ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Καρδουλάκης Νικόλαος
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Δ. Πλεξουσάκης

Χ. Κονδυλάκης (Επιβλέπων)

Πέμπτη , 29 Οκτωβρίου 2020 ,ώρα 10:00 π.μ.

Τηλεδιάσκεψη (μέσω του συστήματος e:Presence), Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Διεύθυνση μετάδοσης (url): http://video.ucnet.uoc.gr/live/show/326

Κανάλι YouTube του Τμήματος
https://www.youtube.com/channel/UC7uE3QiMTQjkrpByB_Gnt6Q/live

"HInT: Υβριδική και Αυξητική Ανακάλυψη Τύπων για Μεγάλα RDF Δεδομένα"

Περίληψη

Η ταχεία ανάπτυξη των διασυνδεδεμένων δεδομένων έχει οδηγήσει στη δημιουργία πολλών πηγών δεδομένων με ασθενή και ελλιπή δομή, στις οποίες οι δηλώσεις των τύπων απουσιάζουν μερικώς ή ολικώς. Από την άλλη πλευρά, η πληροφορία σχετικά με τους τύπους είναι απαραίτητη για ένα πλήθος εργασιών, όπως η απάντηση ερωτήσεων,

η ολοκλήρωση δεδομένων, η δημιουργία συνόψεων και ο κατακερματισμός πηγών δεδομένων σε τμήματα. Οι υπάρχουσες προσεγγίσεις για ανακάλυψη τύπων είτε αγνοούν εντελώς τους ορισμούς τύπων που είναι διαθέσιμοι στα δεδομένα (τεχνικές έμμεσης ανακάλυψης τύπων), είτε βασίζονται στη μερική διαθεσιμότητα αυτών των τύπων, προκειμένου να τους συμπληρώσουν (τεχνικές ρητού εμπλουτισμού τύπων). Οι τεχνικές έμμεσης ανακάλυψης τύπων βασίζονται σε ομαδοποίηση των οντοτήτων, η οποία προϋποθέτει την εξαντλητική μεταξύ τους σύγκριση. Η διαδικασία αυτή είναι κοστοβόρα και μη αυξητική. Από την άλλη, οι τεχνικές ρητού εμπλουτισμού τύπων αδυνατούν να επεξεργαστούν σύνολα δεδομένων που περιέχουν πληροφορία σχετικά με τη δομή τους σε μικρό ή μηδενικό βαθμό.

Σε αυτήν την εργασία, παρουσιάζουμε το HInT, το πρώτο αυξητικό και υβριδικό σύστημα για ανακάλυψη τύπων σε συλλογές δεδομένων RDF. Η προσέγγισή μας πετυχαίνει την ανακάλυψη τύπων τόσο σε περιπτώσεις που η πληροφορία σχετικά με τους τύπους των οντοτήτων είναι μερικώς διαθέσιμη, όσο και σε εκείνες που είναι ολικώς απούσα. Για να επιτευχθεί αυτό, αναγνωρίζουμε αυξητικά τα μοτίβα των διαφόρων οντοτήτων, τα δεικτοδοτούμε και τα ομαδοποιούμε προκειμένου να αναγνωρίσουμε τους τύπους. Κατά την επεξεργασία μίας οντότητας, η τεχνική μας αξιοποιεί την πληροφορία σχετικά με τους τύπους της οντότητας αυτής, εάν υπάρχει διαθέσιμη, για να βελτιώσει την ποιότητα των ανακαλυφθέντων τύπων, καθοδηγώντας την κατηγοριοποίηση της νέας οντότητας στη σωστή ομάδα, βελτιώνοντας παράλληλα τα σύνολα που έχουν ήδη δημιουργηθεί. Επιβεβαιώνουμε αναλυτικά και πειραματικά ότι το σύστημά μας κυριαρχεί σε επίπεδο αποτελεσματικότητας και κυρίως αποδοτικότητας, σε σύγκριση με ανταγωνιστές και από τις δύο κατηγορίες, της έμμεσης ανακάλυψης τύπων και του ρητού εμπλουτισμού τύπων.

University of Crete

Computer Science Department

M.Sc. Thesis presentation / examination

Kardoulakis Nikolaos

Master's Thesis Supervisor: Professor, D. Plexousakis

X. Kondilakis (Thesis Co- Advisor)

Thursday, 29 October 2020, 10:00 a.m

Teleconference (will use the e: Presence system), Computer Science Department, University of Crete

(url) : http://video.ucnet.uoc.gr/live/show/326

YouTube channel :

https://www.youtube.com/channel/UC7uE3QiMTQjkrpByB_Gnt6Q/live

"HInT: Hybrid and Incremental Type Discovery for Large RDF Data Sources"

Abstract

The rapid explosion of linked data has resulted into many weakly structured and incomplete data sources, where type declarations are completely or partially missing. On the other hand, type information is essential for a number of tasks such as query answering, integration, summarization and partitioning. Existing approaches for type discovery, either completely ignore type declarations available in the dataset (implicit type discovery approaches), or have to rely on partial availability of those types, in order to complement them (explicit type enrichment approaches). Implicit type discovery approaches are based on instance grouping, which requires an exhaustive comparison between the instances. This process is expensive and not incremental. Explicit type enrichment approaches on the other hand, can not process data sources that have little or no schema information.

In this thesis, we present HInT, the first incremental and hybrid type discovery system for RDF datasets. It enables type discovery in datasets where type declarations are either partially available or completely missing. To achieve this goal, we incrementally identify the patterns of the various instances, we index and then group them to identify the types. During the processing of an instance, our approach exploits its type information, if available, to improve the quality of the discovered types by guiding the classification of the new instance in the correct group and by refining the groups already built. We analytically and experimentally show that our approach dominates in terms of effectiveness and most importantly efficiency, competitors from both worlds, implicit type discovery and explicit type enrichment.