'une méthode utilisant des propriétés d'interpolation géométriques des données.

Nous nous sommes aussi intéressés à la modélisation des données statistiques, du type tableau avec des pourcentages, moyennes, etc. Nous avons proposé un modèle basé sur le modèle relationnel, dans lequel la sémantique des relations est représentée par des formules de Horn avec des contraintes sur les réels. Nous étudions, en particulier, la possibilité de dériver de nouvelles informations statistiques d'une base de données statistiques.

Les données semi-structurées font partie du domaine d'expertise de Verso depuis maintenant plusieurs années. Cela s'est concrétisé par la publication d'un livre faisant une présentation générale du domaine. Les données accessibles sur la toile sont un bon exemple de données semi-structurées. Leur nombre, en croissance constante, pose de sérieux problèmes de stockage et, en pratique, l'information disponible n'est jamais complète. Nous étudions des mécanismes compacts d'identification des noeuds d'un document semi-structuré et des techniques de représentation et d'interrogation de données semi-structurées incomplètes.

3.4 Vues Actives pour le commerce électronique

Mots clés : règle de production, base de données active, parallélisme, déduction, datalog, workflow, commerce électronique, calcul relationnel, non-déterminisme.

Il s'agit de mieux comprendre des applications mettant en jeu du partage de données entre des clients qui interagissent *via* un système de "vues actives". Des vues actives permettent à un client de voir les données qui le concernent et d'intéragir avec le serveur et d'autres clients par un mécanisme de notification. Les aspects novateurs de nos travaux s'articulent autour de : (i) l'utilisation de données semi-structurées, (ii) l'étude de descriptions logiques des connaissances (celles du catalogue et des profils d'utilisateurs), et (iii) une spécification déclarative des notifications (Voir Projet RNRT GAEL.)

Ces travaux se situent dans le cadre de la spécification déclarative d'applications distribuées.

4 Domaines d'applications

Les bases de données n'ont pas de champs d'application privilégiés. En effet, toute application mettant en jeu une quantité importante de données ou d'informations se doit d'utiliser des

bases de données. Verso choisit de cibler des applications qui présentent des défis particuliers. C'est le cas pour les données spatiales et temporelles que l'on retrouve notamment dans des applications géographiques sur lesquelles nous travaillons (voir 7.2.3), ou pour le répertoire des documents XML du Web (voir 3.1), au-dessus duquel nous envisageons de fournir des services.

Nous nous contentons de mentionner plus en détail quatre applications à titre d'illustation. La première porte sur l'exploitation intelligente des données du Web. La seconde concerne également le Web, mais dans une vision Intranet. Nous nous intéressons notamment à des sous-Web pour des communautés spécifiques (CWeb), en particulier pour gérer la mémoire d'entreprise. La troisième considère la gestion de données culturelles pour lesquelles nous voulons créer des médiateurs *intelligents* s'appuyant sur des descriptions sémantiques des sources et du domaine (ontologies, thésauri). La quatrième, enfin, illustre les applications des bases de données multidimensionnelles.

Pour conclure cette excursion dans les domaines d'applications, nous noterons que Verso étudie, aussi, une application très à la mode : le commerce électronique. Certains aspects de cette application mêlent la gestion de données hétérogènes distribuées à des notions de transactions, familières dans une problématique bases de données. D'autres aspects plus originaux mettent en jeu des spécifications déclaratives des flots d'information entre les acteurs de telles applications (Voir 5.4 et Projet RNRT GAEL.)

4.1 Xyleme : exploitation des données du Web

Mots clés : internet, World Wide Web.

Depuis son adoption par le W3C comme format d'échange standard pour les données de l'Internet, XML a séduit tous les grands acteurs du marché informatique.

Contrairement aux données HTML, il est possible d'extraire la structure d'un document XML. Cette structure peut être utilisée de multiples façons. Notamment, elle permet d'interroger les documents de façon plus intelligente. Par exemple, il est concevable en XML d'évaluer une requête de type : "donner les documents contenant des produits dont le prix est supérieur à 10.000 Euros". Sur un ensemble de documents HTML, la requête la plus proche de celle-ci serait : "donner les documents contenant les mots produits et prix". Il est également possible d'envisager l'intégration automatique d'un ensemble de documents portant sur le même sujet mais dont les structures sont légèrement différentes. Ceci facilite l'interrogation mais également la manipulation des données par des applications (par exemple, une application sur des données génomiques collectées en différents endroits de la planète).

Verso parie sur XML. Nous pensons que ce format va gagner encore en popularité et que dans un futur relativement proche, la majorité des données publiées sur le Web seront XML. Nous comptons être alors présent avec des outils permettant une bonne exploitation de ces données. Nous nous proposons plus particulièrement de construire une base de données géante intégrant toutes les données XML du Web. Cette base servira de support à un ensemble de services tels que : interrogation conviviale et pertinente, abonnement de requêtes (e.g. prévenez-moi si un nouveau projet apparaît sur le site de l'INRIA) ou requêtes temporelles (e.g. quelle

est l'évolution du nombre de projets à l'INRIA depuis 1980).

Une version alpha de ce système est maintenant prête. La société Xyleme vient d'être créée et va prendre en charge ce prototype pour en faire un produit.

4.2 Intranets et mémoire d'entreprise

Mots clés : intranet, WWW, communauté.

De plus en plus d'entreprises utilisent les technologies de l'Internet pour stocker, gérer et échanger leurs informations internes ; c'est ce qu'on appelle *l'intranet*. On peut considérer que l'intranet de la plupart des sociétés contient déjà plus d'informations que leurs bases de données (c'est certainement le cas à l'INRIA). Il est important de savoir gérer les intranets de façon sûre et efficace pour en faire une vraie *mémoire d'entreprise* réutilisable.

Le projet CWeb a développé une plate-forme générique, fondée sur des standards ouverts (en particulier XML et RDF), pour créer des échanges sur le Web adaptés et spécialisés pour une communauté ou une entreprise particulière (voir 7.2.1). Cette plate-forme est maintenant évaluée par deux applications similaires dans le contexte de la préparation d'expositions grand public de musées scientifiques (voir 7.2.2).

4.3 Applications culturelles

Mots clés : culturel, musée, patrimoine, bibliothéque électronique.

Les applications culturelles concernent principalement la gestion de documents électroniques dans les domaines de l'art (e.g., musées), du patrimoine national (e.g., monuments), des bibliothèques, etc. Le but est d'ouvrir ces collections de documents aux chercheurs et au public le plus large. On retrouve les pôles d'intérêt de Verso : grande quantité d'informations multimédia (image, graphique, texte), distribuées géographiquement sur des supports (systèmes et formats) divers.

Verso s'appuie dans ce domaine sur sa maîtrise de la technologie objet. Dans le cadre du contrat Inventaire97 avec le ministère de la Culture, Verso travaille sur la navigation à l'aide de cartes géographiques (IGN, cadastre) pour accéder efficacement à une base de documents du patrimoine national. Nous avons également commencé à travailler sur un médiateur intelligent qui s'appuie sur une description sémantique de ces collections et du domaine considéré (voir 7.2.1).

4.4 Applications multidimensionnelles

Mots clés : temps, espace, objets mobiles..

Les applications des bases de données impliquent de plus en plus souvent des données multidimensionnelles, comme les cartes (par exemple géographiques), les plans d'occupation des sols, les descriptions 3D du sous-sol pour la géologie et l'exploitation minière, les objets mobiles (comme le déplacement des voitures), ou encore pour les objets de formes variables (comme les fleuves). La modélisation de ces données et le développement de langages permettant d'extraire

et de manipuler efficacement l'information stockée constituent un véritable défi au cœur des pôles d'intérêt de Verso : modélisation, langages de requête, optimisation.

Deux applications tournent aujourd'hui sur le prototype DEDALE [21]. La première, cartographique, porte sur une carte de la région d'Orange fournie par l'IGN et permet de répondre à un ensemble de requêtes typiques des systèmes d'information géographique. La seconde application est multidimensionnelle et consiste en des objets mobiles dans une station de ski, sur des données fournies par le LAMA de Grenoble (voir 7.2.4). L'interrogation se fait via une interface graphique dont la nouvelle version a été entièrement écrite en Java afin de permettre son utilisation à distance. D'autres applications sont à l'étude avec des contraintes temporelles plus fortes.

Le projet TMR Chorochronos visait explicitement les applications spatio-temporelles. L'objectif de ce projet était le développement d'une architecture idéale pour la modélisation et l'interrogation des données multidimensionnelles (voir 7.2.3).

Le projet GDR-PSIG MOB vise plus spécifiquement les applications portant sur des objets mobiles. L'objectif étant la mise au point d'un outil efficace permettant l'interrogation et la représentation d'une base de données contenant des objets mobiles.

5 Résultats nouveaux

5.1 Xyleme

Participants : Serge Abiteboul, Sophie Cluet, Guy Ferran, Vincent Aguiléra, Grégory Cobena, Amélie Marian, Laurent Mignet, Benjamin Nguyen, Pierangelo Veltri, Fanny Wattez.

Mots clés : Xyleme.

En collaboration avec l'université de Mannheim et l'équipe IASI du LRI, Verso a développé le système Xyleme, un entrepôt dynamique de toutes les données XML du Web. Xyleme soulève de nombreux et intéressants problèmes. Parmi les résultats notables obtenus par Verso, citons la gestion des aspects dynamiques, des vues et des requêtes.

5.1.1 Acquisition et maintenance de données

La gestion de l'acquisition de nouvelles pages XML dans Xyleme ainsi que leur raffraichissement sont présentées dans [39]. Il s'agit de guider la création d'un fond de pages XML trouvées sur le Web et de le maintenir à jour dans un Web changeant en permanence. On tient compte de facteurs tels que les importances respectives des pages ou les souhaits des utilisateurs du système.

5.1.2 Versions et souscriptions de requêtes

La représentation de versions en utilisant des "deltas" est présentée dans [38]. Il s'agit d'obtenir une représentation relativement compacte de l'histoire de certains documents facilitant des requêtes comme le calcul du changement depuis une version donnée. Un système de souscriptions à des requêtes a aussi été étudié. Il s'agit de pouvoir être notifé, par mail ou sur

une page Web, lors d'événements survenus sur la base (e.g. nouveau document, changement du contenu d'un document, etc.)

5.1.3 Gestion de vues

Notre objectif est de permettre aux utilisateurs du Web de formuler des requêtes précises (par exemple, le nom et l'adresse des responsables des ventes des sociétés informatiques de la région parisienne). Pour ce faire, nous nous proposons de découper le Web en domaines (la culture, le tourisme, les affaires, etc.), chacun étant décrit par une structure simple (comparable à un formulaire avec imbrication de champs). Une interrogation consiste, alors, à annoter cette structure.

Chaque site du Web a une structure différente. Il serait illusoire d'imaginer que le passage à XML va changer cet état de fait. Pour associer à chaque domaine une structure unique, il est donc nécessaire de comprendre les correspondances qui existent entre leur structure réelle et celle proposée par Xyleme. Ce travail a été réalisé par l'équipe IASI du LRI. Notre tâche a consisté à rendre ces correspondances persistantes et utilisables par l'évaluateur de requêtes. Ces dernière années, nous avions travaillé sur un système de vues XML [32] très puissant mais qu'il n'est pas concevable d'utiliser dans le contexte Xyleme, où le passage à l'échelle est primordial. En effet, la taille d'une vue (ensemble des correspondances) est fonction du nombre de structures réelles. Ce nombre tend à croître de façon exponentielle et il n'y a, a priori, pas de limite à cette extension. Cependant, pour permettre une évaluation rapide des requêtes, il est nécessaire de stocker dans une seule mémoire, une synthèse de la vue permettant notamment de déterminer les machines qui vont permettre de répondre à la requête. Egalement, il est important de comprendre comment traduire une requête portant sur une structure abstraite en des requêtes sur les documents réels de façon à minimiser la communication entre machines. Dans [43], nous proposons une solution à ces problèmes, solution que nous avons implantée dans le système Xyleme.

5.1.4 Evaluation de requêtes

Xyleme doit pouvoir répondre en un temps raisonnable à des milliers de requêtes concurrentes portant sur des milliards de documents répartis sur de nombreuses machines. Pour ce faire, nous avons choisi de partitionner les documents suivants des critères sémantiques afin de limiter le nombre de machines impliquées dans une requête. Egalement, nous avons étendu la technique d'indexation plein-texte afin de prendre en compte la structure des documents. Plus particulièrement, nous associons à chaque occurrence d'un mot dans un document un code permettant, étant donnés deux mots, de comprendre leur position relative dans le document. Cette technique permet une évaluation extrêmement rapide des requêtes de type formulaire imbriqué introduites précédemment. Ces travaux sont décrits dans [29, 42, 11].

5.2 CWeb

Participants : Bernd Amann, Irini Fundulaki, Michel Scholl, Anne-Marie Vercoustre.

Mots clés : Hétérogénéité, intégration, ontologie, thésaurus, linéarisation de thésaurus.

La description de ressources Web est un thème fédérateur pour différentes disciplines comme la représentation de connaissances, les bases de données et les systèmes d'information distribués. Dans ce contexte, nous nous sommes intéressés aux problèmes de la représentation et de l'interrogation de méta-données.

5.2.1 Conception de Schéma de Description

Nous avons développé une nouvelle approche pour la création contrôlée de méta-données ou description de documents Web. Cette approche est fondée sur la réutilisation de structures sémantiques qu'on appelle généralement des ontologies et thésaurus. Nous avons montré qu'il est possible de créer et d'adapter des schémas de description à partir d'une ontologie (schéma conceptuel) et de différents thésaurus (hiérarchies et termes), qui peuvent être choisis par rapport aux besoins de l'utilisateur. Un tel schéma peut facilement être représenté sous forme d'un schéma base de données classique, mais également sous forme d'un schéma RDF [1][14, 30]

5.2.2 Linéarisation de thésaurus

Un problème intéressant du point de vue base de données est l'optimisation de requêtes sur des hiérarchies de termes (thésaurus). L'idée est de coder les termes de thesaurus, de traduire des relations hiérarchiques entre termes en intervalles (fenêtres) dans un espace à une (plusieurs) dimension(s) et d'utiliser des index standards comme les arbres B+ (pour une dimension) et les arbres R (pour plusieurs dimensions) . Ainsi, un parcours d'arbre pour trouver, par exemple, tous les descendants d'un terme devient une requête intervalle sur les codes obtenus par la traduction [14].

Nous avons implanté et évalué différents codages dans un SGBD orienté-objet (O2) [44].

5.2.3 Prototype

Les idées ci-dessus ont été expérimentées au moyen d'un prototype implanté avec l'interface Java du SGBD orienté-objet O2. Nous avons construit un schéma pour la description de ressources culturelles à partir d'une ontologie du Comité International pour la Documentation du Conseil International des Musées (ICOM/CIDOC) et le thésaurus AAT (Art and Architecture Thesaurus) de l'institut Getty (J. Paul Getty Trust). Les descriptions sont stockées sous forme d'objets, ce qui permet une interrogation directe avec le langage de requêtes OQL de O2. Nous avons également développé une interface de programmation d'applications Web qui utilise le standard XML pour l'échange de données. Cette interface a été utilisée avec succès dans un autre projet.

1. RDF est une recommandation W3C pour la description de ressources Web.

5.3 Données complexes

Mots clés : multidimensionnel, semi-structuré, contrainte, topologie, statistique, langage de requêtes..

Participants : Stéphane Grumbach, Leonid Libkin, Michel Scholl, Luc Segoufin, Pierangelo Veltri, Victor Vianu.

5.3.1 Données semi-structurées

Les données accessibles sur le réseau vont du très structuré, dans des bases relationnelles, au totalement non structuré, dans des fichiers de texte. De plus, l'intégration de données est aussi souvent source d'irrégularité, des données similaires étant souvent représentées avec des structures différentes dans des sources indépendantes. On utilise le terme *semi-structuré* pour ces données qui ne sont pas vraiment structurées mais qui présentent une certaine structure même si celle-ci est peu régulière et implicite. Une présentation générale du domaine des bases de données semi-structurées est proposée en [13].

Dans [28], nous étudions des mécanismes d'identification compactes des nœuds d'un document semi-structuré. Nous analysons le "coût" de plusieurs techniques telles que des codages à la Huffman. Dans [41], nous étudions la représentation et l'interrogation de données semi-structurées incomplètes.

5.3.2 Bases de données avec contraintes

Les données multidimensionnelles et, en particulier, les données spatiales conduisent à des représentations infinies (e.g. sous-espace du plan réel) mais admettant une représentation effective finie. De telles données peuvent être représentées à l'aide de contraintes sur des domaines numériques, par exemple, les contraintes polynomiales sur les réels ou les contraintes linéaires sur les rationnels. Les bases de données avec contraintes généralisent les bases de données relationnelles et un certain nombre d'outils du modèle relationnel peuvent être utilisés dans ce contexte étendu. Un livre faisant un état de l'art sur les bases de données avec contraintes a été publié [10]. Le projet Verso a participé à la rédaction de nombreux chapitres de ce livre : [25, 15, 16, 17, 20, 21, 24, 19, 26].

La validité du modèle contrainte a été apportée par le projet Verso grâce à l'implémentation d'un prototype DEDALE en collaboration avec le CNAM (M. Scholl, P. Rigaux) [21, 22]. Le défi actuel est de réaliser l'optimisation de ces requêtes qui sont exprimées indépendamment de leur stratégie de calcul. On propose, pour cela, des méthodes basées sur des critères tels que la dimension, la géométrie et la topologie des données [20]. Dans [34], on montre comment il est possible de modéliser et d'interroger des données interpolées de manière complètement transparente pour l'utilisateur. On montre aussi que l'évaluation des requêtes sur de telles données se ramene à la manipulation d'objets de dimension 2, réduisant la complexité, exponentielle en la dimension, en une complexité linéaire. Dans [35], on propose une évaluation en deux étapes, la première rapide mais approximative servant de filtre à la seconde, plus coûteuse en temps.

5.3.3 Représentation de la topologie et requêtes topologiques

Dans certaines applications, la forme des régions est importante alors que, dans d'autres, on ne s'intéresse qu'à leurs propriétés *topologiques*. Dans [27, 24, 33], on étudie les requêtes topologiques sur des bases de données spatiales en dimension 2, où les régions sont spécifiées par des inégalités polynomiales avec des coefficients entiers. Un résumé des résultats connus sur ce sujet est présenté dans [24].

Toute l'information topologique d'une base de données spatiale de dimension 2 peut être regroupée dans une base de données relationnelle finie appelée "invariant" topologique et vue comme une annotation topologique des données spatiales brutes. Il est alors possible de répondre à n'importe quelle requête topologique en utilisant cet invariant topologique. Comme il a généralement une taille bien moins importante que la base de donnée spatiale elle-même, cette méthode induit une stratégie d'évaluation des requêtes topologiques potentiellement plus efficace. La traduction d'une requête sur des données spatiales en une requête équivalente sur l'invariant topologique est examinée dans [27]. Il est montré que le langage "fixpoint+counting" exprime exactement les requêtes topologiques PTIME sur l'invariant topologique. Cela suggère que les invariants topologiques sont des structures particulièrement bien adaptées à la logique descriptive.

Dans [33], on montre que, sans le mécanisme de point fixe ci-dessus, le pouvoir expressif du language au premier ordre est très limité. On caractérise celui-ci dans plusieurs cas particuliers.

5.3.4 Langages de requêtes

Un langage de requêtes pour données spatiales doit posséder deux propriétés fondamentales et difficiles à obtenir : clôture et vitesse d'évaluation rapide. Un langage est clos pour un modèle de données si le résultat de toutes les requêtes exprimables dans celui-ci est représentable dans le modèle données de départ. Cette propriété est essentielle car elle permet de réutiliser le résultat d'une requête afin de l'interroger à nouveau, de la visualiser ou de la stocker dans la base. Les modèles utilisés pour les données spatiales sont le modèle semi-linéaire et le modèle semi-algébrique.

Les langages de requêtes actuels sont basés sur des adaptations de SQL au domaine spatial. Ils sont clos et leur vitesse d'évaluation se fait en temps raisonnable (PTIME), mais ils sont très limités dans les requêtes qu'ils peuvent exprimer : ils ne possèdent pas de mécanisme de récursion et ne peuvent pas tester des propriétés de connexité [10]. Un simple mécanisme de récursion suffit pour rendre ces langages non clôs et impossibles à évaluer.

Dans [31], on montre qu'il est possible d'ajouter un prédicat de connexité à SQL, et en fait, n'importe quel prédicat topologique, tout en préservant la propriété de clôture. Enfin, on spécifie dans [31] un langage puissant, baptisé *Path Logic*, capable d'exprimer de nombreuses requêtes liées à la connexité, possédant la propriété de clôture et pouvant s'évaluer en temps polynomial en la taille des données. Path Logic peut servir de base à une nouvelle génération de puissants langages de requêtes spatiaux.

5.3.5 Vues et données statistiques

De nombreuses applications reposent sur des vues matérialisées (incomplètes) de données initiales qui ne sont pas accessibles. C'est le cas en particulier dans les applications mobiles, mais aussi pour certains types de données dans les entrepôts de données ou pour les données statistiques. Nous nous sommes intéressés au problème du calcul de la réponse à une requête, posée sur une base de données, en utilisant seulement les vues incomplètes [37, 23]. Le problème consiste à réécrire la requête sur les données initiales, dans une requête sur les vues. Nous avons, en particulier, caractérisé le contenu en information des vues.

Une application particulièrement intéressante du problème d'interrogation à l'aide des vues est fournie par les données statistiques. On suppose une population initiale (micro données) pour laquelle on a des informations sur les individus et des vues dérivées qui ne donnent que des informations statistiques (macro données). Le problème est, alors, de savoir si on peut répondre à une requête statistique sur la population initiale, à l'aide des vues seulement. Nous montrons des résultats négatifs, qui, dans certains cas, permettent de dire que l'information contenue dans les vues n'est pas suffisante. Dans [36], nous montrons comment dériver les requêtes sur les vues.

5.4 Vues actives pour le commerce électronique

Participants : Serge Abiteboul, Vincent Aguiléra, Bernd Amann, Sophie Cluet, Amélie Marian, Laurent Mignet, Anne-Marie Vercoustre.

Mots clés : règle de production, base de données active, parallélisme, déduction, datalog, workflow, commerce électronique, calcul relationnel, non-déterminisme.

Nous nous intéressons ici aux applications Internet nécessitant la manipulation de gros volumes de données et une forte interaction entre différents utilisateurs (par exemple, le commerce électronique).

Nous avons poursuivi cette année nos travaux sur les "vues actives" [12]. Il s'agit de pouvoir offrir des vues d'une base de données définies déclarativement et qui incluent des aspects actifs comme d'être notifié de certains évènements ou d'être averti quand certaines valeurs de la vue ont (peuvent avoir) été modifiées dans la base de données. Le but à long terme est d'offrir un tel outil sur le Web. Nous travaillons sur des modules comme : (i) le maintien de trace de l'activité du système interrogeable avec le langage de requête standard, (ii) un système de régles actives permettant au système de réagir à des notifications, (iii) la génération automatique de telles applications.

Nous considérons, en particulier, deux grandes applications. D'abord, le développement de catalogues de vente électronique sur le Web. Ceci fait l'objet du projet GAEL labellisé RNRT en commun avec le LRI (Marie-Christine Rousset) et MatchVision une start-up. Il s'agit de faciliter à des non-informaticiens le développement de boutiques sur le Web en s'appuyant sur des langages déclaratifs. La seconde concerne les systèmes de gestion d'information pour la fabrication. Ceci est l'objet d'un projet commun AFIRST avec l'Université de Tel Aviv (Tova Milo).

Un prototype du système AViews a été démontré l'an dernier à la conférence VLDB. Nous

avons poursuivi ces travaux notamment en réalisant un compilateur qui permet de générer de telles applications automatiquement à partir d'une spécification déclarative.

6 Contrats industriels (nationaux, européens et internationaux)

Xyleme La société Xyleme SA a été créée en Septembre 2000 en grande partie à partir de travaux du projet Verso.

MatchVision Nous collaborons avec MatchVision dans le cadre du projet RNRT GAEL, voir plus loin.

Informix Depuis sa création, nous collaborons avec la société Informix (rachat de O_2Technology par Ardent Software, puis Informix) qui développe et commercialise le système de gestion de bases de données objet O_2.

Fanny Wattez qui bénéficiait d'une bourse CIFRE ArdentSoftware a présenté le résultat de ses travaux dans la session industrielle de la dernière conférence Sigmod[40]. Sophie Cluet continue son action de conseil sur l'optimiseur de requêtes du système O_2 pour la société Informix.

7 Actions régionales, nationales et internationales

7.1 Actions nationales

7.1.1 Livre

La traduction [8] du livre d'Abiteboul, Vianu et Hull sur la théorie des bases de données ((Addison-Wesley) a été publiée par Vuibert, traduit de l'anglais par Patrick Cégielski.

7.1.2 Projet RNRT GAEL

Ce projet, qui réunit une start-up MatchVision, l'équipe de M.C. Rousset au LRI (U. Paris Sud) et le projet Verso, est entré dans sa seconde année.

L'objectif du projet est de concevoir un générateur de catalogues électroniques et de services associés, permettant à des non informaticiens de développer leurs propres catalogues (boutiques, galeries marchandes) électroniques sans gros effort de programmation et en les personnalisant à leurs besoins. Pour ce faire, on combine des techniques de représentation des connaissances et d'agents intelligents. L'effort de recherche porte principalement sur une spécification déclarative des connaissances relatives au commerce électronique (les différents types de produits, de profils clients, d'actes commerciaux), ainsi que des informations présentées aux clients en XML. Nous utilisons des agents intelligents pour doter les catalogues électroniques de comportements dynamiques permettant de réagir de façon adéquate aux achats en cours des clients en fonction du contenu du catalogue, de la stratégie commerciale spécifiée de façon déclarative par le vendeur, ainsi que du profil des clients.

Le responsable Verso pour GAEL est S. Abiteboul.

7.1.3 Autres Collaborations Nationales

Des liens étroits existent avec le Labri (N. Bidoit, B. Courcelle), le LRI (C. Delobel, E Waller, M.C. Rousset), le Cedric au CNAM (M. Scholl, B. Amann, P. Rigaux, D. Vodislav), et l'équipe Caravel de l'INRIA (F. Llirbat).

7.2 Actions financées par la commission européenne

7.2.1 Projet CWeb

Le projet CWeb était un projet européen qui comprenait comme partenaires l'INRIA (l'action de développement Mediaculture et les équipes Verso et Acacia), ICS/Forth et EDW, une PME italienne spécialisée dans les systèmes de gestion d'informations d'entreprise, fondés sur SGML/XML. Dans ce projet de recherche d'une durée très courte (un an), il s'agissait de définir une plate-forme générique et ouverte pour la gestion de Webs communautaires, ou de mémoire d'entreprise. Le projet successeur démarrera début Janvier 2001 sous le nom de MESMUSES (voir 7.2.2).

Le serveur de CWeb peut être consulté à `http://cweb.inria.fr`.

Le responsable Verso pour CWeb était A.M. Vercoustre.

7.2.2 Projet MESMUSES

Le Projet Européen MESMUSES (Metaphor for Science Museums) va démarrer début Janvier 2001. Il inclut tous les partenaires du projet CWeb (voir 7.2.1), l'ENST Bretagne, deux autres partenaires industriels (Valoris/Euroclid, Paris et Finsiel Multimedia Services, Italie) et deux musées scientifiques comme utilisateurs finaux (Cité des Sciences et de l'Industrie, Paris et Istituto e Museo di Storia della Scienza, Florence).

L'objectif est l'implantation de la plate-forme CWeb et l'installation de deux applications dans le contexte de la préparation d'expositions scientifiques. L'objectif principale est la réutilisation optimale de sources électroniques (documents, images, son, vidéo) créées pendant la préparation d'une exposition ou disponibles dans les médiathèques pour la création de sites Web interactifs destinés au grand public.

Le responsable Verso pour MESMUSES est B. Amann.

7.2.3 Projet TMR Chorochronos

Chorochronos est un programme d'échanges européen TMR (Training and Mobility of Researchers) sur les bases de données spatiales et temporelles. La collaboration entre les 10 nœuds du réseau porte sur les thèmes suivants : (i) structure et représentation de l'espace et du temps, (ii) modèles et langages pour les SGBD spatio-temporels (SGBDST), (iii) interfaces pour SGBDST, (iv) exécution et optimisation de requêtes, (v) structures de données et indexation spatio-temporelle, et (vi) architecture d'un SGBDST. Dans ce projet, Verso est le noeud français. Ses contributions essentielles, en collaboration avec le groupe bases de données Vertigo du Cedric au CNAM, sont l'étude de modèles et langages pour les données multidimensionnelles, l'optimisation de l'évaluation des langages et le développement du prototype Dedale pour manipuler des données multidimensionnelles. Un certain nombre d'applications

spatio-temporelles sont étudiées pour valider ces études. Le projet TMR Chorochronos s'est terminé en Juillet 2000.

Les responsables Verso pour Chorochronos sont S. Grumbach et M. Scholl.

7.2.4 GDR-PSIG Cassini

Cassini est un programme de recherche national pluridisciplinaire (informaticiens, géographes) financé par le CNRS et l'Institut Géographique National (IGN) sur l'information géographique. Dans ce programme, le projet Verso, en collaboration avec le CNAM, le LSR de l'IMAG et le LAMA à Grenoble, étudie les problèmes de modélisation d'objets évoluant au cours du temps.

Le responsable Verso pour Cassini est M. Scholl.

7.3 Réseaux et groupes de travail internationaux

Verso participe à Pastel, un groupe de travail financé par la communauté européenne dont l'objectif est la réalisation de systèmes pour des applications persistentes. Ce groupe de travail fait suite au projet Esprit B Fide2 auquel Verso participait.

Verso est membre du réseau d'excellence Compulog (logic programming) et DELOS (European Digital Libraries) et participe au groupe "Bases de données" de l'Ercim. Verso participe également à un working group d'Ercim sur la programmation avec contraintes en cours de création. A.M Vercoustre participe aussi au groupe de travail pour l'internalisation du Dublin Core. A.M.Vercoustre a représenté l'INRIA dans les groupes ERCIM pour les Bibliothèques Electroniques et a participé au projet DELOS pour la création de ETRDL (ERCIM Technical Reference Digital Library). Une démonstration a été présentée à Amsterdam, pour les 10 ans d'ERCIM, en novembre 99.

7.4 Relations bilatérales internationales

7.4.1 Europe

Nous collaborons avec l'Université de Mannheim (G. Moerkotte) et l'Université de Mainz (Thomas Schwentick).

7.4.2 Moyen-Orient

Financé par l'Association Franco-Israélienne pour la Recherche Scientifique et Technique (AFIRST), le projet "Les Usines du Futur" nous permet de collaborer étroitement avec l'Université de Tel Aviv. Le contrat porte sur l'utilisation des vues actives pour permettre le contrôle de fabrication (voir 5.4 et [12]) et nous a déjà permis de développer un prototype au-dessus du système AXIELLE développé par la société Ardent Software. Nous continuons à travailler sur ce prototype, notamment sur les aspects d'optimisation de requêtes.

7.4.3 Amérique du Nord

En Amérique du Nord, des travaux en commun sont en cours avec l'Université de Stanford (J. Widom), Pennsylvanie (P. Buneman), UC Santa Barba (J. Su) [19], UC San Diego (V. Vianu), ATT (Sihem Amer-Yahia, Divesh Srivastava, Dan Suciu), Lucent-Bell (Jérôme Siméon).

7.4.4 Asie et océan Pacifique

Après plusieures années de collaboration avec le CSIRO à Melbourne, A.-M. Vercoustre est partie pour 2 ans travailler dans cette équipe.

7.5 Accueil de chercheurs étrangers

Cette année, nous avons accueilli :

- Leonid Libkin, chercheur à Bells Labs, USA (9 mois).
- Catriel Beeri, professeur, U. jerusalem, (3 mois)
- Victor Vianu, professeur, UC San Diego (4 mois)

Nous avons également accueilli pour de courtes visites d'autres chercheurs et professeurs étrangers avec lesquels nous avons des collaborations suivies : Vassilis Christophides (chercheur, FORTH-Crète, 3 semaines), Tova Milo (professeur, U. Tel Aviv, 15 jours), Torsten Schlieder, U. Berlin, 20 jours), Agnès Voisard (1 mois).

8 Diffusion de résultats

8.1 Actions d'enseignement

S. Abiteboul est professeur à temps partiel à l'Ecole Polytechnique.
B. Amann est maître de conférence au CNAM-Paris.
C. Delobel est professeur à l'Université de Paris 11.
M. Scholl est professeur au CNAM.
V. Vianu est professeur à UCSD.

Les cours suivants ont été assurés par plusieurs membres de l'équipe.

SGBD relationnels,
CNAM-Paris, Cycles A et B, B. Amann
MIAGE et nouvelle formation d'ingénieurs, Paris 11, C. Delobel.

SGBD objets et avancés,
Polytechnique, S. Abiteboul;
CNAM-Paris, M. Scholl et B. Amann;
DEA Systèmes Informatiques, cohabilité Paris 6-Telecom-CNAM, B. Amann, M. Scholl.

Bases de données avec contraintes, DEA I3 de Paris-Sud, et Université de Calabre, S. Grumbach.

Le standard XML, DEA Informatique de Paris-Dauphine, S. Cluet.

Bases de données semistructurées, DEA I3 de Paris XI, S. Cluet et S. Abiteboul.

Algorithmique de la décision, Polytechnique, S. Abiteboul.

Théorie des modèles finis, DEA de Paris VII, S. Abiteboul et S. Grumbach

Initiation aux Bases de Données Relationnelles, Module de première année, Ecole Nationale des Ponts et Chaussées, V. Aguiléra

XML et Données Semistructurées, Troisième année, voie d'approfondissement Informatique, Ecole Nationale des Travaux Publics de l'Etat, V. Aguiléra

Bases de données, Travaux dirigés, Maîtrise d'informatique, Université de Versailles, F. Wattez

Base de données, Travaux dirigés, Cycle B, CNAM Paris, L. Mignet, I. Fundulaki

Base de données, Travaux dirigés, Première Année, Institut d'Informatique d'Entreprise - CNAM, I.Fundulaki

Systèmes d'Information Automatisés, Travaux dirigés, Seconde Année, Institut d'Informatique d'Entreprise - CNAM, I.Fundulaki

SGBD sous Oracle, Travaux Pratiques, Seconde Année d'ingénieur, EPF, L. Mignet

Informatique, Travaux dirigés et Méthodologie, Première Année (MIAS), Institut d'informatique Galilée - Université de Paris XIII, Villetaneuse, P. Veltri

8.2 Participation à des colloques

L'équipe a eu de nombreuses publications dans des conférences internationales et des colloques (voir la bibliographie). Enfin, certains membres du projet ont participé à des comités de programmes. La liste en est donnée ci-dessous.

S. Abiteboul

- 2000 : ACM-SIGMOD International Conference on the Management of Data (SIGMOD2000). ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS2000). NEC Research Symposium 2000, San Francisco (NEC2000). International Workshop on Web Dynamics

S. Cluet

- 2000 : SIGMOD (ACM Conference on the Management Of Data). DOOD (Deductive and Object Oriented Databases), WWW 2001 (World Wide Web International Conference)

S. Grumbach

- 2000 : WAIM'00 (International Conference on Web-Age Information Management). steering committee ASIAN'00 (Asian Computing Science Conference). ASDM'00 (International Workshop On Advanced Spatial Data Management). ICADL'00 (3rd International Conference of Asian Digital Library). WISDM'00 (Workshop on Information Systems and Data Management with High-speed Networks).
- 2001 : SSDBM'01 (Thirteenth International Conference on Scientific and Statistical Database Management).

L. Libkin

- 2000 : DaWaK (International Conference on Data Warehousing and Knowledege Discovery), LPAR'00 (Logic for programming and automated reasoning), PODS'00 (ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems). LICS'00 (IEEE Symposium on Logic in Computer Science). CIKM'00 (International Conference on Information and Knowledge Management). FOIKS'00 (Foundations of Information and Knowledge Systems).

M. Scholl

- 2000 : ER00 (Entity Relation International Conference) ASDM00 (Advanced spatial Database Management International workshop) ECDL2000 semantic web workshop CAISE'01 (11th Conference on Advanced Information Systems), Sigmod'01 (ACM conference on the management of Data), COSIT'01 (Conference on Spatial Information Theory),

L. Segoufin

- 2001 : ICDT'01 (International Conference on Database Theory), PODS'01 (ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems).

A.-M. Vercoustre

- 2000 : WWW9 (9th Internationnal World Wide Web Conference), HT2000 (ACM Conference on Hypertext and Hypermedia), ECDL2000 (European Conference on Digtal Libraries).

V. Vianu

- 2000 : International Conference on Very Large Data Bases (VLDB), Workshop on Deductive Object-Oriented Databases (DOOD).

8.2.1 Conférences invitées, tutoriels, cours, etc.

S. Abiteboul et A.-M. Vercoustre ont co-édité un numéro spécial du *International on Digital Libraries* [9] à partir d'articles de ECDL99.

En 2000, S. Abiteboul est devenu General Chair de ACM International Symposium on Principles of Database Systems, et membre des comités exécutifs de Extended Database Technology Endowment et ACM SIGMOD executive committee.

S. Cluet a donné un cours à l'université de tous les savoirs sur le thème : *le Web, du texte à la connaissance*[18].

L. Libkin est éditeur de "SIGMOD Record", et éditeur invité de "ACM Transactions on Computational Logic" pour un numéro spécial de LICS'00.

M. Scholl a présenté une conférence invitée sur les Bases de Données contraintes au Workshop ERCIM Constraint programming, Juin, Padova.

V. Vianu a présenté plusieurs conférences invitées sur les requêtes topologiques dans les bases de données topologiques géographiques : au Plenary Session of the Computational Logic Conférence 2000, London et à l'Université de Washington, Seattle.

V. Vianu est éditeur d'une colonne sur la théorie des bases de données dans SIGACT News. Il est membre du comité de lecture du "Journal of Discrete Mathematics and Theorical Computer Science", du "SIGMOD Digital Reviews", ainsi que du "ACM Transactions on Computational Logic".

8.2.2 Animations scientifiques

S. Cluet a été nommée au conseil d'administration du VLDB Endowment (Very Large Databases), un organisme basé aux Etats-Unis et dont la vocation est la promotion à travers le monde de la recherche en base de données.

S. Abiteboul est membre du comité de coordination des STIC (Ministère de la Recherche).

M. Scholl est membre de la commission d'évaluation du RNTL et expert auprès de la mission scientifique universitaire du MENRT

L. Libkin est membre du comité d'organisation de la conférence Logic in Computer Science (LICS).

V. Vianu est membre des Comités Executifs de PODS et de ICDT. Il a été General Chair de PODS et membre du Comité Exécutif de SIGMOD jusqu'en Juin 2000.

9 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] S. Abiteboul, P. Buneman, D. Suciu, *Data on the Web: From Relations to Semistructured Data and XML*, Morgan-Kaufman, New York, 1999.

[2] S. Abiteboul, R. Hull, V. Vianu, *Foundations of Databases*, Addison-Wesley, 1995.

[3] S. Abiteboul, V. Vianu, « Computing With First-Order Logic », *Journal of Computer and System Sciences(JCSS) 50(2)*, 1995, p. 309–335.

[4] S. Grumbach, P. Rigaux, L. Segoufin, «The DEDALE System for Complex Spatial Queries», *in : sigmod*, 1998.

[5] S. Grumbach, J. Su, « Finitely representable databases. », *Journal of Computer and System Sciences(JCSS) 55*, 2, 1999, p. 273–298.

[6] S. Grumbach, C. Tollu, « On the Expressive Power of Counting », *Journal of Theoretical Computer Science (TCS)*, 1995.

[7] M. Scholl, A. Voisard, J.-P. Peloux, L. Raynal, P. Rigaux, *SGBD Géographiques*, International Thomson Publishing, 1996.

Livres et monographies

[8] S. Abiteboul, R. Hull, V. Vianu, *Fondements des bases de données*, Vuibert, Paris, 2000.

[9] S. Abiteboul, A.-M. Vercoustre (éditeurs), *International on Digital Libraries*, *3*, 3, Springer, 2000.

[10] G. Kuper, L. Libkin, J. Paredaens, *Constraint Databases*, Springer-Verlag, 2000.

Thèses et habilitations à diriger des recherches

[11] F. Wattez, *Optimisation de requêtes sur des structures d'arbre : application à l'objet et à XML*, thèse de doctorat, Université Paris 1, janvier 2001.

Articles et chapitres de livre

[12] S. Abiteboul, B. Amann, S. Cluet, L. Mignet, T. Milo, « Declarative Specification of Electronic Commerce Applications», *IEEE Data Enginerring Bulletin 23*, 1, 2000, p. 37–42.

[13] S. Abiteboul, «Gestion de données pour le web», *Technique et Science de l'Informatique 19*, 1, 2000.

[14] B. Amann, I. Fundulaki, M. Scholl, «Integrating ontologies and thesauri for RDF schema creation and metadata querying», *Intl. Journal of Digital Libraries (JODL) 3*, 3, October 2000.

[15] M. Benedikt, L. Libkin, « Expressive Power: the Finite Case», *in : Constraint Databases*, G. Kuper, L. Libkin, et J. Paredaens (éditeurs), Springer-Verlag, 2000.

[16] M. Benedikt, L. Libkin, «Query Safety and Constraints», *in : Constraint Databases*, G. Kuper, L. Libkin, et J. Paredaens (éditeurs), Springer-Verlag, 2000.

[17] J. Chomicki, L. Libkin, « Aggregate Languages for Constraint Databases », *in : Constraint Databases*, G. Kuper, L. Libkin, et J. Paredaens (éditeurs), Springer-Verlag, 2000.

[18] S. Cluet, *Université de tous les savoirs*, Odile Jacob, 2001, ch. Le Web : du texte à la connaissance, A paraître.

[19] S. Grumbach, G. Kuper, J. Su, «Expressive power in the infinite», *in : Contraints Databases*, G. Kuper, L. Libkin, et J. Paredaens (éditeurs), Springer-Verlag, 2000.

[20] S. Grumbach, Z. Lacroix, P. Rigaux, L. Segoufin, «Query Evaluation and Optimization», *in : Contraints Databases*, G. Kuper, L. Libkin, et J. Paredaens" (éditeurs), Springer-Verlag, 2000.

[21] S. Grumbach, P. Rigaux, M. Scholl, L. Segoufin, « The Design and Implementation of a Constraint-Based System: DEDALE », *in : Contraints Databases*, G. Kuper, L. Libkin, et J. Paredaens (éditeurs), Springer-Verlag, 2000.

[22] S. Grumbach, P. Rigaux, L. Segoufin, «Spatio-Temporal Data Handling with Constraints», *GeoInformatica*, 2000, à paraître.

[23] S. Grumbach, L. Tininini, « On the Content of Materialized Aggregate Views », *Journal of Computer and System Sciences*, 2000, à paraître.

[24] B. Kuijpers, V. Vianu, «Topological Queries», *in : Contraints Databases*, G. Kuper, L. Libkin, et J. Paredaens (éditeurs), Springer-Verlag, 2000.

[25] G. Kuper, L. Libkin, J. Paredaens, « Introduction », *in : Constraint Databases*, G. Kuper, L. Libkin, et J. Paredaens (éditeurs), Springer-Verlag, 2000.

[26] G. Kuper, M. Scholl, «Geographic Information Systems», *in : Contraints Databases*, G. Kuper, L. Libkin, et J. Paredaens (éditeurs), Springer-Verlag, 2000.

[27] L. Segoufin, V. Vianu, « Querying Spatial Databases via Topological Invariants », *Journal of Computer and System Sciences*, 2000, à paraître.

Communications à des congrès, colloques, etc.

[28] S. Abiteboul, H. Kaplan, T. Milo, « Compact labeling schemes for ancestor queries », *in : Twelfth ACM-SIAM Symposium on Discrete Algorithms*, 2001.

[29] V. Aguiléra, S. Cluet, P. Veltri, D. Vodislav, F. Wattez, « Querying XML Documents in Xyleme », *in : Working Notes of the ACM-SIGIR Worshop on XML and Information Retrieval*, 2000, `http://www.haifa.il.ibm.com/sigir00-xml/final-papers/xyleme/XylemeQuery/XylemeQuery.html`.

[30] S. Alexaki, V. Christophides, G. Karvounarakis, D. Plexousakis, K. Tolle, B. Amann, I. Fundulaki, M. Scholl, A. Vercoustre, « Managing RDF Metadata for Community Webs», *in : Workshop on the Web and Conceptual Modeling (WCM'2000)*, Salt Lake City, Utah, novembre 2000.

[31] M. Benedikt, M. Grohe, L. Libkin, L. Segoufin, «Reachability and Connectivity Queries in Constraint Databases», *in : Proceedings of ACM Principle of Database Systems (PODS)*, 2000.

[32] V. Christophides, S. Cluet, J. Siméon, «On Wrapping Query Languages and Efficient XML Integration», *in : SIGMOD*, Mai 2000.

[33] M. Grohe, L. Segoufin, « On first-order topological queries », *in : Logic in Computer Science (LICS)*, 2000.

[34] S. Grumbach, P. Rigaux, L. Segoufin, «Manipulating Interpolated Data is Easier than You Thought.», *in : Int. Conf. on Very Large DataBases (VLDB)*, 2000.

[35] S. Grumbach, P. Rigaux, P. Veltri, « Hierarchical Optimization of Linear Constraint Processing», *in : Sistemi Evoluti per Basi di Dati (SEBD)*, 2000.

[36] S. Grumbach, L. Tininini, « Automatic Aggregation using Explicit Metadata », *in : International Conference on Scientific and Statistical Database Management*, Berlin, 2000.

[37] S. Grumbach, L. Tininini, « On the Content of Materialized Aggregate Views », *in : PODS*, 2000.

[38] A. Marian, S. Abiteboul, L. Mignet, « Chance-centric management of versions in an XML Warehouse», *in : Conférence sur les Bases de Données Avancées, Blois 2000*, 2000.

[39] L. Mignet, S. Abiteboul, S. Ailleret, B. Amann, A. Marian, M. Preda, « Acquiring XML pages for a WebHouse », *in : Conférence sur les Bases de Données Avancées, Blois 2000*, 2000.

[40] F. Wattez, S. Cluet, V. Benzaken, G. Ferran, C. Fiegel, « Benchmarking Tree Queries: Learning the Hard Truth the Hard Way », *in : SIGMOD*, Mai 2000. session industrielle.

Divers

[41] S. Abiteboul, V. Vianu, L. Segoufin, « Representing and Querying XML with Incomplete Information », 2000.

[42] V. Aguiléra, S. Cluet, F. Wattez, «Querying The XML Documents Of The Web», submitted to The 10th International World Wide Web Conference.

[43] S. Cluet, P. Veltri, D. Vodislav, « Views in a Large Scale XML Repository ».

[44] S. Lafois, *Interrogation de Données Structurées en Arbre: Application à C-Web*, Mémoire, DEA SI, Université Paris VI, Paris, septembre 2000.