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.4.3

Μέθοδοι εύρεσης ομοιότητας χρηστών υπερχώρων δεδομένων

Σεπτέμβριος 2015

Δράση 4	**Κατανεμημένες Υποδομές Αποθήκευσης, Προσπέλασης και Διαχείρισης Δεδομένων**				
Ομάδα	Ερ. Ομάδα 4	**Έναρξη**	01/06/2012	**Λήξη**	30/11/2015
Συντονιστής ΕΟ4	Παναγιώτης Τριανταφύλλου (Παν. Πατρών)				
Υποδράση: ΥΔ 4.3	Μέθοδοι εύρεσης ομοιότητας χρηστών υπερχώρων δεδομένων				
Συμμετέχοντες	*Μέλη ΚΕΟ*	Παναγιώτης Τριανταφύλλου (Παν. Πατρών), Αντώνιος Δεληγιαννάκης (Πολυτεχνείο Κρήτης) , Βασίλειος Σαμολαδάς (Πολυτεχνείο Κρήτης), Ιωάννης Κωτίδης (ΟΠΑ)			
	Μέλη ΟΕΣ	Δημήτρης Καραμπίνας (Παν. Πατρών), Δημήτρης Μπούσης (Παν. Πατρών), Κυριακή Παναγίδη (Παν. Πατρών), Δημήτρης Σαχαρίδης (ΙΠΣΥ – Ε.Κ. ΑΘΗΝΑ), Κωνσταντίνα Μακρυνιώτη (ΟΠΑ), Κωνσταντίνος Γεωργούλας (ΟΠΑ), Κωνσταντίνος Ζαγγανάς (ΙΠΣΥ – Ε.Κ. ΑΘΗΝΑ), Γεώργιος Ζώης (ΟΠΑ)			
Σύντομη Περιγραφή	Η υποδράση ΥΔ 4.3 μελετάει κλιμακούμενες μεθόδους υπολογισμού της ομοιότητας σε υπερχώρους δεδομένων οι οποίες μπορούν να αξιοποιηθούν κατά την επεξεργασία σύνθετων δεδομένων ενός υπερχώρου. Συγκεκριμένα εξετάσαμε περιπτώσεις όπου η ομοιότητα ανάμεσα στους χρήστες ενός υπερχώρου αποτιμάται μέσω των αποτελεσμάτων συναθροιστικών (αναλυτικών) ερωτημάτων. Επιπλέον εξετάσαμε περιπτώσεις όπου είναι επιθυμητός ο χωρο-χρονικός υπολογισμός της ομοιότητας μεταξύ των χρηστών, μέσω της παρακολούθησης των τροχιών κίνησης τους σε εφαρμογές κινητών αντικειμένων.				
Παραδοτέο	Π.4.3 Μέθοδοι εύρεσης ομοιότητας χρηστών υπερχώρων δεδομένων				

Στόχος στο Τ.Δ.	Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 1 δημοσίευση.
Επίτευξη στόχου	100%

Περιεχόμενα

1 Εισαγωγή

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 την 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 4 με τίτλο «Κατανεμημένες Υποδομές Αποθήκευσης, Προσπέλασης και Διαχείρισης Δεδομένων» σκοπό έχει να παράσχει αρχιτεκτονικές και αλγορίθμους οι οποίοι είτε οι ίδιες θα παρέχουν τρόπους για την κατανεμημένη οργάνωση χώρων δεδομένων στα χαμηλότερα επίπεδα του συστήματος, είτε θα παρέχουν κατάλληλα στατιστικά στοιχεία (όπως συνόψεις δεδομένων που περιγράφουν το φόρτο του συστήματος, ή το ποια δεδομένα ζητούνται από ποιούς χρήστες, τι ρόλο παίζουν διάφοροι χρήστες, κλπ) που να υποβοηθούν την οργάνωση αυτή. Επιπλέον, στόχος είναι οι ίδιοι οι χρήστες να αναβαθμιστούν από απλοί καταναλωτές πληροφορίας σε πρωταγωνιστές που μέσω της συνεργατικότητάς τους να βοηθούν στην κατανόηση των δεδομένων, της σχέσης τους και στην εξατομίκευση των αποτελεσμάτων με βάση το ποιός ερωτά και την «κοινοτική σοφία» αναφορικά με τα περιεχόμενα του οικοσυστήματος.

Η Δράση 4 οργανώνεται στις εξής υποδράσεις: ΥΔ 4.1 Διαχείριση πληροφορίας σε μεγάλης κλίμακας πληροφορικά συστήματα, ΥΔ 4.2 Μέθοδοι για παραγωγή στατιστικών και συνόψεων δεδομένων και ΥΔ 4.3 Μέθοδοι εύρεσης ομοιότητας χρηστών υπερχώρων δεδομένων.

Το παρόν Παραδοτέο Π.4.3 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ 4.3. Στην ενότητα **Error! Reference source not found.** παρουσιάζουμε το ερευνητικό πλαίσιο του προβλήματος όπως αυτό διαμορφώθηκε κατά την εκτέλεση του έργου. Στην ενότητα 3 παρουσιάζουμε συνοπτικά τις τεχνικές και

αλγορίθμους οι οποίες προέκυψαν για την υλοποίηση των στόχων. Τέλος, ανακεφαλαιώνουμε τα αποτελέσματά μας στην ενότητα 4.

2 Διαμόρφωση ερευνητικού πλαισίου

Η Δράση 4 «Κατανεμημένη Υποδομή για Αποθήκευση, Πρόσβαση και Διαχείριση Δεδομένων» απαντά στο ερώτημα: «Πώς αντιμετωπίζουμε την κλιμάκωση του συστήματος (σε αριθμό χρηστών, όγκο δεδομένων και μέγεθος κατανεμημένων υποδομών) μέσω κατανεμημένων τεχνικών;». Σκοπός είναι ο σχεδιασμός αρχιτεκτονικών, αλγορίθμων και υποβοηθητικών δομών δεδομένων (όπως συνόψεις και ευρετήρια) για την κατανεμημένη οργάνωση και επεξεργασία/επερώτηση των δεδομένων του υπερχώρου.

Η υποδράση ΥΔ4.3 αφορά μεθόδους εύρεσης ομοιότητας χρηστών υπερχώρων δεδομένων. Στα πλαίσια της δράσης αυτής εξετάσαμε κλιμακούμενες μεθόδους υπολογισμού της ομοιότητας σε υπερχώρους δεδομένων οι οποίες μπορούν να αξιοποιηθούν κατά την επεξεργασία σύνθετων δεδομένων ενός υπερχώρου. Συγκεκριμένα εξετάσαμε περιπτώσεις όπου η ομοιότητα ανάμεσα στους χρήστες ενός υπερχώρου αποτιμάται μέσω των αποτελεσμάτων συναθροιστικών-αναλυτικών ερωτημάτων στα δεδομένα τους. Επιπλέον εξετάσαμε περιπτώσεις όπου είναι επιθυμητός ο χωρο-χρονικός υπολογισμός της ομοιότητας μεταξύ των χρηστών, μέσω της παρακολούθησης των τροχιών κίνησης τους σε εφαρμογές κινητών αντικειμένων.

Για την πρώτη προσέγγιση, προτείναμε έναν καινοτόμο τρόπο εκτίμησης της ομοιότητας σε χρήστες υπερχώρων υψηλής διάστασης τα οποία προκύπτουν ως αποτέλεσμα συναθροίσεων κατά την ομαδοποίηση και ανάλυση των δεδομένων ενός υπερχώρου. Τέτοια παραδείγματα προκύπτουν κατά την ανάλυση δεδομένων σε αποθήκες δεδομένων. Χρησιμοποιήσαμε μια γνωστή τεχνική παραγωγής συνόψεων δεδομένων την οποία συνδυάσαμε με μια τεχνική κατακερματισμού δεδομένων έτσι ώστε να είναι δυνατή μια αρκετά ακριβής και σε μικρό χρόνο εκτίμηση της ομοιότητας σε μεγάλους υπερχώρους δεδομένων υψηλής διάστασης. Με τον τρόπο αυτό, δημιουργείται ένα σχήμα ευρετηρίασης των δεδομένων σε δύο επίπεδα και επιτυγχάνεται ο εντοπισμός ακραίων τιμών-

έκτοπων χρηστών (outliers). Η συγκεκριμένη διαδικασία βοηθά όλους τους χρήστες να αντιληφθούν πιθανές λανθασμένες εγγραφές δεδομένων στην αποθήκη δεδομένων και να επέμβουν διορθώνοντας και επαναφέροντας τις σωστές τιμές διατηρώντας την ακεραιότητα τους.

Σχετικά με τη δεύτερη προσέγγιση, μελετήσαμε δύο προβλήματα ομοιότητας χρηστών με χωρο-χρονικά κριτήρια. Το πρώτο πρόβλημα το οποίο εξετάσαμε αφορά τον υπολογισμό της ομοιότητας σε εφαρμογές κινούμενων αντικειμένων. Σε αυτή την περίπτωση η ομοιότητα των χρηστών προκύπτει μέσω του υπολογισμού της εγγύτητας των τροχιών τους στο χώρο. Στα πλαίσια της ΥΔ 4.3 εξετάσαμε διαφορετικές μετρικές ομοιότητας βασισμένες στην αρχή του εγγύτερου γείτονα, και προτείνουμε αντίστοιχες αποτελεσματικές μεθόδους. Στο δεύτερο πρόβλημα εξετάσαμε μεθόδους αποτίμησης της ομοιότητας χρηστών με βάση την απόσταση των τροχιών τους από μερικά δοσμένα σταθερά σημεία ενδιαφέροντος.

3 Αναλυτικά Αποτελέσματα

3.1 Υπολογισμός ομοιότητας και έκτοπων τιμών σε ομαδοποιημένα δεδομένα

Στα πλαίσια της υποδράσης ΥΔ 4.3 εξετάσαμε σύγχρονες τεχνικές οι οποίες μπορούν να χρησιμοποιηθούν για τον υπολογισμό της ομοιότητας ανάμεσα σε μεγάλα σύνολα δεδομένων υψηλής διάστασης. Χρησιμοποιήσαμε ένα από τα αποτελέσματα της υποδράσης ΥΔ 4.2, συγκεκριμένα την ανάπτυξη συνόψεων με τη χρήση κατακερματισμού ευαίσθητου στην ομοιότητα (Locality Sensitive Hashing - LSH). Στην συγκεκριμένη υποδράση δείξαμε ότι τέτοιες συνόψεις μπορούν να υπολογιστούν αποδοτικά σε υπερχώρους δεδομένων χρησιμοποιώντας σύγχρονες τεχνικές κατανεμημένης επεξεργασίας όπως οι διεργασίες MapReduce (βλ παραδοτέο Π4.2). Επιπρόσθετα, επιτρέπουν την αποτίμηση διαφορετικών μετρικών υπολογισμού της ομοιότητας (Ευκλείδεια, Jaccard, κα).

Σε αυτή την υποδράση μελετήσαμε πώς οι συγκεκριμένες τεχνικές μπορούν να επεκταθούν για την αποτίμηση της ομοιότητας ή, συμμετρικά, για τον υπολογισμό έκτοπων τιμών, σε δεδομένα υψηλής διάστασης τα οποία δημιουργούνται κατά την ανάλυση πρωτογενών δεδομένων σε υπερχώρους μέσω αναλυτικών επερωτήσεων. Παράδειγμα τέτοιων επερωτήσεων είναι ο τελεστής του κύβου σε αποθήκες δεδομένων.

Ο υπολογισμός έκτοπων τιμών (outliers) στοχεύει στην εύρεση δεδομένων τα οποία διαφέρουν σημαντικά από τα υπόλοιπα δεδομένα του υπερχώρου. Αυτό μπορεί να οφείλεται σε σφάλματα απεικόνισης και επομένως η πληροφόρηση αυτή μπορεί να χρησιμοποιηθεί για την παρουσίαση ποιο αξιόπιστων αποτελεσμάτων στους χρήστες του υπερχώρου.

Στη μελέτη μας σχεδιάσαμε ένα καινοτόμο σύστημα το οποίο επιτρέπει τον εντοπισμό ακραίων τιμών-έκτοπων (outliers) σε αποθήκες δεδομένων. Προτείναμε έναν νέο ορισμό ακραίων τιμών για αποτελέσματα συνάθροισης για να είναι δυνατή η διαχείριση μεγάλων υπερχώρων δεδομένων που είναι αρκετά συνηθισμένοι σε αποθήκες δεδομένων.

Οι τεχνικές μας εκμεταλλεύονται ένα σχήμα ευρετηρίασης δεδομένων σε δύο επίπεδα. Το πρώτο επίπεδο βασίζεται στην τεχνική κατακερματισμού LSH και μας επιτρέπει να αντικαταστήσουμε τις ερωτήσεις εύρους (range queries) που δεν είναι αποδοτικές σε δεδομένα με πολλές διαστάσεις (dimensionality curse) με προσεγγιστικούς υπολογισμούς των κοντινότερων γειτόνων. Το δεύτερο επίπεδο ευρετηρίασης χρησιμοποιεί την τεχνική PAA (Piecewise Aggregate Appoximation), η οποία μειώνει σημαντικά τις ανάγκες σε χώρο για την αποθήκευση των συνόψεων των δεδομένων.

Η μέθοδός μας επιτρέπει τις σταδιακές ενημερώσεις των αναπαραστάσεων-συνόψεων οι οποίες είναι απαραίτητες σε μεγάλου όγκου σύνολα δεδομένων που είναι συνηθισμένα σε αποθήκες δεδομένων. Τα αποτελέσματά μας έχουν δημοσιευτεί στο άρθρο [GeKo12].

3.2 Υπολογισμός ομοιότητας (εγγύτητας) κινούμενων αντικειμένων

3.2.1 Ομοιότητα σε σχέση με κινούμενο αντικείμενο

Η ανάλυση δεδομένων παρακολούθησης για διάφορους τύπους κινούμενων αντικειμένων είναι ένα ενδιαφέρον ερευνητικό πρόβλημα με πολλές πραγματικές εφαρμογές. Αρκετές εργασίες έχουν επικεντρωθεί στην συνεχή παρακολούθηση των πλησιέστερων γειτόνων ενός κινούμενου αντικειμένου, ενώ άλλες έχουν προτείνει μέτρα ομοιότητας για την εύρεση παρόμοιων τροχιών σε βάσεις δεδομένων που περιέχουν ιστορικά δεδομένα παρακολούθησης.

Σε αυτό το έργο, εισάγουμε το πρόβλημα της συνεχούς παρακολούθησης πλησιέστερων τροχιών. Σε αντίθεση με άλλες παρόμοιες προσεγγίσεις, ενδιαφερόμαστε για την παρακολούθηση κινούμενων αντικειμένων, λαμβάνοντας υπόψη σε κάθε χρονόσημο όχι μόνο τις τρέχουσες θέσεις τους, αλλά και την πρόσφατη τροχιά τους σε ένα καθορισμένο παράθυρο χρόνου. Περιγράφουμε πρώτα ένα γενικό βασικό αλγόριθμο για το πρόβλημα αυτό, ο οποίος μπορεί να εφαρμοστεί για κάθε συναθροιστική συνάρτηση που χρησιμοποιείται για τον υπολογισμό αποστάσεων τροχιάς μεταξύ των αντικειμένων, και χωρίς περιορισμούς στην κίνηση των αντικειμένων. Χρησιμοποιώντας αυτό ως πλαίσιο, στη συνέχεια προτείνουμε ένα βελτιστοποιημένο αλγόριθμο για τις περιπτώσεις όπου η απόσταση μεταξύ δύο κινούμενων αντικειμένων σε ένα χρονικό παράθυρο ορίζεται από τη μέγιστη ή την ελάχιστη απόσταση τους σε όλα τα χρονόσημα. Στη συνέχεια προτείνουμε επιπλέον βελτιστοποιήσεις για την περίπτωση που υπάρχει ένα άνω φράγμα για τις ταχύτητες των αντικειμένων. Τέλος, αξιολογούμε την αποτελεσματικότητα των προτεινόμενων αλγορίθμων μας, με τη διεξαγωγή πειραμάτων σε τρεις πραγματικές βάσεις δεδομένων. Τα αποτελέσματά μας έχουν δημοσιευτεί στο άρθρο [SaSS14].

3.2.2 Ομοιότητα σε σχέση με σταθερά σημεία αναφοράς

Τα δεδομένα τροχιών αποτυπώνουν το ιστορικό κινούμενων αντικειμένων, όπως άτομα ή οχήματα. Με τη διάδοση του GPS και της τεχνολογίας εντοπισμού,

τεράστιοι όγκοι τροχιών παράγονται και συλλέγονται. Σε αυτό το πλαίσιο, εφαρμογές όπως η πρόταση διαδρομής και η εξόρυξη της συμπεριφορά ταξιδιού, απαιτούν την αποτελεσματική ανάκτηση τροχιών. Επομένως σε αυτή την εργασία, η ομοιότητα χρηστών του οικοσυστήματος ορίζεται με βάση την απόσταστη της τροχιάς τους από μερικά δοσμένα σταθερά σημεία.

Στην συγκεκριμένη εργασία, οι χρήστες αναπαρίστανται από την τροχιά τους στο χώρο, η μετρική ομοιότητας είναι ως προς ένα δοσμένο σύνολο από σημεία στο χώρο. Αναζητούμε επομένως τους χρήστες που έχουν την μεγαλύτερη ομοιότητα με το δοσμένο σύνολο, και εξετάζουμε διάφορες τεχνικές δεικτοδότησης και υλοποίησης της αναζήτησης. Συγκεκριμένα, δεδομένων μια συλλογής τροχιών και ένα σύνολο σημείων επερώτησης, ο στόχος είναι να ανακτήσουμε τις k τροχιές που περνούν όσο το δυνατόν πλησιέστερα σε όλα τα σημεία του ερωτήματος. Εξελίσσουμε την τρέχουσα τεχνολογική κατάσταση, συνδυάζοντας τις υπάρχουσες προσεγγίσεις σε μια υβριδική μέθοδο. Επιπλέον προτείνουμε μια εναλλακτική, πιο αποτελεσματική προσέγγιση βασισμένη σε ερωτήσεις εύρους. Κατόπιν, προτείνουμε και μελετάμε μια πρακτική παραλλαγή της αναζήτησης με βάση την απόσταση που εισάγει και ένα όριο στην απόσταση, ώστε να λαμβάνει υπόψη τα χρονικά χαρακτηριστικά των τροχιών. Μέσα από μια εκτεταμένη πειραματική ανάλυση με πραγματικά δεδομένα τροχιών, δείχνουμε ότι η προσέγγιση που βασίζεται σε ερωτήσεις εύρους έχει καλύτερα επίδοση από τις προηγούμενες μεθόδους κατά τουλάχιστον μία τάξη μεγέθους. Τα αποτελέσματά μας έχουν δημοσιευτεί στο άρθρο [ShBSM15] το οποίο έλαβε βράβευση καλύτερης εργασίας (Best Paper Award).

4 Ανακεφαλαίωση

Το παρόν παραδοτέο Π4.3 παρουσιάζει τα αποτελέσματα της υποδράσης ΥΔ4.3 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ4.3 ήταν η ανάπτυξη τεχνικών αποτίμησης της ομοιότητας χρηστών σε υπερχώρους δεδομένων. Εξετάσαμε δύο διαφορετικές προσεγγίσεις υπολογισμού της ομοιότητας ανάμεσα στους χρήστες.

Η πρώτη χρησιμοποιεί συναρθοιστικά αποτελέσματα αναλυτικών ερωτημάτων και επιτρέπει τον υπολογισμό της ομοιότητας ανάμεσα σε χρήστες μέσω της γενικότερης συμπεριφοράς τους (όπως αυτή αποτυπώνεται μέσω της επιλογής των κατάλληλων συναθροιστικών ερωτημάτων). Οι προτεινόμενες τεχνικές αξιοποιούν αποτελέσματα της ΥΔ 4.2 και επιτρέπουν τον υπολογισμό της ομοιότητας μέσω σύγχρονων τεχνικών κατανεμημένης επεξεργασίας όπως το MapReduce.

Η δεύτερη προσέγγιση αφορά τον υπολογισμό της ομοιότητας ανάμεσα σε χρήστες υπερχώρων λαμβάνοντας υπόψη χωρο-χρονική πληροφορία σε εφαρμογές κινούμενων αντικειμένων. Σε αυτή την περίπτωση η ομοιότητα προκύπτει μέσω του υπολογισμού της εγγύτητας των τροχιών των αντικειμένων στο χώρο είτε ως προς μια δοσμένη τροχιά είτε ως προς ένα δοσμένο σύνολο σταθερών σημείων. Στα πλαίσια της ΥΔ 4.3 εξετάσαμε διαφορετικές μετρικές ομοιότητας βασισμένες στην αρχή του εγγύτερου γείτονα, και προτείνουμε αντίστοιχες αποτελεσματικές μεθόδους υπολογισμού.

Στα πλαίσια των ερευνητικών δραστηριοτήτων της υποδράσης Υ.Δ 4.3 προέκυψαν 3 δημοσιεύσεις (βλ. παρακάτω πίνακα). Σε αυτές παρουσιάζονται αναλυτικά οι αλγόριθμοι και τεχνικές και μελετάται πειραματικά η απόδοση τους.

Δημοσιεύσεις

[GeKo12] Konstantinos Georgoulas, Yannis Kotidis. Towards Enabling Outlier Detection in Large, High Dimensional Data Warehouses. SSDBM 2012: 591-594

[SaSS14] Dimitris Sacharidis, Dimitrios Skoutas, Georgios Skoumas. Continuous monitoring of nearest trajectories. SIGSPATIAL/GIS 2014: 361-370

[ShBSM15] Shuyao Qi, Panagiotis Bouros, Dimitris Sacharidis, Nikos Mamoulis. Efficient Point-based Trajectory Search. SSTD 2015

Παράρτημα