ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΜΕΛΕΤΗ ΠΕΡΙΛΗΨΕΩΝ (ΣΚΙΤΣΩΝ) ΓΙΑ ΡΕΥΜΑΤΑ ΔΕΔΟΜΕΝΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Δ. Σαχαρίδης, 210 772 1402, dsachar@dblab.ece.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στη μελέτη ορισμένων περιληπτικών δομών (σκίτσα) και τη σχεδίαση-υλοποίηση τους ως βιβλιοθήκη σε C++.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: GNU C++

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Λόγω των περιορισμών χρόνου και χώρου που επιβάλλουν τα ρεύματα δεδομένων, προσεγγιστικές μέθοδοι πολλές φορές χρησιμοποιούνται για να απαντήσουν ερωτήματα με κάποιο ελεγχόμενο παράγοντα σφάλματος. Για παράδειγμα, ας δούμε δύο ερωτήματα, τα οποία μοιάζουν, αλλά μόνο το ένα μπορεί να απαντηθεί ακριβώς σε περιβάλλον ρευμάτων δεδομένων. Καλούμαστε να παρακολουθήσουμε την IP κίνηση που περνά από ένα δρομολογητή. Η πρώτη και απλή ερώτηση είναι "Πόσα πακέτα έχουν περάσει από τον δρομολογητή;", ενώ η δεύτερη και δυσκολότερη είναι "Πόσες διαφορετικές IP διευθύνσεις έχουν στείλει πακέτα σε αυτό τον δρομολογητή;". Ενώ η πρώτη που ζητάει ένα απλό COUNT μπορεί να απαντηθεί με τη χρήση ενός μόνο μετρητή, η δεύτερη (DISTINCT COUNT) απαιτεί στη χειρότερη περίπτωση χώρο τόσων bit όσες και οι πιθανές IP διευθύνσεις (της τάξης των 2^{32}), κάτι που είναι απαγορευτικό για τέτοια περιβάλλοντα.

Τα **σκίτσα (sketches)** είναι απλές δομές που διατηρούν κάποια περίληψη των ρευμάτων, ικανή να απαντήσει δύσκολα ερωτήματα προσεγγιστικά, με μικρό σφάλμα και χαμηλή πιθανότητα σφάλματος. Τα σκίτσα βασίζονται κυρίως σε τεχνικές κατακερματισμού και αθροίσματα τυχαίων μεταβλητών για την απάντηση των ερωτημάτων αυτών. Η διπλωματική αυτή εργασία συνίσταται στην βιβλιογραφική μελέτη των διαφόρων τύπων σκίτσων και στη κατηγοριοποίηση τους με βάση τις ερωτήσεις που μπορούν να απαντήσουν. Στη συνέχεια, θα γίνει η υλοποίηση των διάφορων μεθόδων (μετά τον απαραίτητο σχεδιασμό των κλάσεων) ως μία βιβλιοθήκη γενικής χρήσης (πρότυπος κώδικας σε C υπάρχει διαθέσιμος). Τέλος θα γίνει και αξιολόγηση των μεθόδων ανάλογα με την κατηγοριοποίηση.

ΧΡΗΣΙΜΕΣ ΓΝΩΣΕΙΣ:

Ανάλυση Αλγορίθμων (εύρεση χρονικής και χωρικής πολυπλοκότητας), Βασικές γνώσεις Πιθανοτήτων (π.χ. μέση τιμή, διασπορά, τυχαίες μεταβλητές, αμερόληπτες εκτιμήτριες, κτλ.), Ιδέες Αντικειμενοστραφούς Προγραμματισμού, Καλή γνώση C++ σε περιβάλλον Unix/Linux/Cygwin

ΧΡΗΣΙΜΟΙ ΣΥΝΔΕΣΜΟΙ:

- http://dbpubs.stanford.edu:8090/pub/2002-19 Εισαγωγή στα ρεύματα δεδομένων (σύστημα)
- http://www.cs.rutgers.edu/~muthu/stream-1-1.ps Εισαγωγή στα ρεύματα δεδομένων
- http://www.dblab.ece.ntua.gr/~dsachar/writeups/Data Stream Algorithms.ppt Εισαγωγική παρουσίαση κάποιων σκίτσων
- http://www.cs.rutgers.edu/~muthu/massdal-code-index.html Πρότυπος κώδικας σε C για τα περισσότερα σκίτσα