Bases de Données et Web
Enjeux, problèmes et directions de recherche

Patrick Valduriez
Inria, France

Patrick.Valduriez@inria.fr

Le World Wide Web et son infrastructure Internet constituent un gigantesque système distribué, de dimension mondiale et en croissance constante. Conçu initialement pour accéder des documents hypertextes, faiblement structurés, le Web évolue rapidement pour supporter les nouveaux systèmes d'information distribués (autoroutes de l'information et Intranet), caractérisés par de très nombreuses sources de données, à la fois plus structurées et très hétérogènes.

La nature dynamique du Web et sa dimension croissante rendent la recherche d'informations pertinentes de plus en plus difficile. D'autre part, les outils de recherche d'information existants (knowbots) restent limités aux documents et n'exploitent pas la structure des données. Une voie prometteuse consiste à s'inspirer des techniques développées dans le contexte des bases de données distribuées et les étendre pour prendre en compte les caractéristiques du Web (grand nombre de sources de données, hétérogénéïté, dynamique, etc.).

Pour illustrer l'apport potentiel des techniques bases de données au Web, je rappelle rapidement les éléments essentiels ayant fait le succès du Web. Puis je résume les objectifs et les principes des bases de données réparties en montrant leur apport potentiel pour le Web. Enfin, pour illustrer une direction importante de recherche, je décris les objectifs du projet DISCO à l'Inria.

1. Le Web

Le Web (ou World Wide Web) est un système hypertexte distribué, conçu initialement pour faciliter la coopération internationale des scientifiques en exploitant Internet, le réseau des réseaux. Très tôt, le Web a connu un grand succès grâce à l'interface graphique de ses navigateurs (par exemple, Mosaic et Netscape) qui simplifie la navigation dans les documents distribués. Aujourd'hui, l'interface Web est de plus en plus multi-média et interactive, notamment grâce à des langages dynamiques comme Java.

L'architecture du Web est client-serveur avec une interface bien définie entre client et serveur (HTTP). Cette interface est surtout indépendante de toute plate-forme système et protocole et répond enfin au problème de portabilité des interfaces graphiques des clients. Ainsi, le Web peut fonctionner aussi bien sur Internet que sur des réseaux privés internes aux entreprises (Intranet). Bien sûr, un serveur Web peut être client d'un autre serveur Web distant et ainsi de suite, de sorte qu'un accès au Web peut conduire à une navigation très complexe à retracer. Du point de vue technologie, le Web repose sur l'hypertexte et la recherche d'information textuelle. Contrairement à la recherche de données structurées dans les bases de données qui est précise (par exemple, quels sont les noms et adresses des employés entre 40 et 45 ans), la recherche d'information textuelle est plus flexible et doit

prendre en compte des critères de proximité sémantique (par ex., quels sont les articles récents qui parlent du Web ?).

La dimension mondiale, sans frontière, du Web, conjuguée à sa facilité d'usage et à sa portabilité sont à la base de la future société de l'information. Le Web se développe de façon exponentielle (plus de 100 000 serveurs estimés en 1995 [Cameron, 1996]) et ne cesse de croître en diversité d'applications (éducation, publication en ligne, commerce électronique, etc.) et d'utilisateurs (individuels, entreprises, gouvernements, etc.). Il est devenu incontournable pour toute organisation, publique, privée, qui vise le développement et la croissance. C'est véritablement un phénomène de société mondiale qui est posé, non sans soulever d'importantes questions politiques.

Le Web se développe si rapidement que de nombreux problèmes techniques deviennent exacerbés, en particulier les aspects sécurité et recherche efficace d'informations [Manber, 1996]. La recherche d'informations pertinentes est de plus en plus difficile. D'autre part, les outils de recherche d'information existants (par ex. knowbots comme Alta Vista et Yahoo) restent limités aux documents et n'exploitent pas la structure des données. Enfin, ils ne peuvent pas être déployés pour accéder rapidement de nombreuses sources de données.

2. Apport des techniques bases de données réparties

Le Web a évolué depuis la technologie hypertexte, sans aucune contribution de la part des bases de données. Dans la mesure où il s'agit d'un enjeu commercial important, la plupart des fournisseurs de bases de données (Oracle en tête) offrent aujourd'hui des solutions au stockage des données Web (pages HTML) et à l'interopérabilité avec les navigateurs. Certains systèmes vont jusqu'à intégrer des capacités de recherche d'information textuelle. Mais ces solutions n'apportent pas de réponses aux problèmes durs que rencontre le Web pour pouvoir évoluer.

La technologie des bases de données réparties telle qu'elle a été développée par les chercheurs dans les années 1980 [Özsu et Valduriez, 1991] peut apporter des solutions intéressantes. La raison est que cette technologie a été conçue au départ pour les réseaux longues distances alors que les produits se sont limités aux réseaux locaux et au client-serveur (éventuellement avec quelques serveurs). En d'autres termes, les principes des bases de données réparties conçus pour des grands systèmes correspondent bien au contexte du Web, notamment en généralisant l'architecture client-serveur traditionnelle.

Le principe le plus important est celui de transparence ou d'indépendance à la localisation des données qui est mis en place grâce à un catalogue de la base de données réparties. Une conséquence de ce principe est de pouvoir administrer et manipuler une base de données répartie, constituée de multiple bases de données, assez simplement avec un langage de haut niveau et portable (par ex., SQL).

Les bases de données réparties peuvent répondre au probleme de requêtes sur des sources de données hétérogènes, en étendant les techniques de catalogue de données structurées et d'optimisation de requêtes. Un objectif est de pouvoir accéder différentes sources de données, qu'elles soient structurées ou non, avec un seul langage. Evidemment, le contexte du Web demande des extensions significatives à ces techniques, mais les principes de base devraient être préservés. D'autres apports possibles concernent la sécurité et l'intégrité des données accédées à distance.

3. Le projet DISCO

Le projet DISCO (Distributed Information Search COmponent) [Tomasic et al., 1996] développé à l'Inria s'inscrit dans cette direction. L'objectif est d'adresser les problèmes dûs a l'expansion rapide du Web. D'abord, il est difficile de savoir quelles sont les données pertinentes et leur localisation. Ensuite, même si les sources de données pertinentes sont identifiées, il reste difficile d'accéder et d'intégrer rapidement des données hétérogènes sur un réseau global à partir de critères de recherche complexes.

Pour uniformiser la représentation des données, DISCO met en oeuvre une approche objets répartis [Özsu et al., 1993] et s'appuie sur le modèle standard de l'ODMG en étendant les langages ODL et OQL. DISCO adopte l'architecture récente des systèmes de bases de données hétérogènes qui évite la gestion trop rigide d'un schéma global. Cette architecture repose sur trois types de composants : les Médiateurs qui encapsulent la représentation des sources de données, les Traducteurs qui convertissent les requêtes sur les sources de données locales, et les Catalogues qui répertorient tous les composants du système et leur localisation. Un médiateur est un serveur pour les applications ou d'autres médiateurs.

Pour faciliter l'ajout de sources de données dans un médiateur, DISCO modélise les connexions aux sources de données par des objets. Les médiateurs gèrent aussi des répertoires de méta-données et d'index qui sont utilisés pour optimiser l'accès aux informations. Le traitement des multi-sources de données dans DISCO repose sur des techniques de reformulation et d'optimisation de requêtes [Florescu et al., 1995], et prend en compte l'indisponibilité des sources données pendant l'exécution.

DISCO est en cours de réalisation en Java. Un premier prototype a été réalisé pour accéder diverses sources de données (BibTex, Wais, O2). Un objectif important est de déployer DISCO dans une application de bourse d'affaires électronique pour des sociétés de Bâtiment et Travaux Publics et dans une application environnementale de qualité de l'eau. Ces applications demandent l'accès à de nombreuses sources de données hétérogènes.

Références

[Cameron, 1996] D. Cameron : *The World Wide Web : strategies and opportunities for business*. Computer Technology Research Corporation, Charleston, SouthCarolina, 1996.

[Florescu et al., 1995] D. Florescu, L. Raschid, P. Valduriez : "Using Heterogeneous Equivalences for Query Simplification in MDBMS", *Int. Conf. on Cooperative Information Systems*, Vienna, May 1995.

[Manber, 1996] U. Manber : "Future Directions and Research Problems on the World Wide Web", *ACM Int. Conf. on Principles Of Database Systems (PODS)*, Montreal, Juin 1996.

[Özsu et al., 1993] T. Özsu, U. Dayal, P. Valduriez (eds.) : *Distributed Object Management*. Morgan Kaufmann, 1993.

[Özsu et Valduriez, 1991] T. Özsu, P. Valduriez : *Principles of Distributed Database Systems*. Prentice Hall, Englewood Cliffs, New Jersey, 1991.

[Tomasic et al., 1996] A. Tomasic, L. Raschid, P. Valduriez : "Scaling Heterogeneous Databases and the Design of DISCO", *Int. Conf. of Distributed Computing Systems*, Hong-Kong, Mai 1996.