ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Γιαννάκης – Χαλβατζής Αθανάσιος
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης
Επόπτης Μεταπτ. Εργασίας: Αναπλ. Καθηγητής, Ι. Τζίτζικας

Τρίτη, 05/12/2017, 12:00
Αίθουσα Κ206,Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Βελτιστοποιήσεις για τη Γλώσσα Επερώτησης Διασυνδεδεμένων Δεδομένων SPARQL-LD"

ΠΕΡΙΛΗΨΗ

Τα Διασυνδεδεμένα Δεδομένα (Linked Data) είναι ένας τρόπος δημοσίευσης δεδομένων στο διαδίκτυο που διευκολύνει τη διασύνδεση και την ολοκλήρωσή τους. Υπάρχει ένας διαρκώς αυξανόμενος αριθμός οργανισμών ή επιχειρήσεων που δημοσιεύουν τα δεδομένα τους ως διασυνδεδεμένα δεδομένα (Linked Data) τα οποία είναι επερωτήσιμα μέσω της SPARQL που είναι η στάνταρ γλώσσα επερωτήσεων του Σημασιολογικού Ιστού (Semantic Web). Ωστόσο, η χρήση της SPARQL προϋποθέτει ότι τα δεδομένα προς επερώτηση είναι διαθέσιμα εκ των προτέρων στην κύρια μνήμη ή σε κάποιο σημείο σύνδεσης SPARQL (SPARQL endpoint).

Για να δώσουμε μεγαλύτερη ευελιξία, σε αυτήν την εργασία επικεντρωνόμαστε στην SPARQL-LD, μια επέκταση της SPARQL 1.1 που σχεδιάστηκε στο Ινστιτούτο Πληροφορικής, η οποία προσφέρει τη δυνατότητα επερώτησης σε απομακρυσμένα δεδομένα ακόμα και αν δεν φιλοξενούνται από ένα σημείο σύνδεσης SPARQL. Χρησιμοποιώντας τη SPARQL-LD, μπορεί κανείς να επερωτήσει σύνολα δεδομένων (datasets) που είναι προσβάσιμα ως RDF dumps, ως ενσωματωμένα δεδομένα σε ιστοσελίδες, δηλαδή σε μορφή RDFa, JSON-LD, Microdata ή Microformat, καθώς και σε δεδομένα που αντιστοιχούν σε μερικά αποτελέσματα μιας επερώτησης (δηλαδή, που ανακτήθηκαν κατά το χρόνο εκτέλεσης της επερώτησης), ή που δημιουργήθηκαν δυναμικά από υπηρεσίες διαδικτύου. Αυτή η λειτουργικότητα μπορεί να δώσει

κίνητρο στους ιδιοκτήτες περιεχομένου να υιοθετήσουν τις αρχές των διασυνδεδεμένων δεδομένων και να εμπλουτίσουν το ψηφιακό τους περιεχόμενο και υπηρεσίες τους με RDF, αφού έτσι τα δεδομένα τους θα είναι άμεσα επερωτήσιμα μέσω της SPARQL-LD χωρίς να χρειάζονται να δημιουργήσουν και να συντηρούν ένα λειτουργικό σημείο σύνδεσης SPARQL.

Εν συνεχεία επικεντρωνόμαστε σε βελτιστοποιήσεις που αφορούν τη SPARQL-LD, συγκεκριμένα σε: (α) μεθόδους που διερευνούν συντακτικές παραλλαγές των μοτίβων γράφων (Graph Patterns) μιας επερώτησης SPARQL για την επιλογή του σχεδόν βέλτιστου πλάνου εκτέλεσης χωρίς τη χρήση στατιστικών στοιχείων (για το σκοπό αυτό εφαρμόζουμε τεχνικές αναδιάταξης επερωτήσεων βασισμένων σε τεχνικές εκτίμησης της εκλεκτικότητας νέων - μη δεσμευμένων - μεταβλητών με στόχο την επιτάχυνση καθώς και τη μείωση των απομακρυσμένων κλήσεων), και (β) τεχνικές παράλληλης ανάκτησης RDF δεδομένων από το διαδίκτυο με την χρήση απομακρυσμένων επερωτήσεων της SPARQL-LD για αποδοτικότερη προσωρινή αποθήκευση των δεδομένων.

Τέλος, αναφέρουμε πειραματικές μετρήσεις που πραγματοποιήθηκαν πάνω σε πραγματικά σύνολα δεδομένων για την αξιολόγηση της απόδοσης καθώς και της ποιότητας των προτεινόμενων βελτιστοποιήσεων, όπου παρατηρήθηκε σημαντική βελτίωση της επίδοσης κατά την χρήση απομακρυσμένων κλήσεων σε σχέση με άλλες μεθόδους.

Giannakis – Xalvatzis Athanasios
M.Sc. Thesis

Computer Science Department
University of Crete
Master's Thesis Supervisor: Associate Professor, I. Tzitzikas

Tuesday, 05/12/2017, 12:00
Room K206, Computer Science Dept., University of Crete

"Optimizations for the Query Language SPARQL-LD"

ABSTRACT

Linked Data is a method for publishing structured data in the Web for assisting their linking and integration. A constantly increasing number of organizations and owners publish their data on the Web as Linked Data and SPARQL is the standard query language. However the majority of SPARQL

implementations require the data to be available in advance in main memory or accessible through a SPARQL endpoint.

SPARQL-LD is an extension of SPARQL 1.1 (designed by FORTH-ICS) that overcomes this restriction and allows fetching and querying RDF data from any Web source and format i.e., RDFa, JSON-LD, Microdata and Microformats. Using SPARQL-LD, one can query a dataset corresponding to an RDF dump, or a dataset corresponding to the partial results of a query (i.e., discovered at query execution time), or RDF data that are dynamically created by Web Services. This functionality can motivate content owners to adopt the Linked Data principles and enrich their digital content and services with RDF, for having their data directly queryable via SPARQL–LD without having to create and maintain an operational SPARQL endpoint.

In this thesis we focus on optimizations for SPARQL-LD, in particular on: (a) methods for exploring the syntactic variations of graph patterns in a SPARQL query in order to choose a near to optimal execution plan without the use of statistics (to this end we utilize query reordering techniques, using selectivity estimation procedures on new unbound variables for increasing efficiency and decreasing intermediate results and thus the number of calls to remote sources), and (b) methods for parallelizing RDF data retrieval by using SPARQL-LD for efficient data caching.

Finally, we report experimental results on real datasets for evaluating the efficiency as well as the quality of the proposed optimizations. The results showed improved efficiency on SPARQL queries in comparison to existing methods.