1059633 Νικηφόρος - Γιώργος Παπαγεωργίου
1041815 Εμμανουήλ Μηναδάκης
Αλέξανδρος Ξιάρχος 1059619

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ · ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ & ΥΠΟΛΟΓΙΣΤΙΚΗ ΓΕΩΜΕΤΡΙΑ

PROJECT 1 · 2022-2023

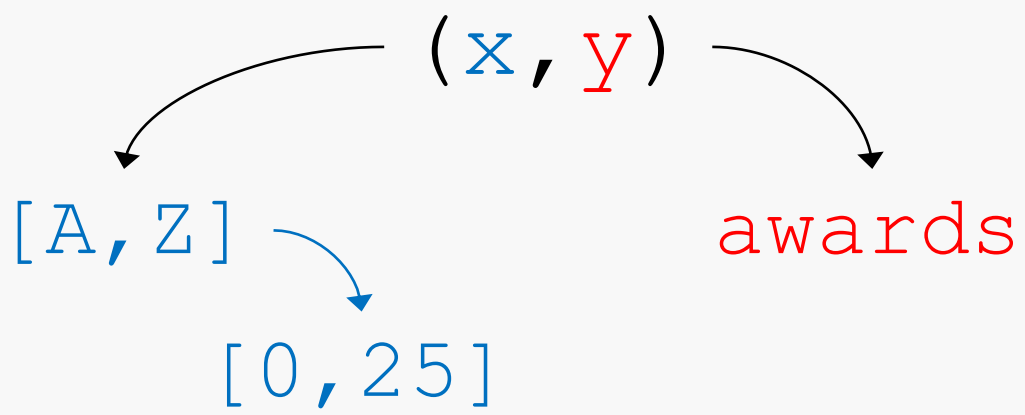
1 ΚΑΤΑΣΚΕΥΗ DATASET

- Κατασκευή δικού μας dataset χρησιμοποιώντας την βιβλιοθήκη `BeautifulSoup` για ανάλυση του HTML περιεχομένου της σελίδας με τη λίστα επιστημόνων της επιστήμης των υπολογιστών και εξαγωγή όλων των επιστημόνων.
- Από τα URLs των επιστημόνων συλλέγονται οι ζητούμενες πληροφορίες (επώνυμο, βραβεία και εκπαίδευση) χρησιμοποιώντας regex expressions, HTML tag parsing και τεχνικές string manipulation. Τα δεδομένα εισάγονται σε ένα `DataFrame`.
- Λόγω της ιδιομορφίας κάθε σελίδας χρησιμοποιήθηκε το διορθωτικό αρχείο `corrections.txt` για τις περιπτώσεις που δεν μπορούσαν να εξαχθούν αυτοματοποιημένα.
- Τα τελικά δεδομένα των 254 επιστημόνων εξάγονται στο αρχείο `scientists_data.csv`.

	surname	awards	education
0	Khan	10	Khan was a Bright Sparks scholar and received ...
1	Aaronson	4	Aaronson grew up in the United States, though ...
2	Abebe	3	Abebe was born and raised in Addis Ababa, Ethi...
3	Abelson	1	Abelson graduated with a Bachelor of Arts degr...
4	Abiteboul	4	The son of two hardware store owners, Abitebou...

2 ΥΛΟΠΟΙΗΣΗ ΚΑΤΑΣΚΕΥΗΣ ΚΑΙ ΑΝΑΖΗΤΗΣΗΣ ΣΕ ΠΟΛΥΔΙΑΣΤΑΤΗ ΜΟΡΦΗ

- Για κάθε δομή υλοποιήθηκαν ξεχωριστά μια συνάρτηση κατασκευής της δομής (`build_tree()`) και μια συνάρτηση αναζήτησης στη δομή (`query_tree()`):
- Η `build_tree()` αφού διαβάσει το `.csv`, δημιουργεί ένα αντιπροσωπευτικό σημείο (`x`, `y`) για κάθε επιστήμονα. Ως x ορίζεται η αριθμητική τιμή του αρχικού του επωνύμου του επιστήμονα, και ως y ο αριθμός των βραβείων που έχει λάβει.
- Σε κάποιες δομές χρησιμοποιήθηκε και ο αριθμός γραμμής (`index`) του επιστήμονα του `DataFrame` για να ξεχωρίσουμε τους επιστήμονες με ίδια (`x`, `y`).
- Με την δημιουργία των σημείων κατασκευάζουμε την κάθε δομή βάσει αυτών.



2 ΥΛΟΠΟΙΗΣΗ ΚΑΤΑΣΚΕΥΗΣ ΚΑΙ ΑΝΑΖΗΤΗΣΗΣ ΣΕ ΠΟΛΥΔΙΑΣΤΑΤΗ ΜΟΡΦΗ

- Η συνάρτηση αναζήτησης δέχεται τέσσερις παραμέτρους:

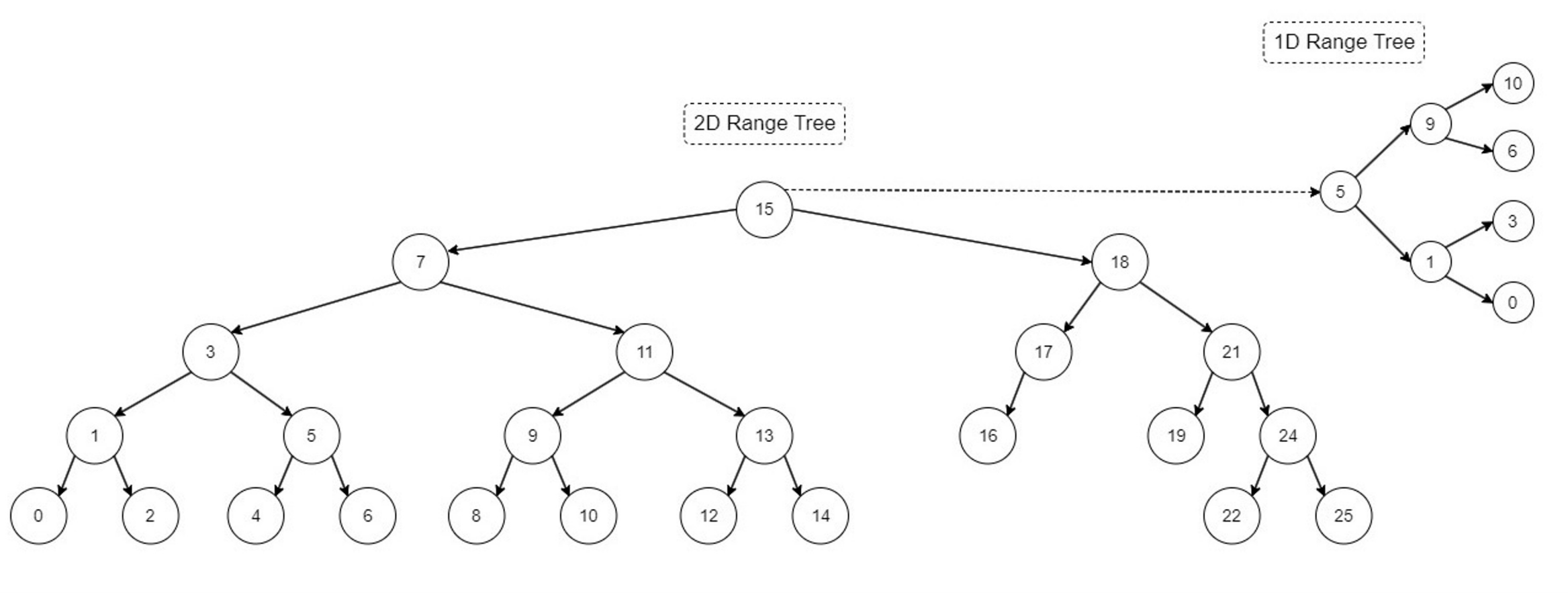
```
query_tree(tree, min_letter, max_letter, num_awards)
```

- `tree`: η πολυδιάστατη δομή που επιστράφηκε από τη συνάρτηση `build_tree()`, δύο γράμματα που αντιπροσωπεύουν το ελάχιστο και το μέγιστο όριο της συντεταγμένης `x` και έναν αριθμό βραβείων που αντιπροσωπεύει το ελάχιστο όριο της `y`.
- Η συνάρτηση αποστέλλει ερώτημα αναζήτησης στη δομή για τα δοθέντα διαστήματα τιμών. Βάσει των αποτελεσμάτων της αναζήτησης ανακτά τα δεδομένα των επιστημόνων από το `.csv` αρχείο και το επιστρέφει στη λίστα `final_results`.

2.1 RANGE TREE

- Η υλοποίηση του 2D Range Tree πραγματοποιήθηκε με την κατασκευή ισοσταθμισμένων δυαδικών δέντρων αναζήτησης (BBSTs).
- Αρχικά κατασκευάζεται ένα κύριο BBST βάσει των συντεταγμένων x των σημείων. Κάθε κόμβος του αποθηκεύει ένα 1D Range Tree (y-tree) με όλα τα σημεία με ίδιο x με τον κόμβο. Κάθε y-tree είναι και αυτό BBST, κατασκευασμένο βάσει των συντεταγμένων y των σημείων που περιέχει.
- Έτσι επιτρέπεται η αναζήτηση σημείων χρησιμοποιώντας το y για σημεία που έχουν ίδιο x.
- Κατά την αναζήτηση ενός εύρους, το κύριο δέντρο προσπελαύνεται πρώτα για να βρεθούν οι κόμβοι με τα x που ανήκουν στο επιθυμητό διάστημα x. Για αυτούς τους κόμβους προσπελαύνεται το αντίστοιχο y-δέντρο τους για τα σημεία y.
- Η συνδυασμένη προσπέλαση των δύο δέντρων επιτρέπει την αποτελεσματική εύρεση όλων των σημείων που βρίσκονται εντός ενός ερωτήματος αναζήτησης (range query).

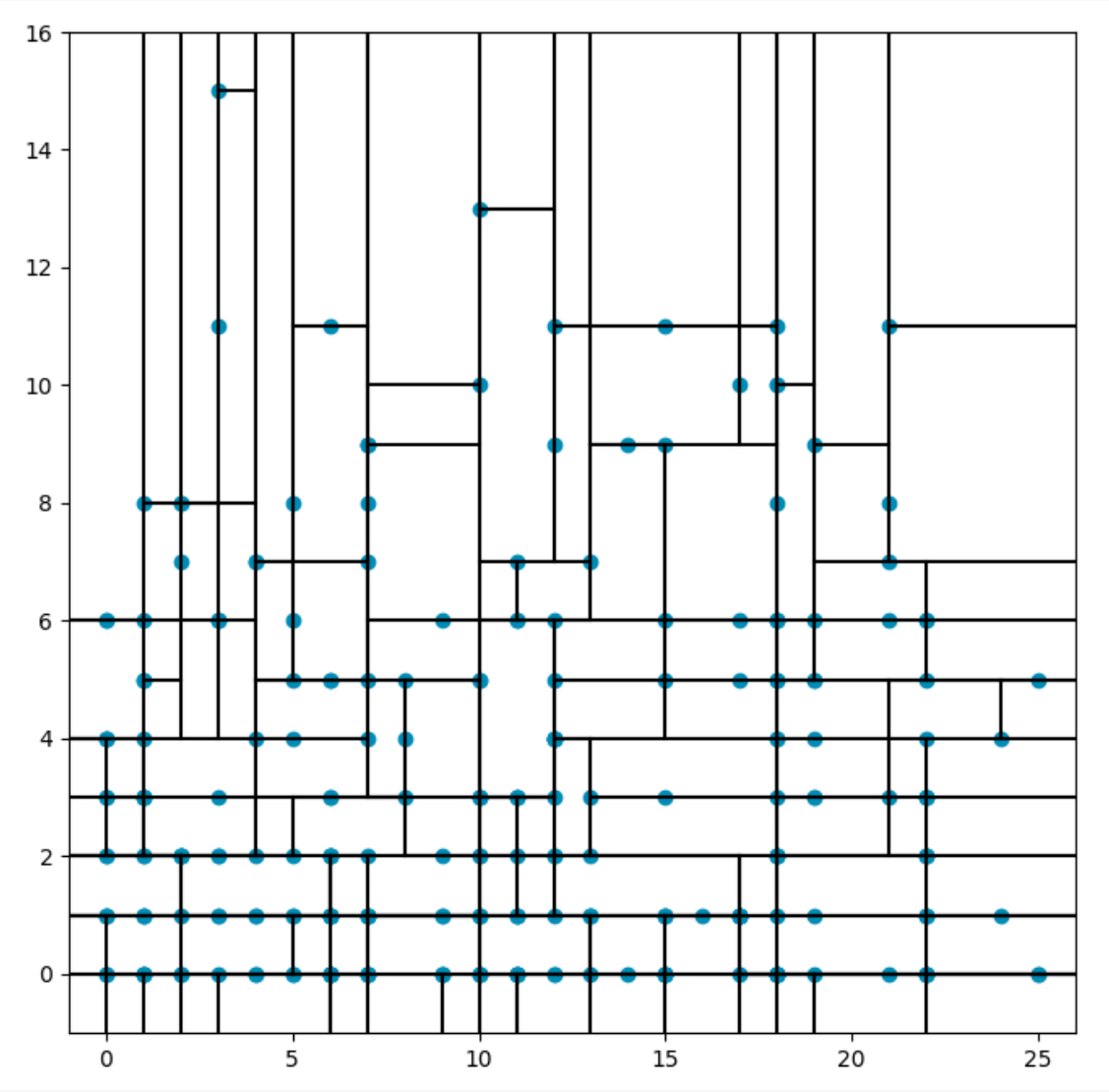
2.1 RANGE TREE



2.2 K-D TREE

- Βρισκόμαστε στον δισδιάστατο χώρο, άρα $K=2$. Τα K-D Trees διαχωρίζουν αυτόν τον χώρο σε ημιεπίπεδα. Η κατασκευή του K-D Tree συνεπάγει την διχοτόμηση αυτού του χώρου σε ημιεπίπεδα από δύο άξονες, τον x και τον y.
- Επιλέγεται ένας αρχικός άξονας και ένα σημείο (x, y) που θα τον τέμνει. Τα υπόλοιπα σημεία θα ανήκουν πλέον σε δύο υποσύνολα, αριστερά και δεξιά του, ανάλογα με τις συντεταγμένες τους. Όσο προστίθενται σημεία, ανάλογα με το βάθος του δέντρου, οι διαχωρισμοί θα εναλλάσσονται διαδοχικά σε κάθετους και οριζόντιους, οι οποίοι αντιστοιχίζονται στον x και στον y άξονα.
- Το αποτέλεσμα της κατασκευής είναι ένα ισοσταθμισμένο δυαδικό δέντρο με κάθε κόμβο του να έχει έναν "προσωπικό" άξονα ο οποίος χωρίζει το χώρο σε όλο και μικρότερα ημιεπίπεδα. Για την αναζήτηση των σημείων οδηγούμαστε όλο και βαθύτερα στο δέντρο, αναζητώντας μόνο τα σημεία που βρίσκονται μέσα στο ορθογώνιο διάστημα που επιθυμούμε.

2.2 K-D TREE



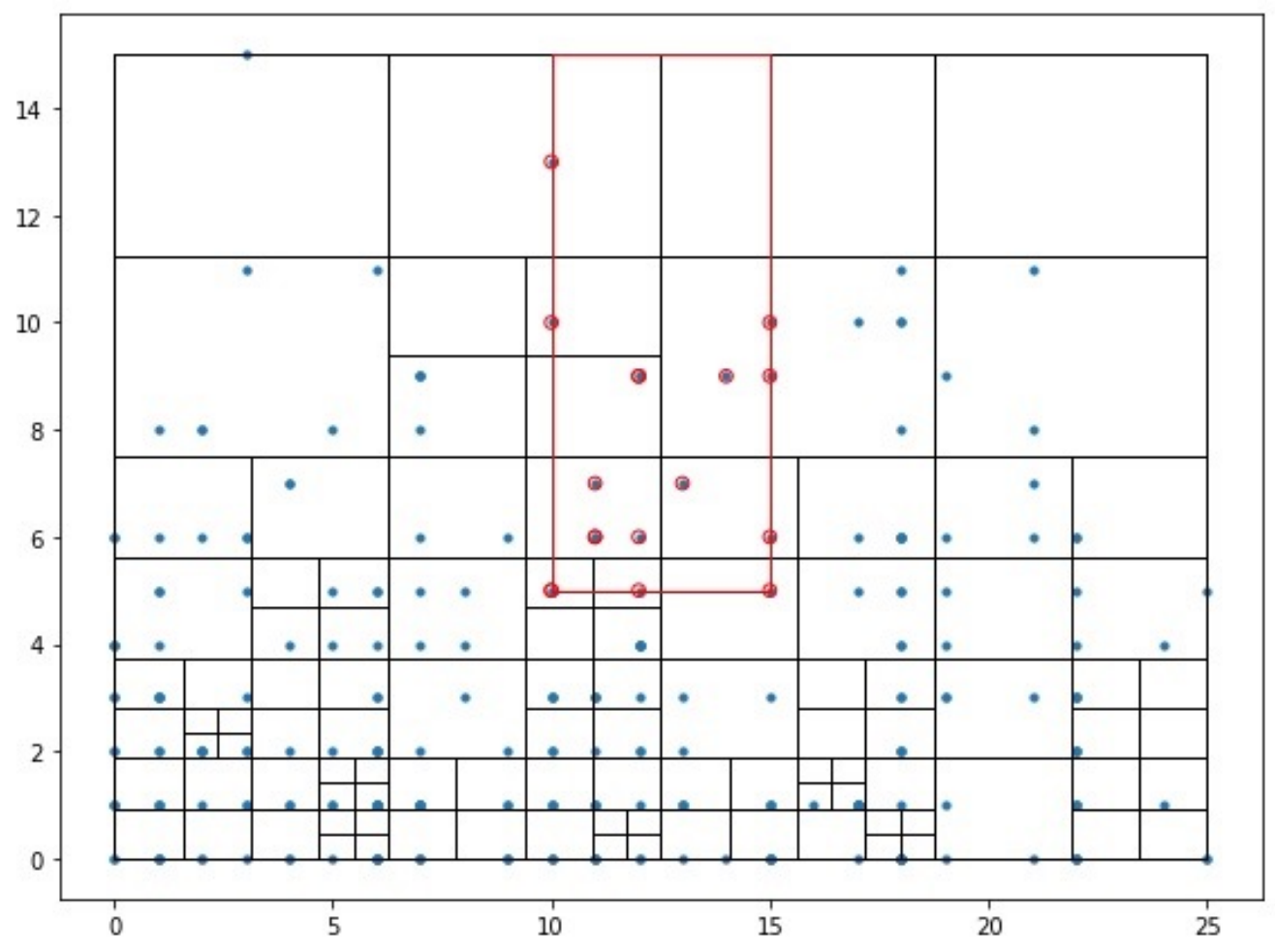
2.3 R-TREE

- Η βασική ιδέα του R-tree είναι να ομαδοποιεί τα δεδομένα σε ορθογώνιες περιοχές, οι οποίες αποτελούν τους κόμβους του δέντρου. Καθώς το δέντρο αναπτύσσεται, τα ορθογώνια μπορεί να υπερκαλύπτονται αλλά προσπαθούν να ελαχιστοποιούν την υπερκάλυψη και το μέγεθός τους.
- Ο κώδικας που υλοποιήσαμε περιλαμβάνει την κλάση `Rtree`, χρησιμοποιώντας τη βιβλιοθήκη `rtree` της `libspatialindex`. Η μέθοδος `insert()` προσθέτει ένα στοιχείο στο δέντρο ψάχνοντας τον πιο κατάλληλο κόμβο για την εισαγωγή του. Αν αυτός ο κόμβος υπερβαίνει το μέγιστο πλήθος στοιχείων, διαιρείται σε δύο νέους κόμβους. Χρησιμοποιούνται αλγόριθμοι που βελτιστοποιούν την διαίρεση ελαχιστοποιώντας την υπερκάλυψη τους ώστε το δέντρο να παραμένει ισορροπημένο.
- Η μέθοδος `search()` χρησιμοποιεί ένα δοθέν bounding box ως όρισμα και καλεί την `intersection()` η οποία βρίσκει όλα τα στοιχεία που τέμνονται με αυτό το bounding box.

2.4 QUAD TREE

- Το Quad Tree χωρίζει το χώρο σε τέσσερα τμήματα (ή κόμβους) και κάθε τμήμα μπορεί να χωριστεί περαιτέρω ανάλογα με το πλήθος των σημείων που περιέχει.
- Κατά την εισαγωγή ενός νέου σημείου το δέντρο ελέγχει σε ποιον κόμβο ανήκει και το προσθέτει σ' αυτόν. Αν ο κόμβος έχει ήδη το μέγιστο επιτρεπόμενο πλήθος σημείων (το ορίζουμε ως 4), ο κόμβος διασπάται και το σημείο προστίθεται στο κατάλληλο υπο-κόμβο απ' αυτούς που προκύπτουν.
- Η αναζήτηση σημείων σε ένα Quad Tree είναι αποτελεσματική, καθώς το δέντρο επιτρέπει την ταχεία πρόσβαση σε συγκεκριμένες περιοχές του χώρου. Αν ζητηθούν να βρεθούν όλα τα σημεία εντός ενός ορθογωνίου, το δέντρο ελέγχει μόνο τους κόμβους που τέμνουν το ορθογώνιο, αγνοώντας όλους τους υπόλοιπους.

2.4 QUAD TREE



Με κόκκινο χρώμα είναι σημειωμένο το ορθογώνιο αναζήτησης
(search boundary) και τα σημεία που περιέχονται σε αυτό.

3 ΥΛΟΠΟΙΗΣΗ LOCALITY SENSITIVITY HASHING (LSH)

- .