Analyse statique et typage de données semi-structurées XML

Dario Colazzo, LAMSADE
U. Paris Dauphine
dario.colazzo@dauphine.fr

Federico Ulliana, Équipe GrapIK
U. Montpellier 2 - LIRMM - INRIA
federico.ulliana@univ-montp2.fr

Contexte

XML est devenu le format standard d'échange et de représentation des données du Web. La vitesse de croissance en terme de taille des données XML publiés dans le Web est désormais vertigineuse, ce qui rende le traitement des données avec les systèmes traditionnels souvent impossible. L'optimisation par analyse statique consiste à raisonner sur l'efficacité et la correction des expressions XML avant de les évaluer, en visant les meilleures stratégies d'exécution. Les benefices de cette approche augmentent avec le volume des données à traiter, dont le fort intérêt à l'exploiter. L'étude des méthodes d'analyse statique exactes et approximées reste l'enjeu principal pour concevoir des nouveaux modules d'optimisation qui intègrent les moteurs d'évaluation de requêtes et mises à jour XML existents.

Sujet

Le typage est une technique classique d'optimisation statique des systèmes de gestion des données, qui consiste à raisonner sur une representation abstraite des données accèdes par une expression [1]. Le typage implique toujours un compromis sur la *précision* de la méthode, afin de mitiger des coûts d'analyse qui montent assez rapidement en complexité pour des langages expressifs, et garder des techniques polynomiaux. Aucune des approches pour XML dans la littérature ne permet pourtant de *régler* la précision de l'inférence par rapport à son coût, et aux ressources du système dont elle tourne.

Ce stage propose l'étude d'un système d'analyse statique qui réalise cet objectif, en combinant une notion de *approximation variable de type* via α-chaînes avec celle de probabilité. Une α-chaîne c'est une chaîne de types de longueur α, qui approxime un ensemble (potentiellement infini) de chemins d'accès aux données dans un arbre XML [2]. Plus la valeur de α est petite, plus l'approximation des données accèdes est importante, et moins cher devient le coût d'execution. En particulier, la complexité de cette inference est un polynôme dont le degré depend de α. Les schémas probabilistes [3], qui considèrent en même temps la structure et la distribution des données, permettent d'affiner ultérieurement l'analyse en fonction de l'instance sur laquelle elle est exécutée, en estimant l'impact de l'α-approximation.

Mesurer i) la précision et ii) le cout de la technique, permettra de définir une analyse statique pour l'optimisation de l'évaluation de requêtes et mises à jour sur des vues matérialisées XML qui peut se régler dynamiquement en fonction d'un modèle de coût qui prend en compte les ressources disponibles, la distribution des données, et la charge du travail du système. L'étude se concentrera sur un fragment simple du langage XPath avec quatre opérateurs de navigation (/,//,[],*), et la possibilité de rajouter des données uniquement. Un approfondissement du point de vue formel, ou pratique (en terme d'implementation), suivra selon les intérêts de l'étudiant.

La direction scientifique du stage sera menée en collaboration entre l'Université Paris Dauphine et l'Université de Montpellier 2.

Résultats attendus

— formalisation de l'inférence de type avec α-chaînes sur schémas probabilistes
— definition des mesure de precision/coût de la technique, et modèle pour le réglage automatique de l'analyse
— (soit) preuves formelles de correction du système (soit) prototype experimental et tests

Références

[1] Véronique Benzaken and Giuseppe Castagna and Dario Colazzo and Kim Nguyen. *Type-Based XML Projection*. VLDB 2006.

[2] Nicole Bidoit-Tollu and Dario Colazzo and Federico Ulliana. *Type-Based Detection of XML Query-Update Independence*. PVLDB 2012.

[3] Michael Benedikt and Evgeny Kharlamov and Dan Olteanu and Pierre Senellart. *Probabilistic XML via Markov Chains*. PVLDB 2010.