

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Αυγουστάκη Αργυρώ

Μεταπτυχιακή Φοιτήτρια

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Καθηγητής Δ. Πλεξουσάκης

Πέμπτη, 18 Δεκεμβρίου 2014, 11:00

Αίθουσα Ε313, τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

"Διαχείριση πληροφορίας προέλευσης σε SPARQL ενημερώσεις"

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια παρατηρείται μια έκρηξη στη δημοσίευση δεδομένων στον Παγκόσμιο Ιστό, κυρίως με τη μορφή Συνδεδεμένων Δεδομένων (Linked Data). Δεδομένα από διάφορες θεματικές περιοχές, π.χ. επιστημονικά, εταιρικά, κυβερνητικά κτλ., διατίθενται για ανοιχτή πρόσβαση και χρήση από εφαρμογές, μεμονωμένους χρήστες ή ακόμα και κοινότητες χρηστών. Δεδομένου του αυξανόμενου όγκου και της ετερογένειας των δεδομένων αυτών κρίνεται επιτακτική η ανάγκη για καταγραφή της πληροφορίας προέλευσης (provenance). Η γνώση της προέλευσης μάς δίνει τη δυνατότητα να υποστηρίξουμε αποτελεσματικά εφαρμογές που σχετίζονται με την αξιοπιστία, την φερεγγυότητα και την επαναληπτικότητα των δεδομένων.

Ένα πλήθος από μοντέλα έχει ήδη προταθεί για την καταγραφή της πληροφορίας προέλευσης των αποτελεσμάτων μιας επερώτησης (query); τα περισσότερα από τα οποία αφορούν RDF ή σχεσιακά (relational) δεδομένα. Αντίθετα, και παρά τη σπουδαιότητα του προβλήματος, η έρευνα για την περίπτωση των ενημερώσεων (updates), και ειδικότερα των SPARQL ενημερώσεων, βρίσκεται ακόμα σε πρώιμο στάδιο.

Στην εργασία αυτή, προτείνουμε ένα νέο μοντέλο για την καταγραφή και διαχείριση της πληροφορίας προέλευσης, σε επίπεδο τριπλέτας (triple) και γνωρίσματος (attribute), των αποτελεσμάτων των SPARQL updates. Το μοντέλο αυτό, το οποίο δανείζεται χαρακτηριστικά

και ιδιότητες από τα ήδη υπάρχοντα μοντέλα του where και how είναι το πρώτο που υποστηρίζει τη χρήση αλγεβρικών εκφράσεων σε ενημερώσεις, ακολουθώντας την προσέγγιση του μοντέλου των provenance semirings.

Από αλγοριθμικής σκοπιάς, παρουσιάζουμε έναν αλγόριθμο, ο οποίος υπολογίζει την πληροφορία προέλευσης για τα αποτελέσματα των SPARQL updates με βάση το προτεινόμενο μοντέλο, καθώς και έναν αλγόριθμο ανακατασκευής (reconstruction), ο οποίος χρησιμοποιεί την πληροφορία προέλευσης μιας τετραπλέτας (quadruple) για να δημιουργήσει ένα SPARQL update, αποδεδειγμένα, συμβατό (compatible) με το αρχικό. Ένα SPARQL update είναι συμβατό με ένα άλλο, αν διαφέρουν μόνο στα ονόματα των μεταβλητών που χρησιμοποιούν, και το πρώτο update περιέχει ένα γνήσιο υποσύνολο των ενώσεων (unions) που εμφανίζονται στο δεύτερο. Η παροχή ενός αλγορίθμου ανακατασκευής κρίνεται απαραίτητη ώστε να μπορέσουμε να περιγράψουμε πλήρως τη διαχείριση της πληροφορίας προέλευσης, καθώς φανερώνει τον καθοριστικό ρόλο της πληροφορίας αυτής στη διατήρηση της συνεκτικότητας (persistence) των αποτελεσμάτων των SPARQL updates.

Avgoustaki Argyro

M.Sc. Thesis

Computer Science Department

University of Crete

Master's Thesis Supervisor: Professor Dimitris Pleksousakis

Thursday, 18/12/2014, 11:00

Room E313, Computer Science dept., University of Crete

"Provenance Management in SPARQL Updates"

ABSTRACT

During the last few years we have witnessed an explosion in the publication of data in the Web, mainly in the form of Linked Data. Scientific, corporate or even governmental data are made available for open access and used by applications, individual users and communities. Given the increasing amount and the heterogeneity of this data, it is of crucial importance to be able to track its provenance. Recording the provenance can help us to effectively support trustworthiness, accountability and repeatability in the Web of Data.

A number of models have already been proposed to capture the provenance information of query results; most of them considering RDF or relational data. On the contrary, despite its

importance, little research has been conducted in the case of updates and especially of SPARQL updates.

In this thesis, we propose a new provenance model that borrows from both how and where data provenance models, and is suitable for capturing the triple and attribute level provenance of SPARQL update results. To the best of our knowledge, this is the first model that deals with the provenance of SPARQL updates using algebraic provenance expressions, in the spirit of the well-established model of provenance semirings.

On the algorithmic side, we introduce an algorithm that records the provenance of SPARQL update results in terms of the proposed model and a reconstruction algorithm that uses the provenance of a quadruple to identify a SPARQL update that is provably compatible to the original one. A SPARQL update is compatible to another if they differ only in the variables names that they employ and the first update contains a genuine subset of the unions that appear in the second one. The latter algorithm is a necessary complement in order to fully describe the provenance management, as it shows the determinant role of provenance information in the persistence of SPARQL update results.