Τμήμα Πληροφορικής Α.Π.Θ.

Θέματα Πτυχιακών Εργασιών 2010-2011

Απόστολος Ν. Παπαδόπουλος

1 Τυχαίοι αλγόριθμοι για την εύρεση κορυφογραμμής

Η εύρεση της *κορυφογραμμής* (skyline) ενός συνόλου πολυδιάστατων δεδομένων αποτελεί πολύ βασική λειτουργία για τα σύγχρονα συστήματα. Ωστόσο, όλοι οι ντετερμινιστικοί αλγόριθμοι που έχουν προταθεί έως τώρα θεωρούν ότι έχουμε πρόσβαση σε οποιοδήποτε δεδομένο με τυχαία προσπέλαση. Κάτι τέτοιο ωστόσο δεν είναι εφικτό σε μερικές εφαρμογές οι οποίες στηρίζονται αποκλειστικά στη σειριακή προσπέλαση των δεδομένων, ακόμη και αν απαιτούνται πολλαπλά περάσματα. Στην εργασία αυτή, στόχος είναι η υλοποίηση και βελτίωση ενός *τυχαίου* (randomized) αλγορίθμου για την εύρεση της κορυφοργραμμής. Απαιτείται καλή γνώση της γλώσσας C++.

2 Ερωτήματα συνάθροισης σε μεγάλα γραφήματα

Στην εργασία αυτή θα υλοποιηθούν ερωτήματα συνάθροισης (aggregation queries) σε μεγάλα γραφήματα. Ένα ερώτημα συνάθροισης εντοπίζει τις k κορυφές του γραφήματος που έχουν το καλύτερο score με βάση μία συνάρτηση βαθμολόγησης (ranking function) που επιδρά στις γειτονικές κορυφές κάθε κορυφής. Η μελέτη των αλγορίθμων θα πραγματοποιηθεί σε μεγάλα γραφήματα είτε συνθετικά, είτε πραγματικά. Απαιτείται καλή γνώση C++.

3 Αλγόριθμοι εξόρυξης από τυχαίες συνόψεις γραφημάτων

Επειδή τα μεγέθη γραφημάτων που χρησιμοποιούνται στις σύγχρονες εφαρμογές είναι μεγάλα, μία τακτική είναι πρώτα να μειώσουμε το μέγεθός τους και στη συνέχεια να εφαρμόσουμε αλγορίθμους εξόρυξης. Η εργασία κινείται σε αυτό ακριβώς το πλαίσιο. Δίνεται ένα σύνολο από γραφήματα των οποίων το μέγεθος πρέπει να μειωθεί, ώστε στη συνέχεια να εφαρμόσουμε αλγορίθμους εξόρυξης. Το ζήτημα είναι ότι επειδή οι αλγόριθμοι εξόρυξης εφαρμόζονται σε συνόψεις των γραφημάτων, αναμένεται ότι θα υπάρχει μία διαφοροποίηση στα αποτελέσματα (δηλαδή μπορεί να χάσουμε πληροφορία). Στον αλγόριθμο που θα υλοποιηθεί υπάρχουν πιθανοτικές εγγυήσεις (probabilistic guarantees). Απαιτείται καλή γνώση C++.

4 Εύρεση συχνών στοιχείων σε ροές δεδομένων

Στην εργασία αυτή καλείστε να αναπτύξετε μεθόδους εντοπισμού συχνών στοιχείων με ένα πέρασμα, σε μία ροή δεδομένων. Το πρόβλημα εμφανίζει πρακτικό και ερευνητικό ενδιαφέρον και χρησιμοποιείται όπου θέλουμε να εντοπίσουμε στοιχεία με μεγάλη συχνότητα εμφάνισης, όπως π.χ. διευθύνσεις IP σε μία ροή δεδομένων από έναν router, ή τηλεφωνικούς αριθμούς σε μία ροή που αναφέρεται σε κλήσεις μεταξύ συνδρομητών. Η βασική δυσκολία του προβλήματος είναι ότι τις περισσότερες φορές δεν επιτρέπεται να έχουμε γραμμικό χώρο σε σχέση με το πλήθος των δεδομένων της ροής. Απαιτείται καλή γνώση C++.

5 Αναζήτηση ακολουθιών DNA

Ένα από τα βασικότερα προβλήματα στη μελέτη βιολογικών δεδομένων είναι η ταυτοποίηση ή ο έλεγχος της ομοιότητας μεταξύ ακολουθιών DNA. Σκοπός της εργασίας είναι η υλοποίηση και πειραματική μελέτη εξειδικευμένων μεθόδων προσπέλασης με στόχο τη γρήγορη αναζήτηση ακολουθιών DNA. Απαιτείται καλή γνώση της γλώσσας C++.

Παρακαλούνται οι ενδιαφερόμενοι να στείλουν αντίγραφο αναλυτικής βαθμολογίας και να δηλώσουν τα θέματα για τα οποία ενδιαφέρονται με ένα μήνυμα στο papadopo@csd.auth.gr. Προθεσμία εκδήλωσης ενδιαφέροντος 1/12/2010.