**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.  Βασίλειου Ευθυμίου**

Την Τετάρτη, 27 Σεπτεμβρίου 2017 και ώρα 15:00 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. Βασίλη Ευθυμίου με θέμα:

**“ Ανάλυση Οντοτήτων στον Παγκόσμιο Ιστό των Δεδομένων”**

**“Entity Resolution in the Web of Data”**

**ΠΕΡΙΛΗΨΗ**

Η ανάλυση οντοτήτων είναι το πρόβλημα της αναγνώρισης περιγραφών των ίδιων οντοτήτων του πραγματικού κόσμου ανάμεσα σε διαφορετικές βάσεις γνώσης. Σε αυτή τη διδακτορική εργασία, μελετάμε το πρόβλημα την ανάλυσης οντοτήτων στον Παγκόσμιο Ιστό των Δεδομένων,στον οποίο οι οντότητες περιγράφονται μέσω RDF γράφων, ακολουθώντας τις αρχές των Διασυνδεδεμένων Δεδομένων. Τα δύο κεντρικά προβλήματα της ανάλυσης οντοτήτων είναι: (α) πώς μπορούμε να υπολογίσουμε την ομοιότητα οντοτήτων αποτελεσματικά, και (β) πώς μπορούμε να αναλύσουμε σύνολα οντοτήτων εντός ή μεταξύ των βάσεων γνώσης αποδοτικά. Σε σχέση με την απαλοιφή διπλοτύπων περιγραφών οντοτήτων σε σχεσιακές βάσεις, οι νέες προκλήσεις για αυτά τα προβλήματα πηγάζουν από την Ποικιλία (πολλαπλοί τύποι οντοτήτων και διαθεματικές περιγραφές), τον Όγκο (χιλιάδες βάσεις γνώσης στον Παγκόσμιο Ιστό με δισεκατομμύρια γεγονότα, που φιλοξενούν εκατομμύρια περιγραφές οντοτήτων), και την Εγκυρότητα (πολλές μορφές ασυνέπειας και λαθών) των περιγραφών οντοτήτων που δημοσιεύονται στον Παγκόσμιο Ιστό των Δεδομένων.

Στον πυρήνα της ανάλυσης οντοτήτων βρίσκεται η διαδικασία λήψης της απόφασης για το αν ένα δοθέν ζευγάρι περιγραφών αναφέρονται στην ίδια πραγματική οντότητα, δηλαδή αν ταιριάζουν (πρόβλημα α). Η απόφαση ταιριάσματος συνήθως εξαρτάται από την εκτίμηση της ομοιότητας δύο περιγραφών, με βάση το περιεχόμενο ή ακόμα και τις γειτονικές τους περιγραφές (για οντότητες διαφορετικών τύπων). Αυτή η διαδικασία είναι συνήθως επαναληπτική, καθώς οι περιγραφές που ταιριάζονται σε μία επανάληψη βοηθούν στη λήψη αποφάσεων ταιριάσματος σε επόμενες επαναλήψεις, χρησιμοποιώντας διάδοση ομοιότητας, έως ότου να μην βρίσκονται άλλες περιγραφές που να ταιριάζουν. Το πλήθος των απαιτούμενων για τη σύγκλιση επαναλήψεων εξαρτάται από το μέγεθος και την πολυπλοκότητα των συλλογών περιγραφών οντοτήτων. Επιπλέον, το ταίριασμα ζευγαριών περιγραφών είναι εκ φύσεως τετραγωνικής πολυπλοκότητας ως προς το πλήθος των περιγραφών και άρα απαγορευτικό στην κλίμακα του Παγκόσμιου Ιστού (πρόβλημα β). Στο πλαίσιο αυτό, η συσταδοποίηση έχει στόχο να αποτρέψει όσο το δυνατόν περισσότερες συγκρίσεις, χωρίς να χαθούν ταιριαστές περιγραφές. Τοποθετεί τις περιγραφές οντοτήτων σε επικαλυπτόμενες ή μη-επικαλυπτόμενες συστάδες, προωθώντας στη φάση ταιριάσματος τις συγκρίσεις μόνο μεταξύ περιγραφών που έχουν τοποθετηθεί σε κάποια κοινή συστάδα. Οι μέθοδοι επικαλυπτόμενης συσταδοποίησης συνοδεύονται από τεχνικές Μετα-συσταδοποίησης, που έχουν ως στόχο την αποτροπή των επαναλαμβανόμενων συγκρίσεων που προτείνονται από πολλαπλές συστάδες, καθώς και των συγκρίσεων μεταξύ περιγραφών που είναι πιθανότερο να μην ταιριάζουν, αλλά έχουν προταθεί λόγω ύπαρξης θορύβου στις περιγραφές οντοτήτων.

Για να αντιμετωπίσουμε το πρόβλημα της ανάλυσης οντοτήτων στην κλίμακα του Παγκόσμιου Ιστού, χρειάζεται να χαλαρώσουμε ένα πλήθος υποθέσεων που υπόκεινται πολλών μεθόδων και τεχνικών, οι οποίες έχουν προταθεί στις ερευνητικές κοινότητες των βάσεων δεδομένων, της μηχανικής μάθησης και του σημασιολογικού Ιστού. Συνολικά, τα χαρακτηριστικά Μεγάλων Δεδομένων που εμφανίζουν οι περιγραφές οντοτήτων στον Παγκόσμιο Ιστό των Δεδομένων απαιτούν νέα συστήματα ανάλυσης οντοτήτων που να υποστηρίζουν: (i) σχεδόν ομοιότητα περιγραφών (αναγνωρίζουν περιγραφές που ταιριάζουν και έχουν χαμηλή ομοιότητα περιεχομένου), (ii) ανεξαρτησία ύπαρξης σχήματος (δεν στηρίζονται στην ύπαρξη ενός συγκεκριμένου συνόλου γνωρισμάτων που χρησιμοποιούνται από όλες τις περιγραφές), (iii) πλήρη αυτοματοποίηση (δεν στηρίζονται σε ειδικούς της εκάστοτε περιοχής για δεδομένα εκμάθησης, αντιστοίχιση σχέσεων, κανόνες ταιριάσματος), (iv) μη-επαναληπτικότητα (οι επαναληπτικές μέθοδοι συγκλίνουν μετά από υπερβολικά πολλές επαναλήψεις στον Παγκόσμιο Ιστό των Δεδομένων), και (v) κλιμακωσιμότητα σε πολύ μεγάλους όγκους δεδομένων (απαιτούνται μαζικά παραλληλοποιήσιμες αρχιτεκτονικές).

Για να ικανοποιήσουμε τις απαιτήσεις ανάλυσης οντοτήτων στην κλίμακα του Παγκόσμιου Ιστού, εισάγουμε το σύστημα MinoanER. Το σύστημά μας εκμεταλλεύεται νέες μετρικές ομοιότητας για την εκτίμηση των ενδείξεων ταιριάσματος τόσο από το περιεχόμενο όσο και από τις γειτονιές των περιγραφών, χωρίς να απαιτεί πρότερη γνώση ή αντιστοίχιση των τύπων των οντοτήτων. Αυτές οι μετρικές επιτρέπουν μια συμπαγή αναπαράσταση των ενδείξεων ομοιότητας που μπορούν να αποκτηθούν από διαφορετικά σχέδια συσταδοποίησης πάνω στα ονόματα και τις τιμές των περιγραφών, καθώς επίσης και στις τιμές των γειτονικών τους περιγραφών. Αυτό επιτρέπει την αναγνώριση σχεδόν όμοιων περιγραφών που ταιριάζουν νωρίς, από το βήμα της συσταδοποίησης. Η σύνθετη αυτή συσταδοποίηση, ακολουθούμενη από μία νέα σύνθετη Μετα-συσταδοποίηση που αποτυπώνει τις ενδείξεις ομοιότητα από διαφορετικού τύπου συστάδες, θέτουν τις βάσεις για ένα μη-επαναληπτικό ταίριασμα. Ο αλγόριθμος ταιριάσματος, σχεδιασμένος με μία μαζικά παράλληλη αρχιτεκτονική, χρησιμοποιεί υπολογιστικά φτηνές ευριστικές μεθόδους για να αναγνωρίσει περιγραφές που ταιριάζουν σε ένα προκαθορισμένο πλήθος βημάτων. Η κύρια συνεισφορά του MinoanER είναι ότι πετυχαίνει τουλάχιστον ισάξια αποτελέσματα σε ομοιογενείς βάσεις γνώσης (που έχουν κοινές πηγές και συνεπώς περιέχουν πολύ όμοιες περιγραφές οντοτήτων), και σημαντικά καλύτερα αποτελέσματα σε ανομοιογενείς βάσεις γνώσης (που έχουν διαφορετικές πηγές και συνεπώς περιέχουν λιγότερο όμοιες περιγραφές), σε σχέση με συστήματα αιχμής στην ανάλυση οντοτήτων, χωρίς να απαιτεί οποιαδήποτε γνώση ενός συγκεκριμένου πεδίου, με μη-επαναληπτικό και εξαιρετικά αποδοτικό τρόπο.

 Επιβλέπων: Καθηγητής, Βασίλης Χριστοφίδης

**ABSTRACT**

Entity resolution (ER) is the problem of identifying descriptions of the same real-world entities among or within knowledge bases (KBs).  In this PhD thesis, we study the problem of ER in the Web of data, in which entities are described using graph-structured RDF data, following the principles of the Linked Data paradigm. The two core ER problems are: (a) how can we effectively compute similarity of Web entities, and (b) how can we efficiently resolve sets of entities within or across KBs. Compared to deduplication of entities described by tabular data, the new challenges for these problems stem from the Variety (i.e., multiple entity types and cross-domain descriptions), the Volume (i.e., thousands of Web KBs with billions of facts, hosting millions of entity descriptions) and Veracity (i.e., various forms of inconsistencies and errors) of entity descriptions published in the Web of data.

At the core of an ER task lies the process of deciding whether a given pair of descriptions refer to the same real-world entity i.e., if they match (problem a). The matching decision typically depends on the assessment of the similarity of two descriptions, based on their content or their neighborhood descriptions (i.e., of related entity types). This process is usually iterative, as matches found in one iteration help the decisions at the next iteration, via similarity propagation until no more matches are found. The number of iterations to converge clearly depends on the size and the complexity of the resolved entity collections. Moreover, pairwise entity matching is by nature quadratic to the number of entity descriptions, and thus prohibitive at the Web scale (problem b). In this respect, blocking aims to discard as many comparisons as possible without missing matches. It places entity descriptions into overlapping or disjoint blocks, leaving to the matching phase comparisons only between descriptions belonging to the same block. For this reason, overlapping blocking methods are accompanied by Meta-blocking filtering techniques, which aim to discard comparisons suggested by blocking that are either repeated (i.e., suggested by different blocks) or unnecessary (i.e., unlikely to result in matches) due to the noise in entity descriptions.

To address ER at the Web-scale, we need to relax a number of assumptions underlying several methods and techniques proposed in the context of database, machine learning and semantic Web communities. Overall, the Big Data characteristics of entity descriptions in the Web of data call for novel ER frameworks supporting: (i) near similarity (identify matches with low similarity in their content), (ii) schema-free (do not rely on a given set of attributes used by all descriptions), (iii) no human in the loop (do not rely on domain-experts for training data, aligned relations, matching rules), (iv) non-iterative (avoiding data-convergence methods at several iteration steps), and (v) scalable to very large volumes of entity collections (massively parallel architecture needed).

To satisfy the requirements of a Web-scale ER, we introduce the MinoanER framework. Our framework exploits new similarity metrics for assessing matching evidence based on both the content and the neighbors of entities, without requiring knowledge or alignment of the entity types. These metrics allow for a compact representation of similarity evidence that can be obtained from different blocking schemes on the names and values of the descriptions, but also on the values of their entity neighbors. This enables the identification of nearly similar matches even from the step of blocking. This composite blocking, accompanied by a novel composite Meta-blocking capturing the similarity evidence from the different types of blocks, set the ground for a non-iterative matching. The matching algorithm, built on a massively parallel architecture, is equipped with computationally cheap heuristics to detect matches in a fixed number of steps. The main contribution of MinoanER is that it achieves at least equivalent results over homogeneous KBs (stemming from common data sources, thus exhibiting strongly similar matches) and significantly better results over heterogeneous KBs (stemming from different sources, thus exhibiting many nearly similar matches) to state-of-the-art ER tools, without requiring any domain-specific knowledge, in a non-iterative and highly efficient way.

Supervisor: Professor, Vasilis Christophides