ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Μουνταντωνάκης Μιχάλης

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Επίκ. Καθηγητής Ι. Τζίτζικας

Δευτέρα, 27/6/2016, 16:00

Αίθουσα B108, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Ευρετήρια και αλγόριθμοι για τη μέτρηση του βαθμού διασύνδεσης των διασυνδεδεμένων δεδομένων"

ΠΕΡΙΛΗΨΗ

Τα Διασυνδεδεμένα Δεδομένα (Linked Data) είναι ένας τρόπος δημοσίευσης δεδομένων που επιτρέπει τη διασύνδεσή τους (μέσω της χρήσης URIs αντί απλών τιμών) και διευκολύνει την ολοκλήρωσή τους. Ήδη υπάρχουν χιλιάδες τέτοια σύνολα δεδομένων, στο εξής πηγές, και ο αριθμός και το μέγεθος τους διαρκώς αυξάνεται. Παρά ταύτα, αυτή τη στιγμή είναι δύσκολο να εκτιμήσει κανείς πόσο συνδεδεμένες είναι αυτές οι πηγές, και συγκεκριμένα είναι δύσκολη (α) η εύρεση όλων των δεδομένων που αφορούν ένα συγκεκριμένο URI, (β) η ανακάλυψη μιας πηγής που σχετίζεται με μία άλλη, (γ) ο υπολογισμός και η οπτικοποίηση του βαθμού διασύνδεσης μεταξύ δύο ή περισσότερων πηγών. Τα παραπάνω είναι αναγκαία στη διαδικασία ολοκλήρωσης σε ένα ανοικτό και εξελισσόμενο περιβάλλον.

Για να απαλύνουμε αυτό το πρόβλημα σε αυτήν την εργασία, παρουσιάζουμε μέτρα, ευρετήρια και αλγορίθμους που επιτρέπουν τη μέτρηση και ποσοτικοποίηση του βαθμού διασύνδεσης πολλών πηγών. Για λόγους κλιμακωσιμότητας προτείνουμε i) ένα ευρετήριο για τα προθέματα των URIs ii) έναν κατάλογο για σχέσεις ισοδυναμίας που λαμβάνει υπ' όψιν του το συμμετρικό και μεταβατικό κλείσιμο των σχέσεων ισοδυναμίας που εμφανίζονται στα σύνολα δεδομένων, iii) ένα σημασιολογικό ευρετήριο στοιχείων (που χρησιμοποιεί τα προαναφερθέντα ευρετήρια),

iv) ένα πλέγμα (lattice) των κοινών στοιχείων που μετράει όλα τα κοινά στοιχεία ενός συνόλου πηγών, και v) δύο αυξητικούς αλγορίθμους που επιταχύνουν τον υπολογισμό του πλέγματος.

Εφαρμόζουμε και αξιολογούμε την προσέγγιση τόσο στο πλαίσιο μιας συγκεκριμένης σημασιολογικής αποθήκης δεδομένων με πληροφορίες για θαλάσσια είδη (όπου εκεί τα μέτρα αυτά χρησιμοποιούνται για την αξιολόγηση της αποθήκης και των συνιστωσών πηγών της, καθώς και για τον έλεγχο της ποιότητας της αποθήκης μετά από ανακατασκευή), καθώς και σε τρακόσες πηγές του νέφους διασυνδεδεμένων δεδομένων. Αναφέρουμε τα αποτελέσματα μετρήσεων που δεν έχουν γίνει στο παρελθόν (όπως το πλήθος της τομής των κοινών URIs μεταξύ τριών ή παραπάνω πηγών, συχνότητα των prefixes, κ.α.), προσφέρουμε νέες υπηρεσίες (όπως εύρεση ισοδύναμων URIs, εύρεση των κοντινότερων πηγών ως προς μία, κ.α.), και τέλος αξιολογούμε την επιτάχυνση που επιτυγχάνεται με τα προτεινόμενα ευρετήρια και αλγορίθμους. Τέλος, προτείνουμε μία επέκταση της οντολογίας VoID που επιτρέπει τη δημοσίευση, το διαμοιρασμό και την αξιοποίηση τέτοιων μετρήσεων.

Mountantonakis Mixalis
M.Sc. Thesis
Computer Science Department
University of Crete
Master's Thesis Supervisor: Assistant Professor I. Tzitzikas

Monday, 27/6/2016, 16:00
Room B108, Computer Science dept., University of Crete

"Indexes and algorithms for measuring the connectivity of Linked Data"

ABSTRACT

Linked Data is a method for publishing structured data that allows them to be interlinked (by using URIs instead of simple values) for assisting their integration. A big number of such datasets, hereafter *sources*, has already been published according to the principles of Linked data and their number and size keeps increasing. However, currently it is not evident how connected these datasets are. In particular, it is difficult (a) to obtain complete information about one particular URI (or a set of URIs), (b) to discover a dataset which is relevant to another one, (c) to compute and visualize the degree of connectivity between two or more datasets. All the aforementioned tasks are important for the integration process in an open and involving environment.

To alleviate this problem in this thesis, we introduce metrics, indexes and algorithms which allow the computation and quantification of connectivity among several datasets. For achieving scalability, we propose (i) a namespace-based prefix index, (ii) a sameAs catalog for computing the symmetric and transitive closure of the sameAs relationships encountered in the datasets, (iii) a semantics-aware element index (that exploits the aforementioned indexes), (iv) a lattice of the common elements of any set of datasets, and (v) two lattice-based incremental algorithms for speeding up the computation of the lattice.

We apply and evaluate the proposed approach in the context of a real and operational semantic warehouse containing information about the marine domain (where the metrics are used for assessing the quality of the semantic warehouse and its underlying sources, and for monitoring the quality of the semantic warehouse after a reconstruction), as well as for three hundred LOD cloud datasets. We report measurements that have not been carried out in the past (like the number of common URIs among three or more datasets, the frequency of prefixes, i.a.), we offer novel services (like finding equivalent URIs, find the most relevant datasets for a specific dataset, i.a.**)** and finally we discuss the speedup obtained by the proposed indexes and algorithms. Finally, we propose an extension of the VoID ontology for publishing, sharing and exploiting such measurements.