Έργο:	**«ΘΑΛΗΣ: Ενίσχυση της Διεπιστημονικής ή και Διιδρυματικής έρευνας και καινοτομίας με δυνατότητα προσέλκυσης ερευνητών υψηλού επιπέδου από το εξωτερικό μέσω της διενέργειας βασικής και εφαρμοσμένης έρευνας αριστείας»**
Τίτλος Υποέργου:	**«ΕΙΚΟΣ»: Θεωρητική και αλγοριθμική θεμελίωση για Προσωποκεντρικά Συνεργατικά Πληροφοριακά Συστήματα**

Παραδοτέο Π.1.2

Τεχνικές ανάκτησης πληροφορίας από πηγές μη-παραδοσιακών δεδομένων

Σεπτέμβριος 2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Δράση 1	**Ανάκτηση και αναπαράσταση μη-παραδοσιακών δεδομένων**				
Ομάδα	Ερ. Ομάδα 1	**Έναρξη**	01/02/2012	**Λήξη**	30/11/2015
Συντονιστής ΕΟ1	Δ. Πλεξουσάκης (Παν. Κρήτης)				
Υποδράση: ΥΔ 1.2	Τεχνικές ανάκτησης πληροφορίας από πηγές μη-παραδοσιακών δεδομένων				
Συμμετέχοντες	*Μέλη ΚΕΟ*	Δ. Πλεξουσάκης, Ι. Θεοδωρίδης, Σ. Σκιαδόπουλος			
	Μέλη ΟΕΣ	Θ. Πάτκος (Παν. Κρήτης), Ν. Πελέκης (Παν. Πειραιώς), Ε. Ζερβάκης (Παν. Πελοποννήσου)			
Σύντομη Περιγραφή	Η Υποδράση ΥΔ1.2 ερευνά μοντέλα και αλγοριθμικές μεθόδους που επιτρέπουν στους χρήστες ενός οικοσυστήματος να εκμαιεύουν την απαραίτητη πληροφορία και μεταπληροφορία για την ένταξη των δεδομένων στον υπερχώρο δεδομένων.				
Παραδοτέο	Π.1.2 Τεχνικές ανάκτησης πληροφορίας από πηγές μη-παραδοσιακών δεδομένων				
Στόχος στο Τ.Δ.	Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 2 δημοσιεύσεις.				
Επίτευξη στόχου	100%				

Περιεχόμενα

1 Εισαγωγή 7

1.1 Πλαίσιο έρευνας 8

1.2 Κίνητρα της έρευνας και κεντρική ιδέα 10

2 Οντολογία για τον εμπλουτισμό τροχιών κινούμενων αντικειμένων 10

3 Ανάκτηση πληροφοριών από πηγές με εξασφάλιση της σημασιολογίας και της ανωνυμίας των δεδομένων 12

4 Ανάκτηση πληροφοριών από δεδομένα γράφων και ιατρικά δεδομένα. 16

5 Ανακεφαλαίωση 17

1 Εισαγωγή

Ο βασικός στόχος του έργου ΕΙΚΟΣ είναι να προσφέρει τη μεθοδολογία, τη θεωρητική θεμελίωση, τις αλγοριθμικές τεχνικές και την αρχιτεκτονική του λογισμικού που απαιτείται ώστε τα πληροφοριακά συστήματα να μπορούν να προσφέρουν στους χρήστες αφενός τη δυνατότητα εξατομίκευσης της παρεχόμενης πληροφορίας και αφετέρου τη δυνατότητα χρήσης ενσωματωμένων ετερογενών δεδομένων, ενδεχομένως διαφορετικής προέλευσης, με διαφανή τρόπο.

Στα πλαίσια του έργου, η Δράση 1 «Ανάκτηση και αναπαράσταση μη-παραδοσιακών δεδομένων» έχει σκοπό να παρέχει την διεπαφή του συστήματος με τα μη-παραδοσιακά δεδομένα. Σε αυτό το πλαίσιο αναγνωρίζονται τρεις βασικές επιδιώξεις οι οποίες θα αποτελούν τις 3 θεμελιώδεις υποδράσεις της συγκεκριμένης δράσεις: (α) η δημιουργία ενός επεκτάσιμου και γενικού μοντέλου για μη-παραδοσιακά δεδομένα, (β) η σχεδίαση τεχνικών ανάκτησης πληροφορίας που θα έχουν τη δυνατότητα να εκμαιεύουν την απαραίτητη πληροφορία και μεταπληροφορία για την ένταξη των δεδομένων στον υπερχώρο δεδομένων και (γ) η ανάπτυξη μηχανισμών δεικτοδότησης των μη-παραδοσιακών δεδομένων και της μεταπληροφορίας που τα συνοδεύει, οι οποίοι θα έχουν την δυνατότητα να υποστηρίξουν γενικού τύπου ερωτήσεις που θα παρέχονται στον υπερχώρο των δεδομένων.

Το παρόν παραδοτέο (Π.1.2) περιλαμβάνει τα αποτελέσματα της Υποδράσης ΥΔ1.2, που αφορά την ανάπτυξη τεχνικών ανάκτησης πληροφορίας από πηγές μη-παραδοσιακών δεδομένων. Οι προκλήσεις που ανακύπτουν στην υποδράση αυτή περιλαμβάνουν το χειρισμό της ετερογένειας των δεδομένων σε όλες τις μορφές της (συντακτική, δομική, σημασιολογική) για την αποδοτική ανάκτηση της πληροφορίας και την εισαγωγή και ολοκλήρωσή της στο οικοσύστημα δεδομένων ώστε να υποστηρίζεται αποδοτική και πλούσια σε σημασιολογία αναζήτηση.

Στην Ενότητα 1 παρουσιάζουμε το γενικότερο πλαίσιο του προβλήματος. Στην Ενότητα 2 περιγράφουμε μηχανισμούς εμπλουτισμού δεδομένων τροχιών

κινούμενων αντικειμένων χρησιμοποιώντας οντολογίες και διασυνδεδεμένα δεδομένα με σαφώς καθορισμένη και ευρέως αποδεκτή σημασιολογία, που είναι ήδη διαθέσιμα στο διαδίκτυο. Στην Ενότητα 3 αναπτύσσουμε τεχνικές ανάκτησης πληροφοριών από πηγές πολυδιάστατων δεδομένων που διαφυλάττουν την σημασιολογία των δεδομένων και τις μεταξύ τους συσχετίσεις αλλά και προστατεύουν την ιδιωτικότητα των εμπλεκομένων. Στην Ενότητα 4 αναπτύσσουμε τεχνικές ανάκτησης πληροφοριών που αποκαλύπτουν ιδιότητες και χαρακτηριστικά στη δομή και στη σημασιολογία ιατρικών δεδομένων και γράφων. Ανακεφαλαιώνουμε τα αποτελέσματά μας στην Ενότητα 5.

1.1 Πλαίσιο έρευνας

Στόχος της Υποδράσης ΥΔ1.2 είναι η ανάπτυξη των μηχανισμών που θα ανακτούν από τις πηγές των δεδομένων την απαραίτητη πληροφορία για την ενσωμάτωση τους στον υπερχώρο δεδομένων. Η αντλούμενη πληροφορία υπακούει στο μοντέλο που προδιέγραψε η Υποδράση ΥΔ1.1 για την αναπαράσταση των δεδομένων και διαφοροποιείται ανάλογα με την πηγή και την πληροφορία που αυτή προσφέρει. Η αντλούμενη πληροφορία μπορεί να δεικτοδοτηθεί με τους τρόπους που περιγράφονται στην Υποδράση ΥΔ1.3.

Στα πλαίσια της Υποδράσης ΥΔ1.2 θα αντλείται πληροφορία σε σχέση με τις ακόλουθες δύο βασικές όψεις των πηγών.

Σε σχέση με τις πηγές καθαυτές. Η πληροφορία που αντλείται αφορά απλές καταγραφές που αφορούν μόνο στην πηγή. Για παράδειγμα, το όνομα της, η τελευταία ώρα ενημέρωσης, περιγραφή των δυνατοτήτων της, το όνομα του ιδιοκτήτη των δεδομένων της κτλ. Αφορά επίσης και πιο σύνθετες σχέσεις της πηγής με άλλες πηγές. Για παράδειγμα, μία πηγή μπορεί να είναι εμφωλευμένη σε μία άλλη, μπορεί να ενημερώνεται ή να ενημερώνει κάποια άλλη, μπορεί να είναι προϊόν επεξεργασίας μίας άλλη πηγής, ή να είναι συναθροιστής διάφορων άλλων πηγών. Οι πληροφορίες είναι ιδιαίτερα χρήσιμες για τη μοντελοποίηση των δεδομένων και πρέπει να επισημειώνονται.

Σε σχέση με τα δεδομένα των πηγών. Οι πηγές έχουν εν γένει ετερόκλητα μοντέλα περιγραφής δεδομένων που κυμαίνονται από πολύ αυστηρά, π.χ.,

σχεσιακό, ως πολύ χαλαρά, π.χ., κείμενο. Οι τεχνικές ανάκτησης πληροφορίας αντλούν και ολοκληρώνουν τα δεδομένα με όλα τα απαραίτητα στοιχεία τους ώστε να ενταχθούν απρόσκοπτα στον υπερχώρο των δεδομένων. Οι τεχνικές που προτείνονται υποστηρίζουν ερωτήσεις όρων-κλειδιών (keyword search), δηλαδή ένας χρήστης που δεν γνωρίζει τίποτε για την οργάνωση των δεδομένων μπορεί να ρωτήσει με τον τρόπο που θα ρωτούσε μία μηχανή αναζήτησης στο διαδίκτυο: δίνοντας απλώς τους όρους που τον ενδιαφέρουν. Αν και η σημασιολογία των απαντήσεων δεν αποτελεί το κύριο αντικείμενο της Υποδράσης ΥΔ1.2, οι τεχνικές ανάκτησης της πληροφορίας μπορούν να συλλέξουν όλους τους όρους που εμφανίζονται στις πηγές καθώς και τη μεταπληροφορία που τις συνοδεύει. Μεγαλύτερη πρόκληση απαιτούν όμως οι πιο σύνθετες πηγές από τις οποίες μπορούν να συναχθούν σημασιολογικές συσχετίσεις μεταξύ των δεδομένων. Τα μη-παραδοσιακά δεδομένα αποτελούν μία από τις δυσκολότερες περιπτώσεις καθώς ενυπάρχει σημασιολογία κρυμμένη συνήθως στη δομή τους (π.χ., μία τροχιά αποτελείται από σημεία, ένα στοιχείο σε μία εγγραφή βρίσκεται εμφωλευμένο μέσα σε ένα άλλο) και πρέπει να αναπτυχθούν ειδικές τεχνικές που θα έχουν τη δυνατότητα να ανακτούν και να αποτυπώνουν αυτήν την πληροφορία. Στα πλαίσια της Υποδράσης ΥΔ1.2. επικεντρωθήκαμε στα χωροχρονικά δεδομένα, στα δεδομένα τροχιών, στα δεδομένα ροών και στα δεδομένα συνόλων και συναλλαγών.

Σε σχέση με την ανάκτηση των όρων, αλλά και την πρωτογενή επεξεργασία τους (διόρθωση τυπογραφικών, ταυτοποίηση λέξεων, συνώνυμα) υπάρχει μεγάλος όγκος έρευνας αλλά και εφαρμογών που προσφέρουν αξιόπιστες λύσεις. Οι μεγαλύτερες προκλήσεις υπάρχουν στην ανίχνευση των σημασιολογικών σχέσεων μεταξύ των μη-παραδοσιακών δεδομένων. Οι σχετικές εργασίες είναι αρκετά περιορισμένες και δεν προσφέρουν ολοκληρωμένη λύση στο πρόβλημα της εξαγωγής σημασιολογίας. Επίσης, οι λύσεις αυτές δουλεύουν κάτω από συγκεκριμένα μοντέλα που δεν ταυτίζονται με αυτό που προδιαγράφεται από την Υποδράση ΥΔ1.1. Τέλος, η συσχέτιση μεταξύ ετερογενών δεδομένων, π.χ., χωροχρονικών, σχεσιακών, συνόλων και συναλλαγών, αποτελεί μία εντελώς νέα πρόκληση.

1.2 Κίνητρα της έρευνας και κεντρική ιδέα

Η τεχνικές ανάκτησης πληροφορίας από πηγές μη-παραδοσιακών δεδομένων δεν έχει λυθεί πλήρως στη διεθνή βιβλιογραφία. Ιδιαίτερη πρωτοτυπία εμφανίζει το γεγονός ότι δεν έχουμε κάποιου είδους ταιριάσματος ανάμεσα σε σχήματα του ιδίου τύπου, αλλά την αναζήτηση σημασιολογικών συσχετίσεων σε κάθε είδους δεδομένων που υποστηρίζονται από τον υπερχώρο δεδομένων. Αυτό σημαίνει ότι, π.χ., χωρικές συσχετίσεις θα πρέπει να εξαχθούν τόσο από τις συντεταγμένες σε χωροχρονικά δεδομένα όσο και από τη σαφή δήλωση τους σε ημιδομημένα δεδομένα, π.χ., με τη χρήση ενός στοιχείου «γείτονες».

Ακολουθώντας τη στρατηγική της Υποδράσης ΥΔ1.1 επικεντρωνόμαστε σε χαρακτηριστικές περιπτώσεις μη-παραδοσιακών δεδομένων όπως: τα χωροχρονικά δεδομένα, τα δεδομένα συνόλων και συναλλαγών και τις ροές δεδομένων και αναπτύσσουμε τεχνικές με τις οποίες εξάγεται η πληροφορία και η μεταπληροφορία που θα ενταχθεί στον υπερχώρο δεδομένων. Πιο συγκεκριμένα, στο παραδοτέο αυτό θα παρουσιάσουμε μια οντολογία για το σημασιολογικό εμπλουτισμό τροχιών κινουμένων αντικειμένων (Ενότητα 2) και μηχανισμούς άντλησης πληροφορίας από πηγές πολυδιάστατων δεδομένων, δεδομένων συνόλων και συναλλαγών με τέτοιο τρόπο ώστε να διατηρείται η σημασιολογία και η χρηστικότητά τους και ταυτόχρονα να διασφαλίζεται η ιδιωτικότητα και ο προσωπικός χαρακτήρας των χρηστών και τον δεδομένων τους (Ενότητα 3). Τέλος αναπτύσσουμε τεχνικές ανάκτησης πληροφοριών που αποκαλύπτουν ιδιότητες και χαρακτηριστικά στην δομή και στην σημασιολογία ιατρικών δεδομένων και γράφων.

2 Οντολογία για τον εμπλουτισμό τροχιών κινούμενων αντικειμένων

Η κατανόηση της κίνησης συχνά απαιτεί περαιτέρω πληροφορία και γνώσεις από ότι μπορεί να ληφθεί από τα χωροχρονική ίχνη που αφήνουν οι χρήστες. Παρά την πρόσφατη πρόοδο στη διαχείριση των δεδομένων τροχιών κινούμενων αντικειμένων, εξακολουθεί να υπάρχει χάσμα μεταξύ των χωροχρονικών δεδομένων και της σημασιολογίας που εμπλέκεται σε αυτά. Το χάσμα αυτό δυσχεραίνει την ανάλυση τροχιών που μπορεί να επωφεληθεί από

την αυξανόμενη συλλογή των διασυνδεδεμένων δεδομένων, με σαφώς καθορισμένη και ευρέως αποδεκτή σημασιολογία, που είναι ήδη διαθέσιμη στο διαδίκτυο. Για το σκοπό αυτό, προτείνουμε τη Baquara, μια πλούσια οντολογία που συνδέεται με την αρχιτεκτονική ενός συστήματος και μια προσέγγιση για να περιοριστεί αυτό το κενό. Η οντολογία Baquara λειτουργεί ως εννοιολογικό πλαίσιο για το σημασιολογικό εμπλουτισμό των δεδομένων κίνησης με επισημειώσεις από τα διασυνδεδεμένα δεδομένα. Η προτεινόμενη αρχιτεκτονική και προσέγγιση αποκαλύπτουν νέες δυνατότητες για την ανάλυση τροχιών, με τη χρήση συστημάτων διαχείρισης βάσεων δεδομένων και triple stores που επεκτείνονται με χωρικά δεδομένα και τελεστές. Η εφικτότητα της πρότασης και η εκφραστικότητα της οντολογίας Baquara μέσω ερωτημάτων διερευνώνται σε μια μελέτη περίπτωσης, χρησιμοποιώντας πραγματικά σύνολα τροχιών και διασυνδεδεμένων δεδομένων.

Συνολικά, η παραπάνω εργασία προτείνει μία πρωτότυπη οντολογία η οποία επιτρέπει τον εμπλουτισμό χωρο-χρονικών δεδομένων που αναπαριστούν τροχιές κινούμενων αντικειμένων, με επισημειώσεις από ελεύθερα διασυνδεδεμένα δεδομένα. Επιπρόσθετα, η οντολογία είναι μέρος μίας αρχιτεκτονικής ενός συστήματος το οποίο επιτρέπει την ανάκτηση πλούσιων σε μετα-δεδομένα τέτοιου τύπου δεδομένα δρώντας ουσιαστικά ως ο ενδιάμεσος κρίκος μεταξύ γλωσσών ερωτήσεων για διασυνδεδεμένα δεδομένα και σημασιολογικά επαυξημένων (από το στάδιο εμπλουτισμού) δεδομένων κίνησης.

Επεκτείνοντας την παραπάνω έρευνα, προτείνουμε το πλαίσιο εργασίας Baquara2, το οποίο εκμεταλλευόμενο γνωσιακές βάσεις, εμπλουτίζει σημασιολογικά και αναλύει δεδομένα κίνησης. Παρέχει ένα οντολογικό μοντέλο για τη δόμηση και αφαίρεση δεδομένων κίνησης σε μία πολλαπλή ιεραρχία από προοδευτικά λεπτομερή τμήματα κίνησης και γενικεύει έννοιες όπως αυτές της τροχιάς, της στάσης και της κίνησης. Η Baquara2 περιλαμβάνει επίσης μια γενική και προσαρμόσιμη διεργασία για την επισημείωση δεδομένων κίνησης με έννοιες και αντικείμενα που περιγράφονται σε οντολογίες και συλλογές ανοιχτών διασυνδεδεμένων δεδομένων. Οι προκύπτουσες σημασιολογικές επισημειώσεις επιτρέπουν ερωτήσεις για την ανάλυση των κινήσεων με βάση την ειδική γνώση της εκάστοτε εφαρμογής ή του πεδίου. Το προκύπτον πλαίσιο

εργασίας χρησιμοποιήθηκε σε πειράματα για το σημασιολογικό εμπλουτισμό δεδομένων κίνησης που συλλέχθηκαν από μέσα κοινωνικής δικτύωσης με γεωαναφερόμενα ανοιχτά διασυνδεδεμένα δεδομένα. Τα αποτελέσματα επιτρέπουν ισχυρά ερωτήματα που αναδεικνύουν τις δυνατότητες του Baquara2.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [FKPTR13] που παρουσιάστηκε στο 32nd International Conference on Conceptual Modeling, ER'13 το 2012 και στο άρθρο [FKPTR15] που θα δημοσιευθεί στο περιοδικό Data & Knowledge Engineering.

3 Ανάκτηση πληροφοριών από πηγές με εξασφάλιση της σημασιολογίας και της ανωνυμίας των δεδομένων

Η διαδικασία ανάκτησης πληροφοριών από τις πηγές και η ένταξή τους στον υπερχώρο δεδομένων ενέχει πολλές προκλήσεις ειδικά σε ότι αφορά πηγές που παρέχουν μη-παραδοσιακά δεδομένα (όπως δεδομένα τροχιών, ιατρικών, ιστορικών, ημερολογίων πιστωτικών καρτών και επίσκεψης ιστοτόπων). Σε πολλές περιπτώσεις, τα δεδομένα αυτά έχουν ξεκάθαρη σημασιολογία και συσχετίσεις που πρέπει να διαφυλαχθούν κατά τη μεταφορά των δεδομένων στον υπερχώρο και να είναι δυνατό να επεκταθούν με την χρήση οντολογιών όπως στην Ενότητα 1.

Συνήθως τα παρεχόμενα από τις πηγές δεδομένα περιέχουν ευαίσθητα προσωπικά δεδομένα (π.χ., ασθένειες, θεραπείες ή αγορές χρηστών). Κατά συνέπεια, είναι πολύ σημαντική η ανάκτηση των δεδομένων αυτών και η δημοσίευσή τους στον υπερχώρο δεδομένων με τρόπο που διασφαλίζει την ανωνυμία των εμπλεκομένων. Ταυτόχρονα όμως, πρέπει να εξασφαλίζεται η χρηστικότητα των αποθηκευμένων δεδομένων στον υπερχώρο με την ελαχιστοποίηση της συνολικής απώλειας πληροφορίας, τόσο στην σημασιολογία τους, όσο και στις μεταξύ τους σχέσεις. Με τον τρόπο αυτό τα δεδομένα στα οποία έχει πρόσβαση ο τελικός χρήστης του υπερχώρου είναι πλούσια σε πληροφορία και μπορούν να παράξουν ποιοτικές αναλύσεις.

Στην πρώτη εργασία μας παρουσιάζουμε μια πρωτοπόρα μέθοδο δημοσιοποίησης των πολυδιάστατων δεδομένων που αφενός εξασφαλίζει την ανωνυμία των χρηστών και αφετέρου διατηρεί την ποιότητα της ανάλυσης των αντλούμενων δεδομένων. Οι υπάρχουσες μέθοδοι για τη διασφάλιση της ανωνυμίας (α) προστατεύουν την ιδιωτικότητα γενικεύοντας ή διαγράφοντας δεδομένα ή προσθέτοντας θόρυβο και (β) θεωρούν τη ξεκάθαρη διάκριση μεταξύ ευαίσθητων και μη-ευαίσθητων δεδομένων. Σε πολλές όμως εφαρμογές οι παραπάνω μέθοδοι δεν μπορούν να εφαρμοστούν. Οι μέθοδοι γενίκευσης και διαγραφής αφαιρούν σημαντική πληροφορία από τα δεδομένα. Επιπρόσθετα, τα σημερινά δεδομένα δεν μπορούν ξεκάθαρα να διαχωριστούν σε ευαίσθητα και μη, καθώς κάποια δεδομένα μπορεί να είναι ευαίσθητα για ένα χρήστη και μη-ευαίσθητα για κάποιον άλλο. Με βάση την παρατήρηση αυτή, παρουσιάζουμε μια μέθοδο ανωνυμίας που την ονομάζουμε αποσυσχέτιση η οποία διατηρεί τα αρχικά δεδομένα ανέπαφα αλλά αποκρύπτει την κοινή τους εμφάνιση. Η ανωνυμία εξασφαλίζεται αποσυσχετίζοντας εγγραφές που συμμετέχουν σε συνδυασμούς που μπορούν να αποκαλύψουν την ταυτότητα των χρηστών. Η μέθοδος που παρουσιάζουμε είναι η πρώτη που χρησιμοποιεί αποσυσχέτιση για την εξασφάλιση της ανωνυμίας. Συμπληρώνουμε την εργασία μας με την υλοποίηση της μεθόδου και την πειραματική σύγκριση με αντίστοιχες μεθόδους γενίκευσης, διαγραφής και εισαγωγής θορύβου.

Στη συνέχεια πραγματευόμαστε την ανάκτηση συνόλων δεδομένων που περιέχουν τόσο σχεσιακά όσο και δεδομένα συναλλαγών. Τέτοια δεδομένα χρησιμοποιούνται σε μια ευρεία γκάμα εφαρμογών που εκτείνονται από την υγειονομική περίθαλψη ως και το μάρκετινγκ. Ωστόσο, η διατήρηση της ιδιωτικότητας και της χρηστικότητας αυτών των συνόλων δεδομένων κατά τη δημοσίευσή τους στον υπερχώρο αποτελεί πρόκληση, καθώς απαιτεί (α) προστασία από κακόβουλους χρήστες, των οποίων η γνώση καλύπτει και τους δύο τύπους χαρακτηριστικών, και (β) την ελαχιστοποίηση της συνολικής απώλειας πληροφορίας. Οι υπάρχουσες τεχνικές δεν μπορούν να εφαρμοστούν σε τέτοια σύνολα δεδομένων και το πρόβλημα δεν μπορεί να αντιμετωπιστεί με βάση τις δημοφιλείς στρατηγικές βελτιστοποίησης πολλαπλών στόχων. Κατά συνέπεια προτείνουμε την πρώτη προσέγγιση για την αντιμετώπιση του προβλήματος αυτού. Συγκεκριμένα αναπτύσσουμε δύο πλαίσια

ανωνυμοποίησης, με προκαθορισμένη απώλεια πληροφοριών στον ένα τύπο δεδομένων και ελάχιστη απώλεια πληροφορίας στον άλλο. Για την εφαρμογή κάθε πλαισίου προτείνουμε αλγορίθμους ανωνυμοποίησης οι οποίοι διατηρούν αποτελεσματικά τη χρηστικότητα των δεδομένων, όπως επαληθεύεται και από εκτεταμένα πειράματα.

Στη συνέχεια αναπτύσσουμε το SECRETA, ένα σύστημα για την ανάλυση της αποτελεσματικότητας και της αποδοτικότητας των διαφόρων αλγορίθμων ανάκτησης και δημοσίευσης πληροφοριών με ανώνυμο τρόπο. Το σύστημα που προτείνουμε επιτρέπει σε όσους επιθυμούν να δημοσιεύσουν δεδομένα στον υπερχώρο, να αξιολογήσουν ένα συγκεκριμένο αλγόριθμο, να συγκρίνουν πολλαπλούς αλγορίθμους και να συνδυάσουν αλγορίθμους για την ανωνυμοποίηση συνόλων δεδομένων τα οποία περιέχουν σχεσιακά δεδομένα και δεδομένα συναλλαγών. Η ανάλυση του(ων) αλγορίθμου(ων) γίνεται με διαδραστικό και προοδευτικό τρόπο και τα αποτελέσματα, συμπεριλαμβανομένων των στατιστικών και διαφόρων δεικτών χρηστικότητας των δεδομένων, μπορούν να συνοψιστούν και να παρουσιαστούν γραφικά.

Ακολούθως πραγματευόμαστε την ανάκτηση πληροφοριών από πηγές τροχιών και μελετάμε την ασφαλή και ανώνυμη δημοσίευσή τους στον υπερχώρο δεδομένων. Η δημοσίευση δεδομένων τροχιών παρέχει νέες κατευθυντήριες γραμμές στη μελέτη της ανθρώπινης συμπεριφοράς, αλλά είναι δύσκολο να πραγματοποιηθεί με έναν τρόπο ο οποίος θα διαφυλάττει την ιδιωτικότητα των ατόμων. Αυτό οφείλεται κυρίως στο ότι η ταυτότητα των ατόμων, των οποίων η κίνηση καταγράφεται στα δεδομένα, μπορεί να αποκαλυφθεί ακόμη και μετά την αφαίρεση προσδιοριστικών πληροφοριών. Οι υφιστάμενες μέθοδοι προστασίας της ιδιωτικότητας παρέχουν ανωνυμία, αλλά με υψηλό κόστος στην σημασιολογία και χρηστικότητα των δεδομένων. Αυτό συμβαίνει επειδή συνήθως δεν παράγουν αληθή δεδομένα, που είναι ιδιαιτέρως σημαντικά σε πολλές εφαρμογές. Προτείνουμε μια νέα προσέγγιση που αντιμετωπίζει τις ανεπάρκειες αυτές, χρησιμοποιώντας το μοντέλο της km-ανωνυμίας σε δεδομένα τροχιών. Για να υλοποιήσουμε την προσέγγισή μας, έχουμε αναπτύξει τρεις αποδοτικούς και αποτελεσματικούς αλγόριθμους ανωνυμίας, οι οποίοι βασίζονται στην αρχή Apriori. Αυτοί οι αλγόριθμοι στοχεύουν στη διατήρηση διαφορετικών χαρακτηριστικών των δεδομένων, όπως η απόσταση μεταξύ

σημείων και η σημασιολογική ομοιότητα. Επίσης στοχεύουν στη διατήρηση διαφόρων κριτηρίων χρηστικότητας που ορίζονται από το χρήστη, τα οποία πρέπει να πληρούνται για να διασφαλιστεί ότι τα δημοσιευμένα στοιχεία μπορούν να αναλυθούν ουσιαστικά. Τα εκτεταμένα πειράματά μας, χρησιμοποιώντας συνθετικά και πραγματικά δεδομένα, επαληθεύουν ότι οι προτεινόμενοι αλγόριθμοι είναι αποδοτικοί και αποτελεσματικοί στη διατήρηση της χρηστικότητα των δεδομένων.

Σε συνέχεια για την ανάκτηση πληροφοριών από πηγές τροχιών προτείνουμε ένα νέο πλαίσιο για την ανωνυμοποίηση δεδομένων τροχιών, το οποίο αποτρέπει την αποκάλυψη πληροφοριών τόσο για την ταυτότητα όσο και για τις ευαίσθητες περιοχές που επισκέφτηκε ο χρήστης, διατηρώντας παράλληλα τη χρηστικότητα των δεδομένων. Το πλαίσιό μας περιλαμβάνει: (i) την επιλογή παρόμοιων τροχιών με τη χρησιμοποίηση είτε ενός αλγορίθμου Z-ordering ή χρησιμοποιώντας προβολές δεδομένων στις συχνά εμφανιζόμενες υποτροχιές, (ii) την ομαδοποίηση των επιλεγμένων τροχιών σε προσεκτικά κατασκευασμένες ομάδες και (iii) την ανωνυμοποίηση κάθε τέτοιας ομάδας ξεχωριστά. Έχουμε αναπτύξει αλγορίθμους για την υλοποίηση του πλαισίου μας οι οποίοι είναι αποτελεσματικοί και αποδοτικοί, όπως επαληθεύεται και από εκτεταμένα πειράματα.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [TLMS12] που παρουσιάστηκε στο Very Large Data Bases (VLDB) Conference (5) το 2012, στο άρθρο [PLGS13] που παρουσιάστηκε στο European Conference on Machine Learning and Knowledge Discovery in Databases, ECML/PKDD (3) το 2013, στο άρθρο [PGLST13] που παρουσιάστηκε στο 17th International Conference on Extending Database Technology, EDBT το 2014, στο άρθρο [PSLG13] που παρουσιάστηκε στο PriSMO: Privacy and Security for Moving Objects το 2013, στο άρθρο [PSLG13a] που παρουσιάστηκε στο IEEE International Workshop on Privacy Aspects of Data Mining (PADM) και στο άρθρο [PSLG14] που δημοσιεύτηκε στο περιοδικό Transactions on Data Privacy το 2014.

4 Ανάκτηση πληροφοριών από δεδομένα γράφων και ιατρικά δεδομένα.

Κατά τη διαδικασία ένταξης και δημοσίευσης πολυδιάστατων δεδομένων (π.χ., γράφοι, ιατρικά δεδομένα) στον υπερχώρο δεδομένων είναι πολύ σημαντική η ταυτόχρονη εξαγωγή χρήσιμων ιδιοτήτων και χαρακτηριστικών της δομής και τοπολογίας τους που καθορίζουν πολλές από τις ιδιότητες και τη σημασιολογία τους. Για το σκοπό αυτό μελετάμε δεδομένα γράφων και αναζητούμε σε αυτούς υπογράφους και πρότυπα με συχνή εμφάνιση. Επίσης, μελετάμε ιατρικά δεδομένα και εξετάζουμε τρόπους και μηχανισμούς για την εξαγωγή πληροφορίας τόσο ως σύνολο εγγραφών όσο και γραφικά.

Πιο συγκεκριμένα, η αναζήτηση συχνών υπογράφων είναι ένας πολύ σημαντικός τελεστής σε γράφους που επιστρέφει όλους τους υπογράφους που εμφανίζονται περισσότερες φορές από ένα κατώφλι το οποίο ορίζεται από το χρήστη. Οι περισσότερες λύσεις στο παραπάνω πρόβλημα υποθέτουν μια βάση με πολλούς μικρούς γράφους. Οι σύγχρονες εφαρμογές όμως, όπως τα κοινωνικά δίκτυα, οι γράφοι δημοσιεύσεων και διάδρασης πρωτεϊνών, αναπαρίστανται ως ένας μεγάλος γράφος. Για το σκοπό αυτό, παρουσιάζουμε το GRAMI, ένα νέο πλαίσιο για την αναζήτηση συχνών υπογράφων σε ένα μεγάλο γράφο. Το GRAMI εισάγει μια πρωτοπόρα μέθοδο που βρίσκει μόνο το ελάχιστο σύνολο στιγμιοτύπων που ικανοποιεί το κατώφλι. Με την προτεινόμενη μέθοδο αποφεύγουμε την ακριβή απαρίθμηση όλων των στιγμιοτύπων την οποία χρειάζονται οι παλαιότερες εργασίες. Συνοδεύουμε την πρότασή μας με ευρετική συνάρτηση και βελτιστοποιήσεις που βελτιώνουν σημαντικά την απόδοση. Επιπρόσθετα, παρουσιάζουμε μια επέκταση που αναζητεί συχνά πρότυπα. Σε σχέση με τους υπογράφους τα πρότυπα είναι μια πιο δυνατή και ευέλικτη εκδοχή ταιριάσματος που αναπαριστά μεταβατικές σχέσεις μεταξύ των κόμβων ενός γράφου (όπως για παράδειγμα ο φίλος ενός φίλου) που είναι ιδιαιτέρως συχνές σε σύγχρονες εφαρμογές. Τέλος παρουσιάζουμε το CGRAMI, μια εκδοχή που υποστηρίζει δομικούς και σημασιολογικούς περιορισμούς και το AGRAMI, μια προσεγγιστική εκδοχή που υπολογίζει τα αποτελέσματα χωρίς λάθος-θετικά (false positives) αποτελέσματα. Τα πειράματά μας σε πραγματικά δεδομένα καταδεικνύουν ότι το πλαίσιό μας είναι 2 τάξεις μεγέθους ταχύτερο και ανακαλύπτει πιο ενδιαφέροντα πρότυπα από υπάρχουσες τεχνικές.

Σε ότι αφορά τα ιατρικά δεδομένα, ανάμεσα στα βασικά εργαλεία έρευνας της (βιο)ιατρικής επιστήμης είναι οι επιδημιολογικές έρευνες που τυπικά εμπλέκουν πολλά νοσοκομεία, κλινικές, και ερευνητικά κέντρα εξακτινωμένα σε ολόκληρο τον πλανήτη, οι οποίες καλούνται πολυ-κεντρικές μελέτες. Η αποτελεσματικότητα και η σημασία των πολυ-κεντρικών μελετών αυξάνεται με τον αριθμό των συμμετεχόντων κέντρων και τον αριθμό των εγγεγραμμένων ασθενών. Ταυτόχρονα όμως, η φυσική κατάτμηση στη διαξεγωγή της έρευνας απαιτεί σύνθετους μηχανισμούς και υποδομές διαχείρισης δεδομένων και γνώσης που να υποστηρίζουν τους συμμετέχοντες φορείς. Η υποδομή αυτού του είδους είναι ακριβή να κατασκευαστεί και να διατρηθεί, αλλά επίσης είναι πολύ δύσκολο να επαναχρησιμοποιηθεί καθώς συνήθως είναι διαμορφωμένη για μια συγκεκριμένη έρευνα. Στην εργασία μας παρουσιάζουμε ένα σύστημα βασισμένο στην τεχνολογία νεφών που επιτρέπει σε χρήστες χωρίς τεχνικές γνώσεις πληροφορικής να σχεδιάσουν, αναπτύξουν και διαχειριστούν πλατφόρμες για αξιοποίηση, διαμοιρασμό και ανάλυση κλινικών δεδομένων από πολυ-κεντρικές μελέτες. Το προτεινόμενο σύστημα παρέχει με σχεδόν μηδενικό κόστος διαχείρισης και συντήρησης ένα εργαλείο διαχείρισης δεδομένων και γνώσης που (α) αυξάνει την επαναχρησιμοποίηση εισάγοντας πρότυπα μελετών, (β) υποστηρίζει τις (βιο)ιατρικές ανάγκες διαμέσου ειδικών δομών δεδομένων που αναπαριστούν εξειδικευμένη πληροφορία, όπως θεραπείες και αγωγές και (γ) υποστηρίζει την εξαγωγή και αναζήτηση δεδομένων ενός απλού αλλά πλούσιου σε δυνατότητες γραφικού περιβάλλοντος αναζήτησης.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [EASK14] που παρουσιάστηκε στο Very Large Data Bases (VLDB) Conference (7) το 2014 και στο άρθρο [TTSZ15] που παρουσιάστηκε στο ACM-SIGAI 8th International Conference on Knowledge Capture (K-Cap) το 2015.

5 Ανακεφαλαίωση

Το παρόν παραδοτέο Π1.2 παρουσιάζει τα αποτελέσματα της Υποδράσης ΥΔ1.2 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ5.2 ήταν να δώσει αλγοριθμικές μεθόδους που να επιτρέπουν στους χρήστες ενός οικοσυστήματος να αντλήσουν πληροφορία από τις πηγές της και να την εισάγουν και δημοσιεύσουν στον

υπερχώρο δεδομένων. Στα πλαίσια της έρευνάς μας, λοιπόν, επιτύχαμε να ανταποκριθούμε στο στόχο της υποδράσης με τους ακόλουθους τρόπους:

1. Εμπλουτίσαμε τα δεδομένα τροχιών κινούμενων αντικειμένων χρησιμοποιώντας οντολογίες και διασυνδεδεμένα δεδομένα με σαφώς καθορισμένη και ευρέως αποδεκτή σημασιολογία, η οποία είναι ήδη διαθέσιμα στο διαδίκτυο.
2. Αναπτύξαμε τεχνικές ανάκτησης πληροφοριών από πηγές πολυδιάστατων δεδομένων που διαφυλάττουν τη σημασιολογία τους και τις μεταξύ τους συσχετίσεις και ταυτόχρονα προστατεύουν την ιδιωτικότητα των εμπλεκομένων.
3. Αναπτύξαμε τεχνικές ανάκτησης πληροφοριών που αντλούν ιδιότητες και χαρακτηριστικά για τη δομή και τη σημασιολογία ιατρικών δεδομένων και γράφων.
4. Το λογισμικό που παράγαμε,δημοσιεύεται στα
   http://www.uop.gr/~poulis/SECRETA/
   http://www.uop.gr/~trifon/CloudStudy/

Δημοσιεύσεις

[FKPTR13] R. Fileto, M. Krüger, N. Pelekis, Y. Theodoridis, C. Renso: “Baquara: A Holistic Ontological Framework for Movement Analysis with Linked Data”, Proceedings of the 32nd International Conference on Conceptual Modeling, ER’13, Hong Kong, November 2013. Springer. *Best paper award*.

[FKPTR15] R. Fileto, C. May, C. Renso, N. Pelekis, D. Klein, Y. Theodoridis: “The Baquara2 Knowledge-based Framework for Semantic Enrichment and Analysis of Movement Data”, Data & Knowledge Engineering, to appear. Elsevier.

[TLMS12] M. Terrovitis, J. Liagouris, N. Mamoulis, S. Skiadopoulos: Privacy Preservation by Disassociation. Proceedings of the Very Large Data Bases (VLDB) Conference, 5(10): 944-955 (2012).

[PLGS13] G. Poulis, G. Loukides, A. Gkoulalas-Divanis, S. Skiadopoulos: Anonymizing Data with Relational and Transaction Attributes. Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), 2013: 353-369.

[PGLST13] G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, Christos Tryfonopoulos: SECRETA: A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms. Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014: 620-62.

[PSLG13] G. Poulis, S. Skiadopoulos, G. Loukides, A. Gkoulalas-Divanis: Distance-Based k^m-Anonymization of Trajectory Data. Proceedings of PriSMO: Privacy and Security for Moving Objects (a workshop in conjunction with the 16th IEEE International Conference on Mobile Data Management - MDM) 2013: 57-62.

[PSLG13a] G. Poulis, S. Skiadopoulos, G. Loukides, A. Gkoulalas-Divanis: Select-Organize-Anonymize: A Framework for Trajectory Data Anonymization. Proceedings of the IEEE International Workshop on Privacy Aspects of Data Mining (PADM) of the 13th IEEE International Conference on Data Mining (ICDM) 2013: 867-874.

[PSLG14] G. Poulis, S. Skiadopoulos, G. Loukides, A. Gkoulalas-Divanis: Apriori-based algorithms for km-anonymizing trajectory data. Transactions on Data Privacy 7(2): 165-194 (2014).

[EASK14] M. Elseidy, E. Abdelhamid, S. Skiadopoulos, P. Kalnis: GRAMI: Frequent Subgraph and Pattern Mining in a Single Large Graph. Proceedings of the Very Large Data Bases (VLDB) Conference (7): 517-528 (2014).

[TTSZ15] A. Tsafara, C. Tryfonopoulos, S. Skiadopoulos, L. Zervakis: Cloud-Based Data and Knowledge Management for Multi-Centre Biomedical Studies. Proceeding of the ACM-SIGAI 8th International Conference on Knowledge Capture (K-Cap), 2015.

Παράρτημα