

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ
Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΑΝΑΠΤΥΞΗ ΕΥΡΕΤΗΡΙΩΝ ΓΙΑ ΣΥΝΘΕΤΑ ΔΕΔΟΜΕΝΑ

ΠΛΗΡΟΦΟΡΙΕΣ: Μ.Τερροβίτης, 210 772 1402, mter@dblab.ece.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στην υλοποίηση μιας υβριδικής δομής ευρετηρίου για τιμές-σύνολα. Η δομή θα βασίζεται στο συνδυασμό δομών από τον τομέα του Information Retrieval και του Data Mining και κλασσικών δομών βάσεων δεδομένων.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: C/C++/Linux

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Δεδομένα σε αρκετά ελεύθερη μορφή συναντάμε όλο και συχνότερα σε όλο και μεγαλύτερες ποσότητες. Η απλούστερη και συνηθέστερη μορφή τους είναι σύνολα από ένα απροσδιόριστο αριθμό τιμών. Παραδείγματα τέτοια είναι οι δοσοληψίες σε καταστήματα λιανικής πώλησης, π.χ. supermarket, δεδομένα εταιριών τηλεπικοινωνιών ακόμη και κείμενα, ως σύνολα λέξεων .

Σκοπός της διπλωματικής εργασίας είναι η κατασκευή μία δομής ευρετηρίου που θα συνδυάζει trie trees με inverted file ευρετήρια. Τα trie trees έχουν αναπτυχθεί και χρησιμοποιηθεί ευρύτατα σε περιοχές όπως το profiling και το information retrieval. Το πλεονέκτημα τους είναι η ταχύτητα με την οποία απαντούν τις διάφορες ερωτήσεις. Το κόστος αυτής της ταχύτητας είναι τεράστιες απαιτήσεις σε μνήμη που καθιστούν αδύνατη την χρήση τέτοιων δομών για πολύ μεγάλα σύνολα τιμών-συνόλων. Στα πλαίσια της διπλωματικής θα αναπτυχθεί μία νέα δομή που θα χρησιμοποιεί ένα συνδυασμό trie trees και inverted files ευρετηρίων ώστε να μειώνονται οι απαιτήσεις σε μνήμη με όσο το δυνατόν μικρότερη επίδραση στην ταχύτητα.. Αναλυτικότερα θα αναπτυχθούν τα εξής:

- Η δομή του ευρετηρίου. Για την υλοποίηση αυτού του μέρους υπάρχει η δυνατότητα να χρησιμοποιηθούν έτοιμα εργαλεία που υπάρχουν.
- Μηχανισμός αποτίμησης των ερωτήσεων
- Μηχανισμός καταγραφής των επιδόσεων.

Στα πλαίσια αυτής της διπλωματικής θα βρεθεί κανείς αντιμέτωπος με αλγοριθμικά προβλήματα και σημαντική προγραμματιστική δουλειά.