



ΕΛΙΔΕΚ.
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

Περιγραφή Χρηματοδοτούμενου Ερευνητικού Έργου
2η Προκήρυξη Ερευνητικών Έργων ΕΛ.ΙΔ.Ε.Κ. για την
ενίσχυση Μεταδιδακτορικών Ερευνητών/τριών

Τίτλος Ερευνητικού Έργου: Responsible by Design Entity Resolution

Επιστημονικός/ή Υπεύθυνος/η: Βασίλης Ευθυμίου

Φιλικός προς τον αναγνώστη τίτλος: ResponsibleER

Επιστημονική Περιοχή: Μαθηματικά και Επιστήμες της Πληροφορίας

Φορέας Προέλευσης και Χώρα: IBM Research, ΗΠΑ

Φορέας Υποδοχής: Ίδρυμα Τεχνολογίας και Έρευνας (ΙΤΕ)

Συνεργαζόμενος Φορέας: Πανεπιστήμιο Tampere, Φινλανδία

**Ιστοσελίδα προβολής του Έργου
(αν υφίσταται): <https://isl.ics.forth.gr/ResponsibleER/>**



Ποσό Χρηματοδότησης: 130.000 €

Διάρκεια Χρηματοδότησης: 36 μήνες

Σύνοψη Ερευνητικού Έργου

Η διαχείριση μεγάλων δεδομένων υπόσχεται να επιφέρει σημαντική βελτίωση στη ζωή μας, επιταχύνοντας την ανακάλυψη γνώσης, την έρευνα και την καινοτομία. Ωστόσο, τα τελευταία χρόνια υπάρχει μια αυξανόμενη ανησυχία σχετικά με την έλλειψη αμεροληψίας, ποικιλομορφίας και διαφάνειας αλγορίθμων που βασίζονται σε δεδομένα που υποστηρίζουν τη λήψη αποφάσεων, αυξάνοντας την ανάγκη για αυτοματοποιημένα συστημάτων λήψης αποφάσεων που είναι εκ σχεδιασμού υπεύθυνα. Μέχρι στιγμής, οι προσπάθειες για υπεύθυνη λήψη αποφάσεων έχουν ως επί το πλείστον επικεντρωθεί στους αλγόριθμους Μηχανικής Μάθησης, υποθέτοντας ότι αυτοί έχουν εκπαιδευτεί σε δεδομένα υψηλής ποιότητας, αγνοώντας τους υποκείμενους περίπλοκους αγωγούς που μπορεί να έχουν παράγει τέτοια δεδομένα. Ένας βασικός αγωγός δεδομένων για την παραγωγή τέτοιων δεδομένων είναι η ανάλυση οντοτήτων (ER), η οποία ανακαλύπτει και ενοποιεί περιγραφές που αντιστοιχούν στις ίδιες πραγματικές οντότητες. Σε αυτό το έργο, στοχεύουμε συστήματα ER που είναι υπεύθυνα από το σχεδιασμό, ιδίως όταν οι αποφάσεις σχετικά με το ποιες περιγραφές οντοτήτων πρέπει να αναλυθούν πρώτα λαμβάνονται σε σχέση με ένα δοθέντα προϋπολογισμό. Οι στόχοι του ResponsibleER είναι: (α) να εμπλουτιστεί η ποικιλομορφία των αναλυμένων οντοτήτων, (β) να διασφαλιστεί η δικαιοσύνη των αναλυμένων οντοτήτων και (γ) να ενισχυθεί η διαφάνεια των συστημάτων ER. Για τα (α) και (β), μας ενδιαφέρει να διασφαλίσουμε ότι όλες οι ομάδες οντοτήτων εκπροσωπούνται δίκαια στα αποτελέσματα του ER. Για το (γ), πρέπει να επεκτείνουμε τους δείκτες που χρησιμοποιούνται στα (α) και (β) και να παρέχουμε ουσιαστικές εξηγήσεις σχετικά με τις ενδιάμεσες αποφάσεις που λαμβάνονται κατά τη διάρκεια μιας διαδικασίας ER (π.χ. ευρετηρίαση). Ενώ και τα τρία προβλήματα έχουν οριστεί ως σημαντικές προκλήσεις πρόσφατα από τις ρυθμιστικές αρχές της ΕΕ και των ΗΠΑ, από όσο γνωρίζουμε, δεν υπάρχει άλλη εργασία στο ER που να έχει μελετήσει αυτούς τους στόχους.

Πρωτοτυπία του Ερευνητικού Έργου

Τα υπάρχοντα συστήματα ER στοχεύουν στη μεγιστοποίηση του πλήθους των σωστών αντιστοιχιών, ευνοώντας την ανάλυση πολύ παρόμοιων, έναντι λιγότερο όμοιων, περιγραφών οντοτήτων, αγνοώντας τα ποιοτικά χαρακτηριστικά των δεδομένων που προκύπτουν. Αυτά τα ζητήματα καθίστανται ακόμη σημαντικότερα στο προοδευτικό ER, όπου μόνο ένα υποσύνολο των οντοτήτων εισόδου, εκείνα που αναμένεται να μεγιστοποιήσουν ένα στοχευόμενο όφελος (π.χ., αριθμός σωστών αντιστοιχιών), μπορούν να επιλεγούν για επίλυση προκειμένου να τηρηθεί ένας δοθείς προϋπολογισμός. Υποστηρίζουμε ότι ένα υπεύθυνο ER απαιτεί να διασφαλιστεί η ποικιλομορφία και η αμεροληψία κατά την ανάλυση ενός γραφήματος οντοτήτων, καθώς και η διαφάνεια των βασικών μονάδων επεξεργασίας ER, όπως η δημιουργία ευρετηρίου οντοτήτων και η αντιστοίχιση. Η κύρια καινοτομία του ResponsibleER είναι η αντιστοίχιση περιγραφών χαμηλής ομοιότητας, επιτρέποντας τον καλύτερο έλεγχο του βάθους (δηλ. μονοπάτια οντοτήτων) και του εύρους (δηλ. χαρακτηριστικών των οντοτήτων) της γνώσης που κωδικοποιείται στα διαγράμματα οντοτήτων που έχουν αναλυθεί.

Σε αυτήν την εργασία μας ενδιαφέρει να εξαλείψουμε τις διακρίσεις στην αντιστοίχιση οντοτήτων στα πλαίσια ενός προοδευτικού αλγορίθμου ER. Από όσο γνωρίζουμε, δεν υπάρχει εργασία στο ER που να στοχεύει την αμεροληψία από το σχεδιασμό. Προσπαθούμε επίσης να γεφυρώσουμε το χάσμα μεταξύ του προβλήματος βελτιστοποίησης της εύρεσης των αντίστοιχων οντοτήτων και των ποιοτικών χαρακτηριστικών του γραφήματος οντοτήτων που προκύπτουν μετά τη συγχώνευσή τους. Τέλος, εστιάζουμε στην παροχή ρητών, αιτιολογικών εξηγήσεων των αποφάσεων ER, με μεγάλη συνοχή. Στόχος μας είναι να ενσωματώσουμε έναν επεξηγηματικό σχεδιασμό για τη φάση προγραμματισμού του προοδευτικού ER, στο οποίο η επιλογή του πλάνου σύγκρισης (δηλ. ποια σύγκριση θα εκτελεστεί στη συνέχεια) θα πρέπει να αιτιολογείται σωστά έναντι εναλλακτικών πλάνων.

Αναμενόμενα αποτελέσματα & Αντίκτυπος του Ερευνητικού Έργου

Παρά τα κοινωνικά οφέλη του σημερινού κόσμου που βασίζεται στα δεδομένα, οι ρυθμιστικές αρχές της ΕΕ και των ΗΠΑ, καθώς και πολιτικοί και νομικοί μελετητές υποστηρίζουν ότι οι εταιρείες που συλλέγουν και χρησιμοποιούν τα δεδομένα μας έχουν την ευθύνη να διασφαλίσουν τη δίκαιη μεταχείριση των χρηστών. Είναι σαφές ότι μεροληπτικά αποτελέσματα αλγορίθμων ανάλυσης μπορούν να προσβάλουν ή ακόμη και να βλάψουν τους χρήστες και να προκαλέσουν δυσπιστία και ενδεχομένως απώλεια εσόδων. Το ResponsibleER στοχεύει στην ερμηνεία των εσωτερικών αναπαραστάσεων και αποφάσεων ER μοντέλων που λειτουργούν ως μαύρα κουτιά και θέτει τα θεμέλια για συστήματα ER που από το σχεδιασμό να μπορούν να ελεγχθούν για αμεροληψία. Αναμένεται το ResponsibleER να έχει βραχυπρόθεσμο και μακροπρόθεσμο αντίκτυπο στην επιστημονική κοινότητα, αλλά ευελπιστούμε ότι θα έχει και κοινωνικο-οικονομικό αντίκτυπο μακροπρόθεσμα. Το όραμά μας είναι να επισημάνουμε στο κοινό και στους ερευνητές, τους πιθανούς κοινωνικούς κινδύνους που μπορεί να φέρει το ER και να θέσουμε τα θεμέλια για μια νέα γενιά υπεύθυνων συστημάτων ER που ο καθένας να μπορεί να εμπιστευθεί. Όσον αφορά τον επιστημονικό αντίκτυπο, το ResponsibleER είναι ένα φιλόδοξο έργο από πολλές πτυχές, καθώς αναμένεται να: α) εισαγάγει το προοδευτικό ER που βασίζεται στην ποικιλομορφία, ένα πρόβλημα που δεν έχει μελετηθεί ποτέ, β) είναι το πρώτο έργο ER που εκ σχεδιασμού στοχεύει την αμεροληψία, γ) είναι το πρώτο έργο ER που εκ σχεδιασμού παρέχει ρητές αιτιολογικές εξηγήσεις, και δ) θέσει τις βάσεις για υπεύθυνο ER.

Καθώς τα αποτελέσματα του έργου μετατρέπονται σε ώριμα εργαλεία λογισμικού, σταδιακά θα προσφέρουν δυνατότητες εμπορικής εκμετάλλευσης. Η εφαρμογή του πρωτοτύπου μας θα είναι διαθέσιμη ως λογισμικό ανοιχτού κώδικα, με άδεια εμπορικής χρήσης, ώστε επιχειρήσεις να μπορούν να τα χρησιμοποιούν για την ανάπτυξη υπεύθυνων εφαρμογών ER.

Η σημασία της χρηματοδότησης

Ως ΕΥ αυτού του έργου, η χρηματοδότηση του ResponsibleER έχει βαθύ αντίκτυπο στην καριέρα μου για δύο βασικούς λόγους. Πρώτον, θέτει τις βάσεις για έρευνα που στοχεύει να αλλάξει την τρέχουσα εστίαση του ER και να συμπεριλάβει μια πτυχή υπευθυνότητας, η οποία έλειπε από το πεδίο, οδηγώντας σε σοβαρές κοινωνικές επιπτώσεις. Ως ερευνητές, είναι σημαντικό να εργαζόμαστε για έργα που πιστεύουμε ότι θα φέρουν μια θετική αλλαγή στον (ψηφιακό) κόσμο γύρω μας. Δεύτερον, η χρηματοδότηση αυτού του έργου, μου έδωσε την ευκαιρία να επιστρέψω στην Ελλάδα σε ένα σημείο καμπής στην καριέρα μου, μετά από ένα μεταδιδακτορικό στην IBM Research - Almaden, ΗΠΑ, όταν η εναλλακτική λύση ήταν να συνεχίσω την καριέρα μου στο εξωτερικό (πιθανόν μόνιμα). Η ευκαιρία που δόθηκε από τον ΕΛΙΔΚ να συνεχίσω την έρευνά μου στην Ελλάδα ως ΕΥ αυτού του ερευνητικού προγράμματος και να αποκτήσω νέες γνώσεις, εμπειρίες και δεξιότητες ήταν καθοριστικής σημασίας.



ΕΛΙΔΕΚ.
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

ΕΠΙΚΟΙΝΩΝΙΑ

Λ. Συγγρού 185 & Σάρδεων 2
ΤΚ. 17121, Νέα Σμύρνη, Ελλάδα
210 64 12 410, 420
communication@elidek.gr
www.elidek.gr