UNIVERSITE PARIS-SUD

ÉCOLE DOCTORALE INFORMATIQUE DE PARIS-SUD (ED 427)

Laboratoire de Recherche en Informatique (LRI)

***DISCIPLINE** INFORMATIQUE*

THÈSE DE DOCTORAT

presentée en vue d'obtention du titre de docteur

par **Asterios KATSIFODIMOS**

Techniques efficaces basées sur des vues matérialisées pour la gestion des données du Web : algorithmes et systèmes

Directeur de thèse :	Ioana Manolescu	Inria Saclay and Université de Paris-Sud
Composition du jury :		
Rapporteurs :	Yanlei Diao	University of Massachusetts Amherst, U.S.A.
	Philippe Rigaux	Conservatoire National des Arts et Métiers
Examinateurs :	Alain Denise	Université de Paris-Sud
	Patrick Valduriez	Inria Sophia Antipolis
	Vasilis Vassalos	Athens University of Economics and Business

Résumé

"Techniques efficaces basées sur des vues matérialisées pour la gestion des données du Web : algorithmes et systèmes"

Asterios Katsifodimos

Le langage XML, proposé par le W3C, est aujourd'hui utilisé comme un modèle de données pour le stockage et l'interrogation de grands volumes de données dans les systèmes de bases de données. En dépit d'importants travaux de recherche et le développement de systèmes efficace, le traitement de grands volumes de données XML pose encore des problèmes des performance dus à la complexité et hétérogénéité des données ainsi qu'à la complexité des langages courants d'interrogation XML.

Les vues matérialisées sont employées depuis des décennies dans les bases de données afin de raccourcir les temps de traitement des requêtes. Elles peuvent être considérées les résultats de requêtes pré-calculées, que l'on réutilise afin d'éviter de recalculer (complètement ou partiellement) une nouvelle requête. Les vues matérialisées ont fait l'objet de nombreuses recherches, en particulier dans le contexte des entrepôts des données relationnelles.

Cette thèse étudie l'applicabilité de techniques de vues matérialisées pour optimiser les performances des systèmes de gestion de données Web, et en particulier XML, dans des environnements distribués.

Mots Clefs : XML, données du web, vues matérialisées, optimisation des requêtes, selection des vues, systèmes d'abonnements, gestion des données.

Table des matières

Chapitre 1

Motivation

Les vues matérialisées sont utilisées depuis longtemps dans les systèmes de bases de données afin d'accélérer les requêtes de base de données [Hal01]. Les vues matérialisées peuvent être considérées comme des résultats des requêtes précalculées qui peuvent être réutilisés pour évaluer une autre requête, et ont été un sujet de recherche dans la communauté de base de données, en particulier dans le contexte de l'entreposage de données relationnelle [GM99]. Dans cette thèse, nous étudions l'applicabilité des techniques basées sur *vues matérialisées* pour optimiser les performances des systèmes de gestion de données du Web. Plus précisément, nous considérons des données XML et des requêtes dans des systèmes distribués.

Dans le cadre de la gestion des données XML, les systèmes distribués sont d'intérêt important pour deux raisons. Tout d'abord, les organisations interagissent de plus en plus, le partage et l'information de la consommation d'une autre, il est souvent le cas que les données (XML) est produit indépendamment par plusieurs sources distribuées. Deuxièmement, un système distribué peut accueillir volumes de données allant bien au-delà de la capacité d'une seule machine ou un cluster.

Le travail présenté dans cette thèse vise à montrer que les vues matérialisées sur les données XML peuvent être utilisés, en particulier dans les systèmes distribués, pour permettre un partage efficace et l'interrogation de grands volumes de données du Web.

Chapitre 2

Sélection des vues matérialisées pour XQuery

Dans ce chapitre, nous considérons le problème de la sélection des meilleures vues à matérialiser dans un espace de stockage donné, afin d'améliorer la performance d'une charge de travail des requêtes. Nous sommes les premiers à considérer un sous-langage de XQuery enrichi avec la possibilité de sélectionner des nœuds multiples et à de multiples niveaux de granularités. La difficulté dans ce contexte vient de la puissance expressive et des caractéristiques du langage des requêtes et des vues, et de la taille de l'espace de recherche de vues que l'on pourrait matérialiser. Alors que le problème général a une complexité prohibitive, nous proposons et étudions un algorithme heuristique et démontrer ses performances supérieures par rapport à l'état de l'art.

Les contributions de ce chapitre sont les suivantes :

— Le travail présenté dans ce chapitre a été le premier à formaliser le problème de sélection pour les requêtes et les vues exprimées dans un riche sous-ensemble de XQuery.

— Nous analysons l'espace des vues candidats potentiels, et nous presentons plusieurs critères de l'elimination efficace de candidats.

— Nous proposons et étudions un algorithme heuristique et de démontrer expérimentalement sa supériorité par rapport à l'état de l'art.

Les travaux décrits dans ce chapitre sont publié dans [KMV12], tandis qu'une version antérieure a été démontré dans [CRKMR10].

Chapitre 3

Gestion de grands corpus XML dans des réseaux pair à pair

Dans ce chapitre, nous considérons la gestion de grands corpus XML dans des réseaux pair à pair, basées sur des tables de hachage distribuées. Nous considérons la plateforme ViP2P dans laquelle des vues XML distribuées sont matérialisées à partir des données publiées dans le réseau, puis exploitées pour répondre efficacement aux requêtes émises par un pair du réseau. Nous y avons apporté d'importantes optimisations orientées sur le passage à l'échelle, et nous avons caractérisé la performance du système par une série d'expériences déployées dans un réseau à grande échelle. Ces expériences dépassent de plusieurs ordres de grandeur les systèmes similaires en termes de volumes de données et de débit de dissémination des données. Cette étude est à ce jour la plus complète concernant une plateforme de gestion de contenus XML déployée entièrement et testée à une échelle réelle.

Les contributions de ce chapitre peuvent être résumée comme suit :

— Nous présentons une architecture complète pour l'évaluation des requêtes, en mode abonnement et en mode instantané. Cette architecture permet la diffusion efficace de réponses aux requêtes de motifs d'arbres (exprimé dans un dialecte XQuery) à des pairs qui sont intéressés en données publiées par d'autres pairs.

— Nous avons mis en œuvre notre architecture, fondé sur FreePastry [Fre],

une infrastructure P2P et nous presentons un ensemble complet d'expériences réalisées dans un WAN, qui démontre la supériorité de ViP2P sur l'état de l'art.

Ce chapitre est une extension des travaux publiés dans [KKMZ12] et [KKMZ11].

Chapitre 4

Système d'abonnements basées sur des vues matérialisées

Dans le dernier chapitre ous présentons une nouvelle approche de dissémination de données dans un système d'abonnements, en présence de contraintes sur les ressources CPU et réseau disponibles ; cette approche est mise en oeuvre dans le cadre de notre plateforme Delta. Le passage à l'échelle est obtenu en déchargeant le fournisseur de données de l'effort de répondre à une partie des abonnements. Pour cela, nous tirons profit de techniques de réécriture de requêtes à l'aide de vues afin de diffuser les données de ces abonnements, à partir d'autres abonnements. Notre contribution principale est un nouvel algorithme qui organise les vues dans un réseau de dissémination d'information multi-niveaux ; ce réseau est calculé à l'aide d'outils techniques de programmation linéaire afin de passer à l'échelle pour de grands nombres de vues, respecter les contraintes de capacité du système, et minimiser les délais de propagation des information. L'efficacité et la performance de notre algorithme est confirmée par notre évaluation expérimentale, qui inclut l'étude d'un déploiement réel dans un réseau WAN.

Les résultats de ce chapitre font partie d'un article soumis pour publication le 1er mai 2013 et qui est actuellement en cours de révision.

Bibliographie

[CRKMR10] Jesús Camacho-Rodríguez, Asterios Katsifodimos, Ioana Manolescu, and Alexandra Roatis. LiquidXML : Adaptive XML Content Redistribution. In *CIKM (demo)*, 2010.

[Fre] Freepastry, an open-source implementation of pastry. http ://freepastry.org/FreePastry/.

[GM99] Ashish Gupta and Inderpal Singh Mumick, editors. *Materialized Views : Techniques, Implementations, and Applications*. The MIT Press, 1999.

[Hal01] Alon Y. Halevy. Answering queries using views : A survey. *VLDB J.*, 10(4), 2001.

[KKMZ11] Konstantinos Karanasos, Asterios Katsifodimos, Ioana Manolescu, and Spyros Zoupanos. The ViP2P Platform : XML Views in P2P. Technical Report RR-7812, INRIA, November 2011.

[KKMZ12] Konstantinos Karanasos, Asterios Katsifodimos, Ioana Manolescu, and Spyros Zoupanos. ViP2P : Efficient XML management in DHT networks. In *ICWE*, 2012.

[KMV12] Asterios Katsifodimos, Ioana Manolescu, and Vasilis Vassalos. Materialized View Selection for XQuery Workloads. In *SIGMOD*, 2012.