ΠΡΟΣ

1) **Όλα τα μέλη ΔΕΠ του Τμήματος Επιστήμης Υπολογιστών**

2) **Τους εκπροσώπους των Μεταπτυχιακών φοιτητών του Τμήματος Επιστήμης Υπολογιστών**

3) **Την Επταμελή Εξεταστική Επιτροπή**

4) **Όλα τα μέλη της Πανεπιστημιακής Κοινότητας**

Πρόσκληση σε Δημόσια Παρουσίαση της Διδακτορικής Διατριβής του

κ. Μουνταντωνάκη Μιχαήλ

Doctoral Dissertation Defense

Mr. Michalis Mountantonakis

Την Παρασκευή, 27/03/2020 και ώρα 14:00 στην αίθουσα Τηλεδιάσκεψης Κ206 του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης στο Ηράκλειο, θα γίνει η δημόσια παρουσίαση και υποστήριξη της Διδακτορικής Διατριβής του υποψηφίου διδάκτορα του Τμήματος Επιστήμης Υπολογιστών κ. Μουνταντωνάκη Μιχαήλ με θέμα:

“ Υπηρεσίες για τη Διασύνδεση και Ολοκλήρωση Μεγάλου Πλήθους Σημασιολογικών Συνολοδεδομένων”

“ Services for Connecting and Integrating Big Number of Linked Datasets”

ΠΕΡΙΛΗΨΗ

Τα Διασυνδεδεμένα Δεδομένα (Linked Data) είναι ένας τρόπος δημοσίευσης δεδομένων που διευκολύνει το διαμοιρασμό, τη διασύνδεση, την αναζήτηση και την επαναχρησιμοποίησή τους. ΄Ηδη υπάρχουν χιλιάδες τέτοια σύνολα δεδομένων, στο εξής πηγές, και ο αριθμός και το μέγεθος τους αυξάνεται. Αν και ο κύριος στόχος των Διασυνδεδεμένων Δεδομένων είναι η διασύνδεση και η ολοκλήρωση τους, αυτός ο στόχος δεν έχει επιτευχθεί ακόμα σε ικανοποιητικό βαθμό. Ακόμα και φαινομενικά απλές εργασίες, όπως η εύρεση όλων των πληροφοριών για μία συγκεκριμένη οντότητα αποτελούν πρόκληση διότι αυτό προϋποθέτει γνώση των περιεχομένων όλων των πηγών, καθώς και την ικανότητα συλλογισμού επί των συναθροισμένων

περιεχομένων τους, συγκεκριμένα απαιτείται ο υπολογισμός του συμμετρικού και μεταβατικού κλεισίματος των σχέσεων ισοδυναμίας μεταξύ των ταυτοτήτων των οντοτήτων και των οντολογιών. Η ανακάλυψη δεδομένων (Dataset Discovery) επίσης αποτελεί μεγάλη πρόκληση, διότι οι τρέχουσες προσεγγίσεις αξιοποιούν μόνο τα μεταδεδομένα των πηγών, και δεν λαμβάνουν υπόψη τα περιεχόμενα τους.

Σε αυτή τη διατριβή, αναλύουμε το ερευνητικό έργο που έχει παραχθεί στον τομέα της Ολοκλήρωσης Διασυνδεμένων Δεδομένων με έμφαση σε τεχνικές που μπορούν να εφαρμοστούν σε μεγάλη κλίμακα. Συγκεκριμένα παραγοντοποιούμε το πρόβλημα σε διαστάσεις που επιτρέπουν την καλύτερη κατανόηση του προβλήματος και τον εντοπισμό των ανοικτών προκλήσεων. Εν συνεχεία προτείνουμε ευρετήρια και αλγορίθμους για την αντιμετώπιση των παραπάνω προκλήσεων, συγκεκριμένα μεθόδους για συλλογισμό επί των ταυτοτήτων των πόρων, για εύρεση όλων των πληροφοριών για μία οντότητα, για ανακάληψη πηγών βάσει περιεχομένου και άλλων. Λόγω του μεγάλου πλήθους και όγκου των πηγών, οι τεχνικές που προτείνονται περιλαμβάνουν αυξητικούς και παράλληλους αλγορίθμους. Δείχνουμε ότι η ανακάλυψη πηγών βάσει περιεχομένου ανάγεται στην επίλυση προβλημάτων βελτιστοποίησης και προτείνουμε τεχνικές για την αποδοτική επίλυσή τους.

Τα παραπάνω ευρετήρια και αλγόριθμοι έχουν υλοποιηθεί στη σουίτα υπηρεσιών που αναπτύξαμε που αναφέρεται με το όνομα LODsyndesis, η οποία προσφέρει όλες αυτές τις υπηρεσίες σε πραγματικό χρόνο. Επιπροσθέτως, παρουσιάζουμε μία εκτενή ανάλυση συνδεσιμότητας για ένα μεγάλο υποσύνολο πηγών του νέφους Ανοικτών Διασυνδεδεμένων Δεδομένων (LOD Cloud). Συγκεκριμένα αναφέρουμε μετρήσεις (συνδεσιμότητας και αποδοτικότητας) που αφορούν 2 δισεκατομμύρια τριπλέτες, 412 εκατομμύρια URIs και 44 εκατομμύρια σχέσεις ισοδυναμίας που προέρχονται από 400 πηγές, χρησιμοποιώντας από 1 έως 96 μηχανήματα για την ευρετηρίαση. Ενδεικτικά, χρησιμοποιώντας 96 μηχανήματα χρειάστηκαν λιγότερα από 10 λεπτά για τον υπολογισμό του συμμετρικού και μεταβατικού κλεισίματος, και 81 λεπτά για την ευρετηρίαση 2 δισεκατομμυρίων τριπλετών. Επιπρόσθετα, χρησιμοποιώντας τα ευρετήρια μαζί με τους προτεινόμενους αυξητικούς αλγορίθμους, κατέστη εφικτός ο υπολογισμός των μετρήσεων συνδεσιμότητας για 1 εκατομμύριο υποσύνολα πηγών σε 1 δευτερόλεπτο (τρεις τάξεις μεγέθους γρηγορότερα σε σχέση με συμβατικές μεθόδους), ενώ οι προσφερόμενες υπηρεσίες έχουν απόκριση δευτερολέπτων. Οι υπηρεσίες αυτές καθιστούν εφικτή και την υλοποίηση υπηρεσιών υψηλότερου επιπέδου, όπως υπηρεσίες εμπλουτισμού πηγών για χρήση από τεχνικές Μηχανικής Μάθησης καθώς και τεχνικές για Διανυσματικές Αναπαστάσεις Γράφων Γνώσης (Knowledge Graph Embeddings) και δείχνουμε ότι ο εμπλουτισμός αυτός βελτιώνει της προβλέψεις σε προβλήματα μηχανικής μάθησης.

Επιβλέπων: Αναπλ. Καθηγητής, Ιωάννης Τζίτζικας

ABSTRACT

Linked Data is a method for publishing structured data that facilitates their sharing, linking, searching and re-use. A big number of such datasets (or sources), has already been published and their number and size keeps increasing. Although the main objective of Linked Data is linking and integration, this target has not yet been satisfactorily achieved. Even seemingly simple tasks, such as finding all the available information for an entity is challenging, since this presupposes knowing the contents of all datasets and performing cross-dataset identity reasoning, i.e., computing the symmetric and transitive closure of the equivalence relationships that exist among entities and schemas. Another big challenge is Dataset Discovery, since current approaches exploit only the metadata of datasets, without taking into consideration their contents.

In this dissertation, we analyze the research work done in the area of Linked Data Integration, by giving emphasis on methods that can be used at large scale. Specifically, we factorize the integration process according to various dimensions, for better understanding the overall problem and for identifying the open challenges. Then, we propose indexes and algorithms for tackling the above challenges, i.e., methods for performing cross-dataset identity reasoning, for finding all the available information for an entity, methods for offering content-based Dataset Discovery, and others. Due to the large number and volume of datasets, we propose techniques that include incremental and parallelized algorithms. We show that content-based Dataset Discovery is reduced to solving optimization problems, and we propose techniques for solving them in an efficient way.

The aforementioned indexes and algorithms have been implemented in a suite of services that we have developed, called LODsyndesis, which offers all these services in real time. Furthermore, we present an extensive connectivity analysis for a big subset of LOD cloud datasets. In particular, we introduce measurements (concerning connectivity and efficiency) for 2 billion triples, 412 million URIs and 44 million equivalence relationships derived from 400 datasets, by using from 1 to 96 machines for indexing the datasets. Just indicatively, by using the proposed indexes and algorithms, with 96 machines it takes less than 10 minutes to compute the closure of 44 million equivalence relationships, and 81 minutes for indexing 2 billion triples. Furthermore, the dedicated indexes, along with the proposed incremental algorithms, enable the computation of connectivity metrics for 1 million subsets of datasets in 1 second (three orders of magnitude faster than conventional methods), while the provided services offer responses in a few seconds. These services enable the implementation of other high level services, such as services for Data Enrichment which can be exploited for Machine-Learning tasks, and techniques for Knowledge Graph Embeddings, and we show that this enrichment improves the prediction of machine-learning problems.

Supervisor: Associate Professor, Ioannis Tzitzikas