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.1.3

Μηχανισμοί δεικτοδότησης 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	**Αναπαράσταση και ανάκτηση μη παραδοσιακών δεδομένων**				
Ομάδα	Ερ. Ομάδα 1	**Έναρξη**	01/02/2012	**Λήξη**	30/11/2015

Συντονιστής ΕΟ1	Δ. Πλεξουσάκης (Παν. Κρήτης)	
Υποδράση: ΥΔ 1.3	Μηχανισμοί δεικτοδότησης μη-παραδοσιακών δεδομένων	
Συμμετέχοντες	*Μέλη ΚΕΟ*	Δ. Πλεξουσάκης (Παν. Κρήτης), Γ. Θεοδωρίδης (Παν. Πειραιώς), Σ. Σκιαδόπουλος (Παν. Πελοποννήσου)
	Μέλη ΟΕΣ	Ν. Πελέκης (Παν. Πειραιώς)
Σύντομη Περιγραφή	Η Δράση 1 αποσκοπεί στον ορισμό μοντέλων αναπαράστασης για μη παραδοσιακούς τύπους δεδομένων που είναι πλούσια σε μεταδεδομένα, καθώς και στο σχεδιασμό μηχανισμών για την αυτόματη εξαγωγή της εν λόγω μετα-πληροφορίας για νέα δεδομένα. Στο πλαίσιο αυτό η Υποδράση ΥΔ1.3 ερευνά το σχεδιασμό νέων τεχνικών ευρετηρίασης δεδομένων, που θα επιτρέψουν την αποδοτική ανάκτηση των δεδομένων με βάση την προαναφερθείσα μετα-πληροφορία.	
Παραδοτέο	Π.1.3 Μηχανισμοί δεικτοδότησης μη-παραδοσιακών δεδομένων	
Στόχος στο Τ.Δ.	Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 2 δημοσιεύσεις.	
Επίτευξη στόχου	100%	

Περιεχόμενα

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 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 1 «Αναπαράσταση και ανάκτηση μη παραδοσιακών δεδομένων» αποσκοπεί στον ορισμό μοντέλων αναπαράστασης για μη παραδοσιακούς τύπους δεδομένων που είναι πλούσια σε μεταδεδομένα, καθώς και στο σχεδιασμό μηχανισμών για την αυτόματη εξαγωγή της εν λόγω μετα-πληροφορίας για νέα δεδομένα. Τέλος αποσκοπεί στο σχεδιασμό νέων τεχνικών ευρετηρίασης των δεδομένων, που θα επιτρέψουν την αποδοτική ανάκτηση των δεδομένων με βάση την προαναφερθείσα μετα-πληροφορία. Η Δράση οργανώνεται σε τρεις θεμελιώδεις δράσεις, εκ των οποίων η πρώτη αφορά στη μοντελοποίηση μη-παραδοσιακών δεδομένων, η δεύτερη σε τεχνικές ανάκτησης πληροφορίας από πηγές μη-παραδοσιακών δεδομένων και η τρίτη σε μηχανισμούς δεικτοδότησης μη-παραδοσιακών δεδομένων.

Το παρόν Παραδοτέο Π.1.3 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ1.3. Στην ενότητα 1 παρουσιάζουμε το γενικότερο πλαίσιο της έρευνας που διεξήχθηκε η οποία ακολούθησε δύο οδούς, η πρώτη από τις οποίες αφορά σε δεικτοδότηση μεγάλων συμβολοσειρών και η δεύτερη αυτή της διαχείρισης σημασιολογικά εμπλουτισμένων βάσεων κινούμενων αντικειμένων. Στις ενότητες 2 και 3 παρουσιάζουμε τα κύρια αποτελέσματα για αυτές τις δύο διαφορετικές προσεγγίσεις.

1.1 Πλαίσιο έρευνας

Ο στόχος αυτής της εργασίας είναι να προτείνει νέους μηχανισμούς ευρετηρίασης για μη παραδοσιακούς τύπους δεδομένων. Οι μηχανισμοί ευρετηρίασης θα πρέπει να μπορούν στηρίξουν την ανάλυση χώρων δεδομένων (dataspaces), δηλαδή την επέκταση των υφιστάμενων τεχνικών ευρετηρίασης για να χειριστεί τις προκλήσεις που τίθενται από το νέο μοντέλο. Το μοντέλο dataspace μπορεί να χρησιμοποιεί τα υπάρχοντα σχήματα δεικτοδότησης, αλλά σε γενικές γραμμές οι νέες δομές δεικτοδότησης πρέπει να μπορούν να εφαρμοστούν για την αντιμετώπιση πολύπλοκων ερωτημάτων σε μη παραδοσιακούς τύπους δεδομένων. Τέλος, αποτελεί στόχο να ληφθεί υπόψιν η παρουσία μεγάλου όγκου πλούσιων μεταδεδομένων η οποία απαιτεί τη δημιουργία εξειδικευμένων δομών ευρετηρίασης που υποστηρίζουν ερωτήματα που περιλαμβάνουν δεδομένα και μεταδεδομένα για όλες τις πηγές που διατίθενται στο dataspace.

Για την επίτευξη των παραπάνω στόχων η έρευνα στα πλαίσια του Π1.3 στόχευσε δύο πολύ συνήθεις μη παραδοσιακούς τύπους δεδομένων, αυτό των συμβολοσειρών και αυτό των σημασιολογικά εμπλουτισμένων δεδομένων κίνησης. Η πρώτη προσέγγιση αυτή μπορεί να μοντελοποιήσει πλήθος μη-παραδοσιακών δεδομένων όπως βιοιατρικά, χρονικά και συμπιεσμένα δεδομένα αλλά και δεδομένα κειμένου, τα οποία είναι συνήθεις πηγές δεδομένων σε χώρους δεδομένων. Ομοίως, η δεύτερη προσέγγιση μπορεί να εφαρμοστεί σε οποιαδήποτε εφαρμογή εμπλέκει χωρο-χρονικο-κειμενικά ακολουθιακά δεδομένα, όπως είναι τα δεδομένα που υπάρχουν σε κοινωνικά δίκτυα βάσει θέσης. Οι επόμενες δύο ενότητες παρουσιάζουν αναλυτικότερα τις δύο προαναφερθείσες προσεγγίσεις.

2 Δεικτοδότηση μεγάλων συμβολοσειρών

Το suffix tree είναι μια δομή για την δεικτοδότηση συμβολοσειρών που χρησιμοποιείται σε πληθώρα εφαρμογών όπως η βιοιατρική, η ανάλυση χρονοσειρών, η ομαδοποίηση, η επεξεργασία κειμένου και η συμπίεση. Όμως

όταν η συμβολοσειρά και το suffix tree της είναι πολύ μεγάλα για να αποθηκευθούν στη μνήμη οι περισσότεροι αλγόριθμοι κατασκευής του suffix tree είναι μη αποδοτικοί. Στην έρευνα αυτή σχαδιάσαμε ένα αλγόριθμο κατασκευής με βάση το δίσκο που ονομάζεται ERA και μπορεί να διαχειριστεί αποδοτικά πολύ μεγαλές συμβολοσειρές. Ο ERA διαχωρίζει τη διαδικασία κατασκευής οριζόντια και κατακόρυφα και ελαχιστοποιεί το I/O προσαρμόζοντας δυναμικά τα οριζόντια τμήματα ανεξάρτητα από τα κατακόρυφα βασιζόμενος στο εξελισσόμενο σχήμα του δέντρου και την διαθέσιμη μνήμη. Όπου είναι δυνατό, ο ERA ομαδοποιεί τα κατακόρυφα τμήματα για να μειώσει τα I/O. Υλοποιήσαμε μια σειριακή έκδοση, μια παράλληλη (σε αρχιτεκτονική shared-memory/shared-disk) και μια παράλληλη (σε αρχιτεκτονική shared nothing). Η μέθοδός μας δεικτοδοτεί σε ένα κοινό οικιακό υπολογιστή με 1CPU μια συμβολοσειρά μεγέθους 3GB σε 19 λεπτά. Για σύγκριση, η προηγούμενη ταχύτερη μέθοδος χρειαζόταν 15 λεπτά στο υπερυπολογιστή IBM BlueGene που διαθέτει 1024CPUs. Η προσέγγιση αυτή μπορεί να μοντελοποιήσει πλήθος μη-παραδοσιακών δεδομένων όπως βιοιατρικά, χρονικά και συμπιεσμένα δεδομένα αλλά και δεδομένα κειμένου.

Η παραπάνω έρευνα εντάσσεται στη Υπο-Δράση Π1.3 καθώς αφορά στην κατασκευή μηχανισμών δεικτοδότησης δεδομένων με αλφαριθμητική αναπαράσταση (string representation).

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [MASK11] που παρουσιάστηκε στο VLDB, το 2011.

3 Διαχείριση και ανάλυση σημασιολογικά εμπλουτισμένων βάσεων κινούμενων αντικειμένων

Η έκρηξη διαθέσιμης πληροφορίας θέσης, συνδεδεμένη με επαγόμενη ή δηλούμενη (από τους χρήστες) σημασιολογία των αντίστοιχων τοποθεσιών, ήδη συνεισφέρει στην εποχή των μεγάλων δεδομένων, θέτοντας νέες προκλήσεις στην ερευνητική κοινότητα της διαχείρισης και εξόρυξης γνώσης από δεδομένα

κίνησης. Σε αυτή την εργασία, εμπνεόμενοι από πρόσφατες εξελίξεις στο πεδίο, παρουσιάζουμε ένα ενοποιημένο πλαίσιο εργασίας για τη διαχείριση και ανάλυση των «βημάτων» (LifeSteps) μας, δηλαδή, αντικειμένων δεδομένων τα οποία αναπαριστούν ταυτόχρονα τα πρωτογενή δεδομένα τροχιών κινούμενων αντικειμένων και την αντίστοιχη σημασιολογική πληροφορία τους. Ειδικότερα, παρέχουμε λύσεις για την ανάπτυξη πραγματικών, σημασιολογικά εμπλουτισμένων βάσεων κινούμενων αντικειμένων και συστημάτων αποθηκών δεδομένων, ενώ σχεδιάζουμε αντίστοιχους αλγορίθμους διαχείρισης επερωτήσεων. Η πειραματική μας μελέτη σε συνθετικά δεδομένα που διατηρούν συγχρονισμένα τα πρωτογενή δεδομένα (GPS log) με τη σημασιολογική πληροφορία (ημερολόγια κίνησης) επαληθεύει την αποτελεσματικότητα και αποδοτικότητα του προτεινόμενου πλαισίου εργασίας.

Η έρευνα αυτή αποτελεί μία καινοτόμο πρόταση στη διαχείριση χωρο-χρονο-κειμενικών ακολουθιακών δεδομένων, όπως μπορούν να θεωρηθούν ότι είναι οι σημασιολογικά εμπλουτισμένες τροχιές κινούμενων αντικειμένων. Προτείνεται ένα γενικού σκοπού πλαισίου εργασίας για τέτοιου τύπου μη-παραδοσιακά δεδομένα, το οποίο υποβοηθά την ανάπτυξη σημασιολογικά εμπλουτισμένων βάσεων κινούμενων αντικειμένων και συστημάτων αποθηκών δεδομένων. Αυτό επιτυγχάνεται με ειδικά σχεδιασμένους αλγορίθμους διαχείρισης επερωτήσεων (όπως χωρο-κειμενικά ερωτήματα εύρους ή ερωτήματα προτύπων) και διαδικασίες εξαγωγής-μεταμόρφωσης-φόρτωσης των πρωτογενών δεδομένων σε κατάλληλες αποθήκες δεδομένων, οι οποίες υποβοηθούνται από ειδικούς μηχανισμούς δεικτοδότησης για χωρο-χρονο-κειμενικά, ακολουθιακά δεδομένα.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [PST15] που παρουσιάστηκε στο IEEE/ACM International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015), το 2015.

4 Δημοσιεύσεις

[MASK11] E. Mansour, A. Allam, S. Skiadopoulos, P. Kalnis: ERA: Efficient Serial and Parallel Suffix Tree Construction for Very Long Strings. PVLDB 5(1): 49-60, Istanbul, Turkey, 2011.

[PST15] N. Pelekis, S. Sideridis, Y. Theodoridis: "Hermes^{sem}: a Semantic-aware Framework for the Management and Analysis of our LifeSteps", IEEE/ACM International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015), Paris, France, 2015.

Παράρτημα