Projects: Μελέτη-Υλοποίηση και Πειραματική Αξιολόγηση Πολυδιάστατων Δομών Δεδομένων με Εφαρμογές τους

Υπεύθυνοι Καθηγητές: Σπύρος Σιούτας (Καθηγητής ΤΜΗΥΠ), Κων/νος Τσίχλας (Αναπληρωτής Καθηγητής ΤΜΗΥΠ).

Στόχος: Στόχος του project είναι η υλοποίηση πολυδιάστατων δομών δεδομένων σε Περιβάλλον Προγραμματισμού της αρεσκείας σας, με προτίμηση τις γλώσσες scala ή python και η πειραματική αξιολόγησή τους με data sets συνθετικά (artificial synthetic-data sets) ή ακόμα και πραγματικά (rea-data sets) στις βασικές πράξεις: Build, Insert, Delete, Update, Searching (Exact Match Queries, Range Queries), *Similarity Queries, kNN Queries*, Interval Queries, Stabbing Queries, Top-k Queries.

PROJECT-1 (100%):
Range and Similarity Queries σε Σύνολα Κειμένων: Υλοποίηση multi-dimensional Index δομής (k-d trees, quad trees, Range Trees και R-trees) που θα δεικτοδοτεί ένα σύνολο από κείμενα που θα προκύψουν από επιστήμονες επιστήμης υπολογιστών https://en.wikipedia.org/wiki/List_of_computer_scientists και θα είναι της μορφής: **(Surname:String, #Awards:Integer, Education:text-vector, #DBLP_Record). Πιο συγκεκριμένα, το index θα δημιουργηθεί ως προς τα τρία (3) πεδία (surname, #awards, #DBLP_Record)** προκειμένου να εντοπίσει τους επιστήμονες που το όνομά τους ανήκει αλφαβητικά σε ένα εύρος τιμών, ο αριθμός βραβείων που έχουν αποσπάσει να είναι μεγαλύτερος από ένα **user_defined_threshold** και ο αριθμός δημοσιεύσεων να είναι επίσης σε ένα έυρος τιμών. Στη συνέχεια, στο σύνολο των κειμένων που θα προκύψει από το παραπάνω πολυδιάστατο index, θα εκτελούνται ερωτήματα ομοιότητας ως προς το τρίτο πεδίο (education) που περιγράφει την εκπαίδευση που έχει λάβει κάθε ένας από αυτούς τους επιστήμονες με βάση τη μέθοδο LSH. Π.χ. σκεφτείτε ερωτήματα της μορφής: ***«Βρείτε τους επιστήμονες της επιστήμης υπολογιστών από τη ΒΔ Wikipedia που το γράμμα τους να ανήκει στο διάστημα [A, G], να έχουν αποσπάσει >4 βραβεία, ο αριθμός δημοσιεύσεων στο DBLP Record να ανήκει στο εύρος [100, 200] και να έχουν ποσοστό ομοιότητας εκπαίδευσης >50%»***. Να συγκριθούν πειραματικά οι 4 μέθοδοι: **k-d + LSH, Quad+LSH, Range+LSH, R-trees + LSH.**

PROJECT-2 (100%):
(A) [30%] Για κάθε επιστήμονα, βάζουμε στο «παιχνίδι» και ένα έξτρα πεδίο, τη λίστα χρονικών διαστημάτων (ή τμημάτων) στα οποία ο συγκεκριμένος επιστήμονας έχει δημοσιεύσεις στο DBLP_Record**.** Για παράδειγμα ο David J. Brown (https://dblp.org/pid/12/1509.html) έχει δημοσιεύσει στα παρακάτω χρονικά διαστήματα (time intervals or time segments): [1996, 1998], [2000, 2000], [2003, 2005], [2008, 2010], [2022, 2022]. Αποθηκεύστε τα παραπάνω intervals (ή segments) σε ένα INTERVAL TREE και σε ένα SEGMENT TREE, προκειμένου να απαντήσετε interval και stabbing Queries της μορφής: «Βρείτε τους επιστήμονες που σημοσίευσαν συνεχόμενα στο χρονικό διάστημα [2008, 2012]», «Βρείτε τους επιστήμονες που σημοσίευσαν τη χρονολογία 2010».

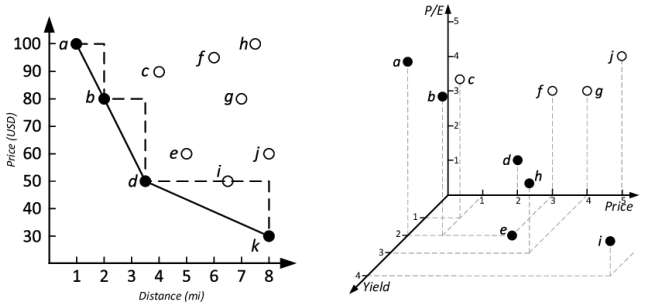
(B) [15%] Θεωρείστε το σύνολο των 3-dimensional points P1={(surname, #Awards, #DBLP_Record)}. Υπολογίστε το κυρτό τους περίβλημα **(Convex Hull)**.

(Γ) [15%] Θεωρείστε το σύνολο των 2-dimensional points P2={(#Awards, #DBLP_Record)}. Υπολογίστε τον SKYLINE OPERATOR ο οποίος είναι υποσύνολο του Convex Hull προβλήματος και χρησιμοποιείται στη γλώσσα SQL ως εξής:

```
SELECT ... FROM ... WHERE ...
GROUP BY ... HAVING ...
SKYLINE OF [DISTINCT] d1 [MIN | MAX | DIFF],
                  ..., dm [MIN | MAX | DIFF]
ORDER BY ...
```

Το ερώτημα επιλέγει όλα τα K-διάστατα σημεία των οποίων οι διαστάσεις d_1, ... d_m (m<=K) ελαχιστοποιούνται, μεγιστοποιούνται ή παίρνουν μία διαφορετική τιμή. Στην περίπτωση μας έχουμε 2 διαστάσεις, d1= #Awards και d2=#DBLP_Records. Μας ενδιαφέρει να υπολογίσουμε τα 4 υποσύνολα που συνθέτουν το κυρτό τους περίβλημα και αυτά είναι:
1ο subset: MIN d1, MIN d2
2ο subset: MIN d1, MAX d2
3ο subset: MAX d1, MIN d2
4ο subset: MAX d1, MAX d2

Για παράδειγμα στο σχήμα 1 (α) d1=distance (m), d2= price ($) και η γραμμή **c b d k** συμβολίζει το 1ο subset του CH(P). (β) Αντίστοιχο παράδειγμα για 3 διαστάσεις, όπου εδώ έχουμε 8 περιπτώσεις (ή subsets) από MIN, MIN, MIN μέχρι MAX, MAX, MAX. Εδώ το CH(P) είναι η επιφάνεια σφαίρας και τα αντίστοιχα 8 subsets είναι οι 8 υπο-φλοιοί που συνθέτουν την τρισδιάστατη κυρτή επιφάνεια της σφαίρας.

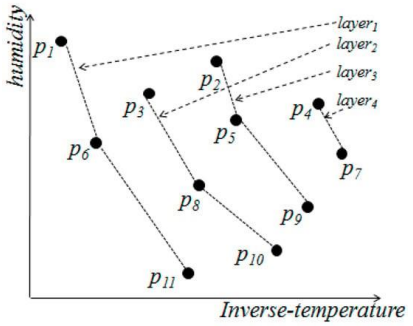


(a) 2-D skyline of hotels of (X = min, Y = min)

(b) 3-D skyline of stocks with attributes (X = min, Y = min, Z = max)

Σχήμα 1. (α) 2D Skyline-Operator (MIN Distance, MIN Price) (β) 3D Skyline Operator (MIN X, MIN Y, MAX Z)

(Δ) [10%] Οι περισσότερες μηχανές αναζήτησης υλοποιούν τον γενικευμένο "skyline_layers" operator ο οποίος στη γλώσσα της υπολογιστής πολυπλοκότητας είναι γνωστός με το όνομα LAYERS OF MAXIMA και αναπαρίσταται από τα παρακάτω σχήμα:



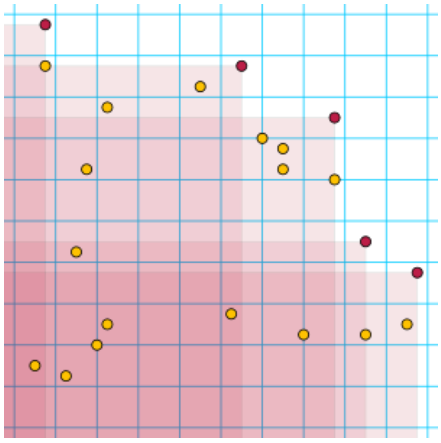
Σχήμα 2: Skyline Layers σε 2 διαστάσεις

Στο παράδειγμα του σχήματος 2, υπολογίζω το 1Ο επίπεδο L1=skyline (P) (min, min) και στη συνέχεια το 2ο L2=skyline (P-L1) (min, min) , το 3ο L3=skyline (P-L1-L2)

(min,min) κ.ο.κ.

Υπολογίστε τα top-k skyline layers, όπου k=user defined threshold, για το σύνολο P2 του (Γ) ερωτήματος.

(Ε) [30%] Για μεγαλύτερη απόδοση, ο skyline operator μπορεί να υπολογιστεί κάνοντας χρήση του maxima of a point set προβλήματος της υπολογιστικής γεωμετρίας καθώς και αποδοτικών πολυδιάστατων δομών δεδομένων, κυρίως k-d trees, R-trees ή priority search trees αν μιλάμε για 2 διαστάσεις. Δείτε το παρακάτω σχήμα:



Σχήμα 3: The five red points are the maxima of the set of all the red and yellow points. The shaded regions show the points dominated by at least one of the five maxima.

Σημείωση: **Maxima Set Problem**, has been studied as a variant of the convex hull. In computational geometry, a point p in a finite set of points S is said to be *maximal* or *non-dominated* if there is no other point q in S whose coordinates are all greater than or equal to the corresponding coordinates of p.

Στο σχήμα 3, έχουμε την περίπτωση του MAX X, MAX Y Skyline Operator. Ξεκινάμε είτε από το πρώτο δεξιά σημείο του skyline set, δηλαδή το red point με την maximum X coordinate ή το πρώτο αριστερά σημείο του skyline set, δηλαδή το red point με την maximum Y coordinate. Υποθέτουμε ότι ισχύει το πρώτο σενάριο. Τα σημεία του συνόλου που γίνονται dominated από αυτό το red point, δηλαδή τα σημεία με μικρότερο X και Y coordinate, είναι αυτά που ανήκουν στην 3-sided περιοχή που είναι σκιασμένη και έχει πάνω διαγώνια κορυφή το συγκεκριμένο red point. Με τη χρήση αποδοτικής πολυδιάστατης δομής, κάνω prune (ή και batch deletion) τα σημεία αυτά από το υπόλοιπο dataset. Συνεχίζω επαναληπτικά με το σημείο που έχει τη μέγιστη X συντεταγμένη στο εναπομείναν σύνολο μέχρι να ακουμπήσω το σημείο με την μέγιστη Y συντεταγμένη (ή μέχρι να αδειάσει η δομή). Έτσι, στο τέλος προκύπτει το σύνολο των red points. Υπολογίστε το skyline operator για το σύνολο P2 του (Γ) ερωτήματος, κάνοντας χρήση του MAXIMA SET προβλήματος καθώς και αποδοτικής πολυδιάστατης δομής δεδομένων της αρεσκείας σας. Επειδή είμαστε σε 2 διαστάσεις, σκεφτείτε τη χρήση του priority search tree.

Προαπαιτούμενες Γνώσεις: Δομές Δεδομένων, Αλγόριθμοι και Πολυπλοκότητα, Βάσεις Δεδομένων, Αντικειμενοστραφής Προγραμματισμός, Συναρτησιακός Προγραμματισμός (Functional Programming).

Παραπομπές:

1. Διαφάνειες Μαθήματος και Βιβλίο Α. Τσακαλίδη
2. https://en.wikipedia.org/wiki/Range_tree
3. https://en.wikipedia.org/wiki/K-d_tree
4. https://en.wikipedia.org/wiki/Quadtree
5. https://en.wikipedia.org/wiki/Interval_tree
6. https://en.wikipedia.org/wiki/Segment_tree
7. https://en.wikipedia.org/wiki/Priority_search_tree
8. https://en.wikipedia.org/wiki/Bloom_filter
9. https://en.wikipedia.org/wiki/MinHash
10. https://en.wikipedia.org/wiki/R-tree
12. https://en.wikipedia.org/wiki/Convex_hull
13. https://en.wikipedia.org/wiki/Skyline_operator
14. https://en.wikipedia.org/wiki/Maxima_of_a_point_set

ΣΗΜΕΙΩΣΗ 1: Το προς επεξεργασία σύνολο δεδομένων (real dataset) θα μπορούσε να αντληθεί είτε manually είτε αυτόματα μέσω ενός προσαρμοσμένου web crawler στο URL: https://en.wikipedia.org/wiki/List_of_computer_scientists. Συνίσταται το δεύτερο (χωρίς αυτό να είναι υποχρεωτικό).

ΣΗΜΕΙΩΣΗ 2: ΠΑΡΑΔΟΤΕΑ ΚΑΙ ΤΡΟΠΟΣ ΕΞΕΤΑΣΗΣ: DEMO ΠΑΡΟΥΣΙΑΣΗ ΚΑΙ ΠΑΡΑΔΟΣΗ ΕΚΤΥΠΩΜΕΝΗΣ ΤΕΧΝΙΚΗΣ ΑΝΑΦΟΡΑΣ ΚΑΘΩΣ ΚΑΙ ΟΛΟΥ ΤΟΥ ΨΗΦΙΑΚΟΥ ΥΛΙΚΟΥ (ZIP/RAR FILE) [**ΕΞΕΤΑΣΤΙΚΗ ΠΕΡΙΟΔΟΣ ΙΑΝΟΥΑΡΙΟΥ – ΦΕΒΡΟΥΑΡΙΟΥ**]. ΑΡΙΘΜΟΣ ΜΕΛΩΝ ΑΝΑ ΟΜΑΔΑ: **4-5**.