Τεχνολογία Βάσεων Δεδομένων
Εργασία 2021-2022
Υλοποίηση Χωρικών Μεθόδων Προσπέλασης

1. Εισαγωγή

Στην εργασία αυτή θα ασχοληθούμε με την οργάνωση και επεξεργασία *δεδομένων χώρου* (spatial data). Τα δεδομένα χώρου στην πιο απλή μορφή τους είναι σημεία σε έναν πολυδιάστατο χώρο, όπως για παράδειγμα σημεία στο επίπεδο ή στον τρισδιάστατο χώρο. Ωστόσο, εύκολα μπορούμε να δημιουργήσουμε σημεία και σε χώρους με περισσότερες διαστάσεις. Στην πράξη, το πλήθος των διαστάσεων είναι παράμετρος η οποία λαμβάνεται υπόψη κατά την οργάνωση των δεδομένων και κατά την επεξεργασία των ερωτημάτων. Στόχος της εργασίας είναι η κατασκευή μίας δομής δεδομένων δευτερεύουσας μνήμης η οποία έχει τη δυνατότητα να οργανώνει πολυδιάστατα δεδομένα και να υποστηρίζει βασικά ερωτήματα όπως: **ερωτήματα περιοχής** (range queries), **ερωτήματα πλησιέστερων γειτόνων** (k-nearest neighbor queries). Επίσης, θα πρέπει να υποστηρίζονται οι λειτουργίες εισαγωγής, διαγραφής και μαζικής κατασκευής του καταλόγου. Θα δουλέψετε σε ομάδες των δύο ατόμων. Η εργασία λαμβάνει το 50% του τελικού βαθμού.

2. Αναλυτικότερα

Η δομή που θα υλοποιήσετε είναι το **R*-tree** και αποτελεί βελτίωση του **R-tree**. Τα δεδομένα θα πρέπει να τα πάρετε από το OpenStreetMap το οποίο σας επιτρέπει να *κατεβάσετε* περιοχές του κόσμου με σημεία ενδιαφέροντος. Μπορείτε να κατεβάσετε δεδομένα από οποιαδήποτε περιοχή θέλετε και οποιουδήποτε τύπου. Οι εγγραφές που θα δημιουργηθούν θα πρέπει να περιέχουν τουλάχιστον κάποιο id, το όνομα του σημείου, και τις LAT-LON συντεταγμένες. Μαζί με την εκφώνηση της εργασίας θα βρείτε και ένα παράδειγμα αρχείου OSM.

Οι εγγραφές αυτές θα πρέπει να αποθηκευθούν σε αρχείο (π.χ., datafile) το οποίο θα αποτελείται από blocks μεγέθους $\boldsymbol{B}$ **= 32KB**. Το κάθε block περιέχει ένα σύνολο από εγγραφές και επίσης έχει ένα μοναδικό *blockid* το οποίο δεν αλλάξει. Το πρώτο block του αρχείου είναι το block0 μέσα στο οποίο μπορείτε να αποθηκεύσετε βοηθητικές πληροφορίες, όπως το πλήθος των εγγραφών του αρχείου, το πλήθος των blocks κλπ. Τα δεδομένα θα πρέπει να αποθηκεύονται από το block1 και μετά.

Ο κατάλογος (index) αποθηκεύεται και αυτός σε ξεχωριστό αρχείο (έστω indexfile) και οργανώνει τις εγγραφές που είναι αποθηκευμένες στο datafile. Επομένως, στα φύλλα του καταλόγου αποθηκεύονται οι συντεταγμένες του σημείου και ένα Record ID που στην ουσία δείχνει σε ποιό block του datafile είναι αποθηκευμένη η συγκεκριμένη εγγραφή και σε ποιό slot. Η δομή του R*-tree που θα υλοποιήσετε πρέπει να υποστηρίζει τις ακόλουθες λειτουργίες:

- εισαγωγή εγγραφής (insertion),
- διαγραφή εγγραφής (deletion),
- ερώτημα περιοχής (range query),
- ερώτημα πλησιέστερων γειτόνων (k-nn query),
- ερώτημα κορυφογραμμής (skyline query),
- μαζική κατασκευή του δένδρου bottom-up

Προσοχή, ο κώδικάς σας θα πρέπει να χρησιμοποιεί το πλήθος των διαστάσεων ως παράμετρο. Δηλαδή, να μην γράψετε κώδικα θεωρώντας ότι αναγκαστικά το πλήθος των διαστάσεων είναι 2. Αυτό σημαίνει ότι ο κώδικας θα πρέπει να μπορεί να τρέχει και για 3 ή περισσότερες διαστάσεις.

3. Παραδοτέα

Θα πρέπει να παραδώσετε τον πηγαίο κώδικα (C++, Java, C#) και μία τεχνική αναφορά στην οποία θα αναλύετε τη μεθοδολογία που χρησιμοποιήσατε, πως υλοποιήσατε τη δομή του R*-tree και επίσης παραδείγματα από την εκτέλεση ερωτημάτων περιοχής, k-nn και κορυφογραμμής, εκτέλεση μαζικής κατασκευής μαζί με χρόνους εκτέλεσης καθώς και συγκρίσεις με τη σειριακή αναζήτηση (δηλαδή πόσο χρόνο χρειάζεται η σειριακή αναζήτηση στο datafile για ερωτήματα περιοχής ή πλησιέστερου γείτονα). Επίσης πρέπει να έχουμε και σύγκριση για τις δύο τεχνικές κατασκευής του καταλόγου, δηλαδή εισαγωγή των στοιχείων ένα-προς-ένα και μαζική κατασκευή. Στην αναφορά σας θα πρέπει να υπάρχουν και ενδεικτικές γραφικές παραστάσεις ή πίνακες που να δείχνουν τους χρόνους εκτέλεσης όσο μεγαλώνει η περιοχή ενδιαφέροντος *R* για να ερωτήματα περιοχής και όσο αυξάνει το *k* για ερωτήματα πλησιέστερων γειτόνων.

Ημερομηνία παράδοσης Κυριακή 26 Ιουνίου 2022.