UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE INFORMATIQUE DE PARIS-SUD (ED 427)
Laboratoire de Recherche en Informatique (LRI)

DISCIPLINE Informatique

THÈSE DE DOCTORAT

soutenue le 12/12/2012 par

Federico Ulliana

Detection de l'independance entre requête et mise à jour XML : une approche basée sur le typage

THÈSE DIRIGÉE PAR

Bidoit-Tollu Nicole	*Professeur, Université Paris Sud*
Colazzo Dario	*Maître de Conférence, HDR, Université Paris Sud*

RAPPORTEURS

Schmitt Alan	*CR, HDR, INRIA Rennes*
Senellart Pierre	*Maître de Conférence, HDR, Telecom ParisTech*

EXAMINATEURS

Manoussakis Yannis	*Professeur, Université Paris Sud*
Talbot Jean-Marc	*Professeur, Université de Provence*

Résumé

Pendant la dernière décennie, le format de données XML est devenu l'un des principaux moyens de représentation et d'échange de données sur le Web. La détection de l'indépendance entre une requête et une mise à jour, qui a lieu en absence d'impact d'une mise à jour sur une requête, est un problème crucial pour la gestion efficace de tâches comme la maintenance des vues, le contrôle de concurrence et de sécurité. Cette thèse présente une nouvelle technique d'analyse statique pour détecter l'indépendance entre requête et mise à jour XML, dans le cas où les données sont typées par un schéma. La contribution de la thèse repose sur une notion de type plus riche que celle employée jusqu'ici dans la littérature. Au lieu de caractériser les éléments d'un document XML utiles ou touchés par une requête ou mise à jour en utilisant un ensemble d'étiquettes, ceux-ci sont caractérisés par un ensemble de chaînes d'étiquettes, correspondants aux chemins parcourus pendant l'évaluation de l'expression dans un document valide pour le schéma. L'analyse d'indépendance résulte du développement d'un système d'inférence de type pour les chaînes. Cette analyse précise soulève une question importante et difficile liés aux schémas récursifs : un ensemble infini de chaînes pouvant être inférées dans ce cas, est-il possible et comment se ramener à une analyse effective donc finie. Cette thèse présente donc une technique d'approximation correcte et complète assurant

une analyse finie. L'analyse de cette technique a conduit à développer des algorithmes pour une implantation efficace de l'analyse, et de mener une large série de tests validant à la fois la qualité de l'approche et son efficacité.

Introduction

Dans la dernière décennie XML est devenu le standard de facto pour la représentation et l'échange des données sur le Web. Une famille de langages d'interrogation et de manipulation des données a été conçue au tour du format XML, dont figurent les W3C standards comme XPath, XQuery et XSLT [SCF+07]. La plus récente standardisation d'un langage pour la mise à jour des documents XML [RCD+11] a produit un gros volume de recherche par la communauté scientifique des bases des données. Des vieilles questions concernant l'optimisation de mise à jour, qui étaient déjà reconnues comme très importantes dans la modèle relationnel, sont ouvertes de nouveau pour les données semi-structurées XML.

L'indépendance entre requête et mise à jour XML est une de ces questions. Cette propriété vaut quand une requête et une mise à jour n'interagissent pas sur les données et, en particulier, aucune opération de la mise à jour ne peut changer le résultat de la requête. Détecter l'indépendance entre requête et mise à jour est très importante car permet d'optimiser le processus de maintenance des vues matérialisées, la concurrence, la sécurité, et en général le contrôle des accès. Étaient les vues matérialisées des requêtes calculées sur la base de données, détecter l'indépendance par rapport à une mise à jour permet d'éviter toute opération de maintenance de la vue lorsque les données évoluent, ce qui dans certains cas peut-être compris seulement après des vérifications qui prennent un temps proportionnel à la taille de la base de données. Concernant le contrôle de la concurrence, si une requête et une mise à jour n'interagissent pas, ils peuvent être exécutée en quelconque ordre sur la base des données, ou en parallèle, en garantissant toujours le même état final de la base des données. Enfin, quand une vue de sécurité spécifique qu'un uti-

lisateur ne peut pas modifier partie des données, l'indépendance peut être utilisée pour assurer que toutes mises à jour de l'utilisateur respectent cette politique des accès.

Une requête et une mise à jour sont indépendantes lorsque, pour toutes bases des données, le résultat de la requête est jamais impacté par l'exécution de la mise à jour. Dans tous contextes où décider cette propriété est fondamentale, les optimisations sont amplifiées si l'indépendance peut être décidée de façon *statique*. Pour être fiable, une analyse statique doit être correcte : si une indépendance statique est détectée alors elle doit être vraie dans tout cas. Pourtant, l'implication inverse n'est pas assurée car le problème est indécidable en général, comme montre dans [BC09].

Cela implique que si un outil d'analyse statique est employée, par exemple, dans un système de maintenance des vues, dans quelque cas les requêtes sont rematérialisées inutilement car l'analyse peut donner lieu à des faux négatifs. Cela est souvent le cas, si un analyseur avec peut de précision est utilisé.

Une technique d'analyse précise peut être conçue en prenant en compte le schéma de la base des données. Malgré le fait qu'une base des donnés XML peut exister sans schéma, ces sont dans plusieurs cas définis par les utilisateurs, ou peuvent être calculée avec techniques à la fois précises et efficientes [BNSV10]. La détection de l'indépendance entre requêtes et mises à jour XML utilisant les schémas a été récemment étudiée. L'état de l'art est constitué par la technique proposée par Benedikt et Cheney [BC09]. Cette technique calcule, à partir du schéma, l'ensemble des types des noeuds outils ou touchés par la requête et la mise à jour. Les deux expressions sont considérées indépendantes si ces deux ensembles sont disjoints. Cette méthode est capable de gérer une sous ensemble d'XQuery,

présente une complexité très basse, et donc peut éviter plusieurs récomputations si utilisée pour des optimisations. Pourtant, la méthode a des limites de précision. Comme montrée dans [BC09], dans certains cas, l'indépendance est exclue à cause de certains approximations apportées par le système d'inference, que sont nécessaires pour concevoir une technique trés rapide.

Contributions

Dans cette thèse nous proposons une nouvelle méthode pour la détection de l'indépendance entre requête et mise à jour XML, en présence d'un schéma. Par rapport aux travaux proposés en [BC09, Che08, CGMS06], notre système infère à partir du schéma des séquences d'étiquettes, que nous appelons chaînes, afin d'analyser statiquement si les les données visitées par la requête sont disjoint de ces modifiés par la mise à jour. Intuitivement, pour chaque noeud d'un document valide par rapport au schéma, qui est nécessaire à l'évaluation d'une expression, notre système infère une chaîne qui représente toutes étiquettes rencontrées dans le chemin dès la racine au noeud, et l'ordre de visite. Cette approche permet de développer une analyse statique très précise.

La contribution majeure de ce travail est un algorithme précis pour détecter l'indépendance entre requêtes et mises à jour XML. Cette méthode porte sur les contributions suivantes.

1. Une notion *statique* d'indépendance entre requête et mise à jour basées sur les chaînes des étiquettes utiles ou touchées par les expressions. À partir de l'ensemble de chaînes associé à une DTD, notre système d'inférence calcule un sous-ensemble des chaînes pour la requête et la mise à jour. Cela représente toutes navigations qui peuvent avenir lorsque les expressions sont évaluées sur un document valide par rapport au schéma. Notre système d'inférence supporte tout type d'axe de navigation XPath. L'indépendance basée sur les chaînes vaut en absence de superposition de chaînes inférées pour la requête et la mise à jour. La preuve formelle de correction de notre système d'inférence est ainsi fournie.

2. Une contribution relevante de notre travail concerne le traitement des schémas récursifs, pour lesquels l'inférence des chaînes peut se retrouver à gérer un nombre infini de chaînes. Dans ce cas, on montre que dans tout cas il suffit de limiter l'analyse à un ensemble fini de chaînes pour détecter l'indépendance. Ce résultat nous permet de donner une analyse *finie* dans tout cas qui est *équivalente* de celle possiblement infinie.

3. Dans la perspective de concevoir une technique efficiente en pratique, on montre qu'en utilisant une technique de représentation basée sur les graphes, notre analyse peut être exécutée en temps et espace polynomiale. On a procédé à l'implantation de notre technique, ainsi qu'à une extensive série de tests, pour valider à la fois la précision et l'efficience de notre approche. Pour ce qui concerne la précision, nos résultats montrent que notre méthode présente une précision supérieure par rapport à la méthode relevante dans l'état de l'art [BC09].

Les résultats de cette thèse ont été publiés dans *International Conference of Very Large Databases 2012* [BTCU12]. Versions précédentes du travail ont été publiés dans *26ème Journée des Bases de Données Avancées 2010* [BTCU10a] et *International Formal Methods Workshop 2010* [BTCU10b].

L'organisation du manuscrit est reportée de suite. Le premier chapitre de la thése introduit le problème de détecter l'indépendance entre requête et mise à jour XML. Le deuxième chapitre révise l'état de l'art pour ce problème et fournit des motivations pour le travail ici fait. Le quatrième chapitre présente notre approche pour détecter l'indépendance basée sur les chaînes. Le cinquième chapitre présente comment réaliser une implantation efficace de la méthode. Le sixième chapitre

présente des tests qui valident l'approche avec chaînes. Le septième chapitre décrit des extensions, et le huitième chapitre présente des directions de travail futures. Le neuvième chapitre contient les preuves formelles de la correction de la méthode.

Bibliographie

[BC09] Michael Benedikt and James Cheney. Schema-based independence analysis for XML updates. In *VLDB*, 2009.

[BNSV10] Geert Jan Bex, Frank Neven, Thomas Schwentick, and Stijn Vansummeren. Inference of concise regular expressions and DTDs. *ACM TODS*, 2010.

[BTCU10a] Nicole Bidoit-Tollu, Dario Colazzo, and Federico Ulliana. Detecting XML query-update independence. *26ème journée des Bases des données Avancées*, 2010.

[BTCU10b] Nicole Bidoit-Tollu, Dario Colazzo, and Federico Ulliana. Detecting XML query-update independence. *International Formal Methods Workshop*, 2010.

[BTCU12] Nicole Bidoit-Tollu, Dario Colazzo, and Federico Ulliana. Type-based detection of XML query-update independence. *PVLDB*, 5(9), 2012.

[CGMS06] Dario Colazzo, Giorgio Ghelli, Paolo Manghi, and Carlo Sartiani. Static analysis for path correctness of XML queries. *Journal of Functional Programming*, 16, 2006.

[Che08] James Cheney. FLUX : FunctionaL Updates for XML. In *ICFP*, 2008.

[RCD+11] J Robie, D Chamberlin, M Dyck, D Florescu, J Milton, and J Simeon. XQuery update facility 1.0. Technical report, W3C Consortium, 2011.

[SCF+07] Jérôme Siméon, Don Chamberlin, Daniela Florescu, Scott Boag, Mary F. Fernández, and Jonathan Robie. XQuery 1.0 : An XML query language. W3C recommendation, W3C, January 2007. http ://www.w3.org/TR/2007/REC-xquery-20070123/.