ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Σαβέτα Τζανίνα

Μεταπτυχιακή Φοιτήτρια

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Καθηγητής Δ. Πλεξουσάκης

Παρασκευή, 31 Οκτωβρίου 2014, 10:00
Αίθουσα Ε313, τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

"SPIMBench: Ένα Κλιμακώσιμο, με Επίγνωση Σχήματος Πλαίσιο Αξιολόγησης Συστημάτων Αντιστοίχισης Στιγμιοτύπων για τη Δημοσίευση Σημασιολογικά Εμπλουτισμένων Δεδομένων"

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια, η αύξηση των διαθέσιμων Συνδεδεμένων Δεδομένων (Linked Data) στον Παγκόσμιο ιστό έχει αποτελέσει τον θεμέλιο λίθο στην ανάπτυξη Συστημάτων Αντιστοίχισης Στιγμιοτύπων (Instance Matching Systems). Όπως για τα συστήματα Βάσεων Δεδομένων, έτσι και εδώ, Πλαίσια Αξιολόγησης Συστημάτων Αντιστοίχισης Στιγμιοτύπων (Instance Matching Benchmarks) έχουν αναπτυχθεί για τον έλεγχο απόδοσης των προαναφερθέντων συστημάτων με βασικό σκοπό τον προσδιορισμό των μειονεκτημάτων τους για την περαιτέρω βελτίωση των λειτουργιών τους. Ένα πλαίσιο αξιολόγησης συστημάτων ταυτοποίησης στιγμιοτύπων θα πρέπει να ελέγχει τη συνολική ποιότητα του συστήματος αντιστοίχισης στιγμιοτύπων με μετρικές όπως η ακρίβεια (precision), η ανάκληση (recall), και το F-measure καθώς και την ικανότητα να χειρίζεται σύνολα δεδομένων μεγάλου όγκου. Πλαίσια αξιολόγησης έχουν ήδη προταθεί για τον έλεγχο της απόδοσης συστημάτων αντιστοίχισης στιγμιοτύπων για δεδομένα XML και δεδομένα σχεσιακών βάσεων και πρόσφατα για τα δεδομένα RDF τα οποία έχουν αρχίσει να επικρατούν στον Παγκόσμιο Ιστό. Τα συστήματα αξιολόγησης που λαμβάνουν υπ' όψιν δεδομένα εκφρασμένα σε RDF είναι τα πρώτα τα οποία εξέτασαν το πρόβλημα της αντιστοίχισης στιγμιοτύπων όταν ένα αντικείμενο του πραγματικού κόσμου έχει

διαφορετικές περιγραφές που χρησιμοποιούν τα ίδια ή διαφορετικά RDFS (ή τα εκφραστικότερα OWL) σχήματα. Αυτό σημαίνει πως εκτός από τις λεξικολογικές διαφορές μεταξύ των στιγμιοτύπων που περιγράφουν την ίδια οντότητα του πραγματικού κόσμου, τα πλαίσια αξιολόγησης λαμβάνουν υπ' όψιν διαφορές σε επίπεδο σχήματος όπως τη διάσπαση ή τη συνάθροιση μίας ιδιότητας ενός στιγμιότυπου. Ωστόσο, σύμφωνα με τη βιβλιογραφία, κανένα από τα προτεινόμενα πλαίσια αξιολόγησης μέχρι σήμερα δεν λαμβάνει υπ' όψιν τις πιο πολύπλοκες δομές σε επίπεδο σχήματος τα οποία μπορούν να εκφραστούν, χρησιμοποιώντας τα πλούσια δομικά στοιχεία της γλώσσας του Σημασιολογικού Ιστού OWL. Οι μετασχηματισμοί που έχουν προταθεί παραμένουν όλοι στο επίπεδο των απλών δομών όπως εκείνες περιγράφονται στην γλώσσα RDFS.
Στην παρούσα εργασία προτείνουμε το Semantic Publishing Instance Matching Benchmark, εν συντομία SPIMBench, ένα πλαίσιο αξιολόγησης εμπνευσμένο από το Semantic Publishing Benchmark (SPB). Το SPIMBench, όπως το SPB, είναι βασισμένο στις οντολογίες όπως έχουν δοθεί από το BBC (http://www.bbc.com/) οι οποίες χρησιμοποιήθηκαν από τον συγκεκριμένο δημοσιογραφικό οργανισμό για την δημοσίευση Σημασιολογικά Εμπλουτισμένων Δεδομένων. Στο SPIMBench προτείνουμε και υλοποιούμε μία α) επεκτάσιμη γεννήτρια δεδομένων, β) ένα σύνολο μετασχηματισμών πού αποτελούνται από τους καθιερωμένους λεξικολογικούς, δομικούς και μετασχηματισμούς σε επίπεδο λογικού σχήματος. Οι τελευταίοι μετασχηματισμοί υπερβαίνουν τα καθιερωμένα δομικά στοιχεία και περιλαμβάνουν εκφραστικά δομικά στοιχεία όπως ισότητα/ανισότητα στιγμιοτύπων, ισοδυναμία των κλάσεων και των ιδιοτήτων σε επίπεδο σχήματος, περιορισμό ιδιοτήτων, περίπλοκους ορισμούς κλάσεων, και τέλος γ) έναν σταθμισμένο χρυσό κανόνα ο οποίος μπορεί να χρησιμοποιηθεί για τον εντοπισμό σφαλμάτων στα συστήματα αντιστοίχισης στιγμιοτύπων.

Saveta Tzanina

M.Sc. Thesis

Computer Science Department

University of Crete

Master's Thesis Supervisor: Professor Dimitris Pleksousakis

Friday, 31/10/2014, 10:00

Room E313, Computer Science dept.,University of Crete

"SPIMBench: A Scalable, Schema-Aware Instance Matching Benchmark for the Semantic Publishing Domain"

ABSTRACT

Instance matching systems and methods need to be tested using well defined and widely accepted benchmarks to determine the weak and strong points thereof and also to motivate the development of more complete systems. A benchmark should test the overall quality of the instance matching system in terms of measures such as precision, recall, and F-measure as well as the ability to handle large and diverse datasets. A number of benchmarks have already been proposed to test the performance of instance matching techniques mostly for XML and relational data but, more recently, also for RDF, the type of data prevalent in the Web of Data. Instance Matching benchmarks for RDF data are the first to consider the problem of instance matching when a real world object is represented in different ways that do not all conform to the same RDFS or OWL schema. Meaning that in addition to lexical differences among entities representing the same object, these benchmarks consider structural differences such as property splitting or aggregation. However, to the best of our knowledge, none of the proposed benchmarks to date considers the more complex logical constructs that can be expressed in terms of rich OWL constructs. The logical transformations proposed by existing benchmarks all remain at the level of simple RDFS constraints. In this thesis we propose the Semantic Publishing Instance Matching Bench- mark, in short, SPIMBench inspired from the Semantic Publishing domain. SPIMBench is based on the BBC (http://www.bbc.com/) ontologies that represent information about creative works (called journalistic assets) created by the publisher's editorial team. SPIMBench proposes and implements i) a scalable data generator, ii) a set of transformations on source data to obtain the target data that include, in addition to the standard value and structural transformations, logical ones that go beyond the standard RDFS constructs and include expressive OWL constructs, namely instance (in)equality, equivalence of classes and properties, property constraints and complex class definitions, a iii) weighted gold standard that can be used for debugging instance matching systems and finally, iv) a set of metrics used to assess the performance of an instance matching system.