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.4.2

Μέθοδοι για παραγωγή στατιστικών και συνόψεων 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 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 4.2	Μέθοδοι για παραγωγή στατιστικών και συνόψεων 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 ΥΔ4.2 επικεντρώνεται στην ανάπτυξη τεχνικών για την παραγωγή στατιστικών και συνόψεων δεδομένων χρησιμοποιώντας όπου αυτό είναι εφικτό σύγχρονες τεχνικές κατανεμημένης επεξεργασίας (πχ MapReduce). Η έρευνα μας εστίασε όχι μόνο σε ντετερμινιστικά δεδομένα αλλά αναπτύχθηκαν τεχνικές κατάλληλες και για πιθανοτικά δεδομένα. Οι τεχνικές που αναπτύχθηκαν έχουν κεντρικό ρόλο στην ανάπτυξη κλιμακώσιμων εφαρμογών καθώς επιτρέπουν την προσεγγιστική αποτίμηση σύνθετων ερωτημάτων αναζήτησης χρησιμοποιώντας σημαντικά λιγότερους υπολογιστικούς πόρους.				

Παραδοτέο	Π.4.2 Μέθοδοι για παραγωγή στατιστικών και συνόψεων δεδομένων
Στόχος στο Τ.Δ.	Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 2 δημοσιεύσεις.
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.4.2 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ 4.2. 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 οικοσυστήματος.

Η υποδράση ΥΔ 4.2 «Μέθοδοι για παραγωγή στατιστικών και συνόψεων δεδομένων» αποσκοπεί στην παραγωγή συνόψεων τόσο για ντετερμινιστικά όσο και για πιθανοτικά δεδομένα. Οι συνόψεις που δημιουργούνται είναι πολύ μικρότερες από τα πρωτογενή δεδομένα και επιτρέπουν την αποτίμηση σύνθετων ερωτημάτων αναζήτησης χρησιμοποιώντας σημαντικά λιγότερους υπολογιστικούς πόρους. Στα πλαίσια της εκπόνησης του έργου μελετήθηκαν τα παρακάτω είδη συνόψεων:

1. Ιστογράμματα για ντετερμινιστικά δεδομένα καθώς και για δεδομένα με αβεβαιότητα: Τα ιστογράμματα αποτελούν τον πιο ευρέως χρησιμοποιούμενο και αποτελεσματικό μηχανισμό σύνοψης δεδομένων. Μελετώντας τους αλγόριθμους κατασκευής ιστογραμμάτων για ντετερμινιστικά δεδομένα, στο πλαίσιο αυτής της εργασίας έγινε δυνατή η προσαρμογή τους για πιθανοτικά δεδομένα, που είναι και το κύριο ζητούμενο της υποδράσης ΥΔ 4.2.
2. Δημιουργία συνόψεων LSH για ερωτήματα ομοιότητας: Ο υπολογισμός της ομοιότητας ανάμεσα σε δεδομένα (πχ αντικείμενα, κείμενα) είναι κεντρικός στη υλοποίηση αποδοτικών αλγορίθμων αναζήτησης σε σύγχρονες εφαρμογές. Στα πλαίσια της δράσης εξετάστηκαν τεχνικές δημιουργίας συνόψεων με τη μορφή πινάκων κατακερματισμού οι οποίοι επιτρέπουν τη γρήγορη αποτίμηση της ομοιότητας (πχ σε ερωτήματα κοντινότερου γείτονα).
3. Δημιουργία συνόψεων κορυφογραμμών: τα ερωτήματα κορυφογραμμής επιτρέπουν τη σύνοψη μεγάλων δεδομένων μέσω του υπολογισμού των πλέον

ενδιαφερουσών από αυτά χρησιμοποιώντας ένα σύνολο ρητά προσδιορισμένων προτιμήσεων στις τιμές των γνωρισμάτων τους. Με αυτό τον τρόπο γίνεται δυνατή η παρουσίαση προσωποποιημένης πληροφόρησης σε συστήματα διαχείρισης δεδομένων μεγάλης κλίμακας.
4. Δημιουργία συνόψεων με τη μέθοδο της διαφοροποίησης: τέτοιες συνόψεις επιτρέπουν τη διαφοροποίηση (diversification) των αποτελεσμάτων αναζήτησης σε ερωτήματα χρηστών.

Επιπλέον των ανωτέρω μελετήθηκε ο υπολογισμός κατάλληλων στατιστικών τα οποία πέρα από τα δεδομένα θα επιτρέπουν την αξιολόγηση των χρηστών του συστήματος και της εκτίμηση του ρόλου τους σε συσχέτιση με τους υπόλοιπους χρήστες του οικοσυστήματος. Ειδικότερα, εστιάσαμε στη αξιοποίηση γεωαναφορών για τον υπολογισμό της κεντρικότητας χρηστών Κοινωνικών Δικτύων με Επίγνωση Θέσης (ΚΔΕΘ), όπου οι χρήστες συνάπτουν κοινωνικές σχέσεις και κάνουν δημοσιεύσεις με γεωαναφορές.

3 Αναλυτικά Αποτελέσματα

3.1 Δημιουργία ιστογραμμάτων για ντετερμινιστικά και πιθανοτικά δεδομένα

Τα ιστογράμματα είναι ένας αποδεδειγμένα πολύ αποτελεσματικός μηχανισμός περίληψης, και χρησιμοποιούνται ευρέως για να προσεγγίσουν κατανομές δεδομένων. Επιπρόσθετα, αποτελούν ένα σημαντικό εργαλείο σε εμπορικές μηχανές επερωτήσεων.

Ας θεωρήσουμε την κατανομή της συχνότητας εμφάνισης πλειάδων σε ένα μεγάλο σύνολο δεδομένων. Παραδοσιακά, η προσέγγιση μιας τέτοιας κατανομής με ένα ιστόγραμμα επιτυγχάνεται με το διαχωρισμό του πεδίου τιμών των δεδομένων σε ένα μικρό αριθμό από διαδοχικές περιοχές τιμών (τους κάδους), και αποθηκεύοντας μόνο συνοπτικά στατιστικά στοιχεία που περιγράφουν τις συχνότητες εμφάνισης των πλειάδων σε έναν κάδο. Τα όρια ενός κάδου επιλέγονται ώστε να ελαχιστοποιείται μια δεδομένη μετρική σφάλματος που μετρά τις ανομοιομορφίες εντός του κάδου, δηλαδή τις διαφορές των

συνοπτικών στατιστικών που έχουν επιλεγεί για τον κάδο (π.χ., η μέση τιμή των συχνοτήτων των πλειάδων μέσα στον κάδο) από τις πραγματικές συχνότητες των πλειάδων, και συναθροιστικά σφάλματα σε όλους τους κάδους (ως άθροισμα ή κατά μέγιστη τιμή). Ο γρήγορος υπολογισμός τέτοιων περιλήψεων είναι μια κρίσιμη παράμετρος σε συστήματα διαχείρισης μεγάλων δεδομένων όσον αφορά την ταχεία λήψη προσεγγιστικών απαντήσεων.

Δεδομένου ότι τα ιστογράμματα είναι από τη φύση τους προσεγγιστικά, ένας επιπλέον παράγοντας προσέγγισης που εισάγεται για επιτάχυνση της διαδικασία κατασκευής τους, μπορεί να γίνει ανεκτός. Στα πλαίσια της υποδράσης ΥΔ 4.2 εκπονήθηκε μία (1+ε)-προσεγγιστική μέθοδος για την κατασκευή ιστογραμμάτων για ντετερμινιστικά δεδομένα.

Ένα άλλο θέμα που εξετάσαμε, είναι η ανάπτυξη συνόψεων με τη μορφή ιστογραμμάτων σε συστήματα που διαχειρίζονται δεδομένα με αβεβαιότητα. Συγκεκριμένα εξετάστηκε η ενσωμάτωση μεθόδων για την κατασκευή πιθανοτικών ιστογραμμάτων πάνω σε πιθανοτικά δεδομένα. Συγκεκριμένα αναπτύχθηκαν αλγόριθμοι Δυναμικού Προγραμματισμού (ΔΠ) δύο φάσεων για την κατασκευή βέλτιστων πιθανοτικών ιστογραμμάτων για μια σειρά από μετρικές σφάλματος.

Στη συνέχεια, προκειμένου να μειωθεί το υψηλό υπολογιστικό κόστος της βέλτιστης κατασκευής πιθανοτικών ιστογραμμάτων, σχεδιάστηκε μία προσεγγιστική μέθοδος στο γενικό πλαίσιο του ΔΠ. Η επιτάχυνση που επιτυγχάνεται οφείλεται στο ότι ο προσεγγιστικός αλγόριθμος λαμβάνει υπόψη μόνο ένα μικρό σύνολο των υπο-προβλημάτων του ΔΠ. Ο προτεινόμενος αλγόριθμος παρέχει εγγυήσεις στο συνολικό σφάλμα του ιστογράμματος και προσφέρει στο χρήστη μια μέθοδο αντιστάθμισης της ακρίβειας του ιστογράμματος και του υπολογιστικού κόστους.

Η πειραματική μελέτη έδειξε ότι η μέθοδος παράγει σχεδόν βέλτιστα πιθανοτικά ιστογράμματα, με πολύ χαμηλότερο υπολογιστικό κόστος από το βέλτιστο αλγόριθμο. Τα αποτελέσματά μας παρουσιάζονται στην εργασία [Ma13].

3.2 Δημιουργία συνόψεων που διατηρούν την ομοιότητα

Στη κοινωνία της πληροφορίας οι πήγες παραγωγής δεδομένων αυξάνονται συνεχώς όπως και η ανάγκη για αποτελεσματική διαχείριση τους. Η ανάγκη για αναζήτηση σε μεγάλα, και συχνά ετερογενή, σύνολα πληροφοριών, έχει οδηγήσει στη χρήση προσεγγιστικών τεχνικών, προκειμένου να μπορούν να απαντηθούν αποδοτικά ερωτήματα ομοιότητας, όπως αυτά των κοντινότερων γειτόνων. Η εμφάνιση του προγραμματιστικού μοντέλου MapReduce καθώς και του αντίστοιχου συστήματος αρχείων HDFS, καθιστούν εφικτή την παράλληλη και κατανεμημένη επεξεργασία μεγάλου όγκου δεδομένων, με χρήση απλών υπολογιστικών κόμβων.

Στα πλαίσια της υποδράσης ΥΔ 4.2 εξετάσαμε το πρόβλημα του υπολογισμού ερωτημάτων κοντινότερων γειτόνων με χρήση προσεγγιστικών τεχνικών και του προγραμματιστικού μοντέλου MapReduce, για σύνολα δεδομένων σε χώρους υψηλής διάστασης. Μελετήθηκαν διαφορετικά μέτρα αποτίμησης της ομοιότητας πολυδιάστατων αντικειμένων όπως η Ευκλείδεια απόσταση, ο Jaccard Index, το μέτρο συνημίτονου και ο correlation coefficient.

Καθώς ο όγκος και η διάσταση των δεδομένων σε σύγχρονες εφαρμογές κλιμακώνονται, συμβατικές τεχνικές οι οποίες χρησιμοποιούν ντετερμινιστικούς αλγορίθμους για τον υπολογισμό της ομοιότητας μεταξύ των δεδομένων και συνεπακολούθως των κοντινότερων τους γειτόνων γίνονται υπολογιστικά ασύμφορες. Ως εναλλακτική εξετάσαμε τη χρήση προσεγγιστικών τεχνικών οι οποίες μπορούν να υπολογίσουν τους κοντινότερους γείτονες οποιουδήποτε επιλεγμένου αντικειμένου με ελεγχόμενη αξιοπιστία. Συγκεκριμένα, εξετάσαμε τη χρήση του κατακερματισμού ευαίσθητου στην ομοιότητα (Locality Sensitive Hashing - LSH). Το LSH είναι μία τεχνική η οποία δημιουργεί μία σύνοψη (ή ευρετήριο) ενός μεγάλου συνόλου δεδομένων με την μορφή ενός πίνακα κατακερματισμού. Ο πίνακας αυτός έχει την ιδιότητα ότι η πιθανότητα δύο αντικείμενα να βρίσκονται στον ίδιο κάδο του είναι ανάλογη της πραγματικής τους ομοιότητας.

Με τη χρήση καταλλήλων συναρτήσεων κατακερματισμού, ανάλογα με το επιθυμητό τρόπο υπολογισμού της ομοιότητας, μπορούμε να αναζητήσουμε τους κοντινότερους γείτονες ενός αντικειμένου εστιάζοντας σε ένα υποσύνολο του πίνακα κατακερματισμού. Δυστυχώς, για μεγάλα δεδομένα η δημιουργία και αναζήτηση στους κάδους του LSH ευρετηρίου είναι απαγορευτικά ακριβή. Στα πλαίσια του ΕΙΚΟΣ εξετάσαμε τη χρήση του προγραμματιστικού μοντέλου MapReduce και του συστήματος αρχείων HDFS, για παράλληλη και κατανεμημένη επεξεργασία μεγάλου όγκου δεδομένων. Τόσο η δημιουργία της LSH σύνοψης όσο και η αναζήτηση σε αυτήν αναπτύχθηκαν μέσω διεργασιών MapReduce επιτρέποντας έτσι την κλιμάκωση της εφαρμογής σε μεγάλα δεδομένα.

Το κύριο αποτέλεσμα αυτής της εργασίας, το οποίο είναι η παράλληλη και κατανεμημένη υλοποίηση της παρασκευής συνόψεων βασισμένα σε LSH, κάνοντας χρήση του συστήματος απεικόνισης-μείωσης (MapReduce) Apache Hadoop, συνεισφέρει τα μέγιστα στο αντικείμενο μελέτης της υποδράσης ΥΔ 4.2. Τα αποτελέσματά μας παρουσιάζονται στην εργασία [Σπ13].

3.3 Δημιουργία συνόψεων μέσω κορυφογραμμών

Οι τεχνικές υπολογισμού κορυφογραμμών επιτρέπουν τη σύνοψη μεγάλων δεδομένων μέσω του υπολογισμού των πλέον ενδιαφερουσών εγγραφών χρησιμοποιώντας ένα σύνολο ρητά προσδιορισμένων προτιμήσεων στις τιμές των γνωρισμάτων τους. Συχνά το σύνολο των εγγραφών που ανήκουν στη κορυφογραμμή είναι πολύ μικρότερο από τα αρχικά δεδομένα, επιτρέποντας έτσι την χρησιμοποίηση τους ως μία "σύνοψη" των πιο ενδιαφερόντων τμημάτων ενός μεγάλου συνόλου δεδομένων.

Στα πλαίσια της υποδράσης ΥΔ 4.2 μελετήθηκε λεπτομερώς μια σημαντική κατηγορία αλγορίθμων κορυφογραμμής, οι αλγόριθμοι κορυφογραμμής βασισμένων στην σάρωση (scan-based skyline algorithms). Η βασική ιδέα των αλγόριθμων αυτών περιγράφεται ως εξής: Ένας αλγόριθμος αυτής της κατηγορίας εκτελεί πολλαπλά περάσματα (passes) πάνω από ένα αρχείο

εισόδου. Στο πρώτο πέρασμα, το αρχείο εισόδου περιέχει ολόκληρη την βάση δεδομένων. Στα επόμενα περάσματα, το αρχείο εσόδου αποτελεί το αρχείο που έχει προκύψει ως έξοδος από το προηγούμενο πέρασμα. Ο αλγόριθμος τερματίζει όταν το αρχείο εξόδου που παρήγαγε στο τέλος ενός περάσματος είναι άδειο. Κατά τη διάρκεια κάθε περάσματος, ο αλγόριθμος διατηρεί στην κύρια μνήμη (main memory) ένα παράθυρο (window) από μη-συγκρίσιμα (incomparable) αντικείμενα, τα οποία χρησιμοποιεί για να απομακρύνει από το αρχείο εισόδου αντικείμενα τα οποία κυριαρχούνται (dominated). Τέλος, κάθε αντικείμενο που δεν κυριαρχείται, εγγράφεται στο αρχείο εξόδου.

Με βάση το μοντέλο δευτερεύουσας μνήμης (external memory), προσαρμόστηκαν τέσσερις δημοφιλείς αλγόριθμοι υπολογισμού κορυφογραμμής βασισμένοι στην σάρωση, και εξετάστηκαν λεπτομερώς διάφοροι παράμετροι υλοποίησης καθώς και θέματα διαχείρισης μνήμης. Επικεντρωθήκαμε στην μελέτη ενός κύριου χαρακτηριστικού των αλγορίθμων αυτής της κατηγόριας, την διαχείριση των αντικειμένων που διατηρούνται στο παράθυρο της κύριας μνήμης. Πιο συγκεκριμένα, εισάγαμε και μελετήσαμε ένα σύνολο από διαφορετικές πολίτικες (policies) σχετιζόμενες με δυο κύριες εργασίες: την διάσχιση (traverse) και την απομάκρυνση (eviction) των αντικειμένων του παραθύρου. Και οι δύο αυτές εργασίες μπορούν να έχουν σημαντικές επιπτώσεις στον αριθμό των I/Os (Εισόδων/Εξόδων) καθώς και στον χρόνο επεξεργασίας (CPU Time). Πραγματοποιήσαμε εκτενή μελέτη των προτεινόμενων πολιτικών. Από την μελέτη προέκυψε, ότι σε δεδομένα με συγκεκριμένα χαρακτηριστικά οι πολιτικές αυτές μπορούν να μειώσουν τον αριθμό των ελέγχων κυριαρχίας (dominance checks) παραπάνω από 50%, οδηγώντας έτσι στην κατασκευή πιο αποδοτικών μεθόδων παραγωγής συνόψεων με βάση τον τελεστή κορυφογραμμής. Η συγκεκριμένη μελέτη δημοσιεύτηκε στο άρθρο [BiSS14].

3.4 Δημιουργία συνόψεων προσαρμοσμένων στις προτιμήσεις των χρηστών με τη μέθοδο της διαφοροποίησης

Ένας τρόπος υπολογισμού συνόψεων προσαρμοσμένες στις προτιμήσεις του εκάστοτε χρήστη είναι η μέθοδος της διαφοροποίησης (diversification) των αποτελεσμάτων αναζήτησης. Λαμβάνοντας υπόψη έναν ελκυστή, ένα σύνολο από απωθητές, και μια συλλογή αντικειμένων, ορισμένα σε ένα μετρικό χώρο, το ερώτημα μέγιστης έλξης, ελάχιστης απώθησης (ΜΕΕΑ) επιστρέφει το αντικείμενο που μεγιστοποιεί τη σταθμισμένη διαφορά της απόστασής του από τον ελκυστή και της συναθροιστικής απόστασης από τους απωθητές, δηλαδή είναι όσο το δυνατό κοντά στον ελκυστή και ταυτόχρονα μακριά από τους απωθητές.

Ερωτήματα ΜΕΕΑ προκύπτουν σε διάφορα προβλήματα βελτιστοποίησης, όπως το πρόβλημα της διαφοροποίησης αποτελεσμάτων αναζήτησης, όπου ο στόχος είναι να ανακτηθεί ένα σύνολο από αποτελέσματα που είναι σχετικά με ένα συγκεκριμένο ερώτημα αναζήτησης και ταυτόχρονα διαφορετικά μεταξύ τους. Στο βασικό ερώτημα ΜΕΕΑ, σε αντιστοιχία με τα ερωτήματα εγγύτερου γείτονα (ΕΓ), ο ελκυστής είναι το βέλτιστο σημείο στο χώρο. Ωστόσο, σε αντίθεση με τα ερωτήματα ΕΓ, η απάντηση σε ένα ΜΕΕΑ ερώτημα θα μπορούσε να είναι πολύ μακριά από τον ελκυστή. Παρόλο που τα ερωτήματα ΜΕΕΑ απαιτούν μόνο γραμμικό χρόνο στη χειρότερη περίπτωση, δείξαμε ότι είναι δυνατόν να κατασκευάσουμε έναν αλγόριθμο που είναι πολύ πιο γρήγορος στην πράξη, μελετώντας τις ιδιότητες της συνάρτησης βελτιστοποίησης και εισάγοντας κριτήρια αποκοπής. Χρησιμοποιώντας μια πολυδιάστατη δομή δεικτοδότησης, που προέρχεται από μια μεγάλη ποικιλία υποστηριζόμενων δομών, είμαστε σε θέση να επεξεργαστούμε ερωτήματα ΜΕΕΑ για μια ευρεία κατηγορία από συναρτήσεις βελτιστοποίησης και μετρικές απόστασης, αποτελεσματικά ως προς τις απαιτούμενες λειτουργίες Ε/Ε, και τάξεις μεγέθους πιο γρήγορα από ό,τι μια απλή γραμμική σάρωση.

Η εργασία μας έχει τις εξής συνεισφορές.

1. Ορίσαμε το ερώτημα μέγιστης έλξης, ελάχιστης απώθησης (MEEA) που μπορεί να χρησιμοποιηθεί ως βασικό εργαλείο για την παραγωγή διαφοροποιημένων αποτελεσμάτων αναζήτησης για μια πληθώρα από μετρικές απόστασης.
2. Μελετήσαμε τα γεωμετρικά χαρακτηριστικά του ερωτήματος και εισάγαμε κριτήρια αποκοπής ώστε να ελαχιστοποιήσουμε το χώρο αναζήτησης και να επιταχύνουμε τη διαδικασία παραγωγής της διαφοροποιημένης σύνοψης.
3. Παρουσιάζουμε έναν αλγόριθμο που εκμεταλλεύεται τα γεωμετρικά χαρακτηριστικά ώστε να κατευθύνει την αναζήτηση προς επιθυμητά αποτελέσματα αναζήτησης.
4. Παρουσιάσαμε μία εκτενή πειραματική μελέτη σε πραγματικά δεδομένα, η οποία αναδεικνύει την αποτελεσματικότητα της προτεινόμενης μεθόδου μας.
Τα αποτελέσματα δημοσιεύτηκαν στο άρθρο [SaD14].

Σαν συνέχεια αυτής της εργασίας, εξετάσαμε και ένα πρόβλημα που γενικεύει τα ερωτήματα MEEA. Συγκεκριμένα, θεωρούμε ότι έχουμε ένα σύνολο από ελκυστές, αντί για έναν, και το ζητούμενο είναι να βρούμε το αντικείμενο που είναι ταυτόχρονα κοντά στους ελκυστές και μακριά από τους απωθητές. Αυτή η επέκταση βρίσκει ευθεία εφαρμογή σε τεχνικές κατασκευής περιλήψεων που βασίζονται στη συσταδοποίηση (clustering) των αντικειμένων ενός οικοσυστήματος. Η συγκεκριμένη εργασία δημοσιεύτηκε στο άρθρο [SaD15].

3.5 Υπολογισμός στατιστικών μέτρων κεντρικότητας χρηστών

Ένας από τους στόχους της υποδράσης ΥΔ 4.2 είναι ο υπολογισμός κατάλληλων στατιστικών τα οποία πέρα από τα δεδομένα θα επιτρέπουν την αξιολόγηση των χρηστών του συστήματος και της εκτίμηση του ρόλου τους σε συσχέτιση με τους υπόλοιπους χρήστες του οικοσυστήματος. Στα πλαίσια αυτής της ενότητας μελετήσαμε ένα σημαντικό στατιστικό μέτρο για την αξιολόγηση διασυνδεδεμένων χρηστών (πχ ενός κοινωνικού δικτύου), αυτό της κεντρικότητας. Ειδικότερα, εστιάσαμε σε δίκτυα όπου γίνονται δημοσιεύσεις με γεωαναφορές, και θέλουμε να υπολογίσουμε μια παραλλαγή του μέτρου αυτού που να λαμβάνει υπόψη την τοποθεσία.

Η πανταχού παρουσία κινητών συσκευών οι οποίες μπορούν να αναγνωρίζουν τη τοποθεσία και ο πολλαπλασιασμός των κοινωνικών δικτύων έχουν συντελέσει στη δημιουργία Κοινωνικών Δικτύων με Επίγνωση Θέσης (ΚΔΕΘ). Σε τέτοια δίκτυα οι χρήστες συνάπτουν κοινωνικές σχέσεις και κάνουν δημοσιεύσεις με γεωαναφορές. Στα πλαίσια της ανάλυσης τέτοιων δικτύων είναι σκόπιμο να εντοπιστούν οι χρήστες που μπορούν να επηρεάσουν ένα μεγάλο αριθμό άλλων σημαντικών χρηστών, μέσα σε μια δεδομένη περιοχή του χώρου. Η επιστροφή μιας λίστας κατάταξης για τοπικά σημαίνοντες χρήστες ΚΔΕΘ είναι χρήσιμη για εφαρμογές viral marketing καθώς και για άλλα σενάρια αναλυτικής ανά περιοχή επεξεργασίας. Στα πλαίσια της δράσης δείξαμε ότι κάτω από ένα γενικό μοντέλο διάδοσης της επιρροής, το πρόβλημα είναι #P-δύσκολο, ενώ γίνεται επιλύσιμο σε πολυωνυμικό χρόνο σε ένα πιο περιορισμένο μοντέλο. Σύμφωνα με το πιο περιοριστικό μοντέλο, στη συνέχεια, δείχνουμε ότι το πρόβλημα μπορεί να μεταφραστεί στον υπολογισμό μιας παραλλαγής της λεγόμενης κεντρικότητας με βάση την εγγύτητα των χρηστών του κοινωνικού δικτύου, και προτείνουμε μια μέθοδο αξιολόγησης. Τα αποτελέσματα δημοσιεύτηκαν στο άρθρο [BoSB14].

4 Ανακεφαλαίωση

Το παρόν παραδοτέο Π4.2 παρουσιάζει τα αποτελέσματα της υποδράσης ΥΔ4.2 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ4.2 ήταν η ανάπτυξη τεχνικών για την παραγωγή στατιστικών και συνόψεων δεδομένων. Η έρευνα μας εστίασε όχι μόνο σε ντετερμινιστικά δεδομένα αλλά αναπτύχθηκαν τεχνικές κατάλληλες και για πιθανοτικά δεδομένα. Οι τεχνικές που αναπτύχθηκαν έχουν κεντρικό ρόλο στην ανάπτυξη κλιμακώσιμων εφαρμογών καθώς επιτρέπουν την προσεγγιστική αποτίμηση σύνθετων ερωτημάτων αναζήτησης χρησιμοποιώντας σημαντικά λιγότερους υπολογιστικούς πόρους. Πέρα των πλεονεκτημάτων στη ταχύτερη διαχείριση των εφαρμογών, οι τεχνικές που αναπτύχθηκαν επιτρέπουν την πραγματοποίηση προσωποποιημένων αναζητήσεων με διαφορετικά κριτήρια, την ευελιξία όσων αφορά τις μετρικές υπολογισμού της απόστασης/ομοιότητας ανάμεσα στα δεδομένα καθώς και την διαφοροποίηση (diversification) των αποτελεσμάτων μιας αναζήτησης. Επιπλέον μελετήθηκαν τρόποι υπολογισμού στατιστικών τα οποία επιτρέπουν

την αξιολόγηση της κεντρικότητα ενός χρήστη με βάση τις καταγεγραμμένες κοινωνικές συσχετίσεις.

Στα πλαίσια των ερευνητικών δραστηριοτήτων της υποδράσης προέκυψαν 4 δημοσιεύσεις και 2 μεταπτυχιακές εργασίες (βλ. παρακάτω πίνακα). Σε αυτές παρουσιάζονται αναλυτικά οι αλγόριθμοι και τεχνικές και μελετάται πειραματικά η απόδοση τους χρησιμοποιώντας κατανεμημένες τεχνολογίες συστημάτων μεγάλων δεδομένων όπως η τεχνική MapReduce.

Δημοσιεύσεις

[Σπ13] Α. Σπηλιόπουλος (επιβλέπων Ι. Κωτίδης). Υπολογισμός Ερωτημάτων Κοντινότερων Γειτόνων με Χρήση Τεχνικών Map-Reduce. Διπλωματική Εργασία ΠΜΣ, ΟΠΑ, 2013.

[Ma13] S-G. Mammas (επιβλέπων Α. Δεληγιαννάκης). Desigining Efficient Algorithms for Approximating Probabilistic Data. Master Thesis, Technical University of Crete, 2013.

[BiSS14] Nikos Bikakis, Dimitris Sacharidis, Timos Sellis. A Study on External Memory Scan-Based Skyline Algorithms. DEXA (1) 2014: 156-170

[BoSB14] Panagiotis Bouros, Dimitris Sacharidis, Nikos Bikakis. Regionally influential users in location-aware social networks. SIGSPATIAL/GIS 2014: 501-504

[SaD14] Dimitris Sacharidis, Antonios Deligiannakis. Maximum Attraction, Minimum Repulsion Queries. HDMS 2014.

[SaD15] Dimitris Sacharidis, Antonios Deligiannakis. Spatial Cohesion Queries. SIGSPATIAL/GIS 2015

Παράρτημα