Department of Informatics
Aristotle University of Thessaloniki
Fall 2016-2017

Multimedia Database Systems

Indexing Part B

Metric-based Indexing Techniques

Περιεχόμενα

- Εισαγωγή
- Μετρικές Δομές Οργάνωσης των Δεδομένων
  - Μετρικοί Χώροι
  - Ερωτήματα Ομοιότητας
  - M-δένδρο
  - Slim-δένδρο
- Προσεγγιστικά Ερωτήματα με χρήση M-trees
- Σχήμα Ταξινόμησης προσεγγιστικών ερωτημάτων

Εισαγωγή

- Αυξανόμενη χρήση της τεχνολογίας -> Αύξηση του όγκου των δεδομένων σε Β.Δ.
- Ζητήματα Απόδοσης σε ερωτήματα αναζήτησης σε Β.Δ.
- Δενδρικές Δομές (Β, Β+ δένδρα) για την αποδοτική δεικτοδότηση του συνόλου των εγγραφών
- Αλλαγές της μορφής των δεδομένων (πολυμεσικών, γεωγραφικών) -> Αλλαγές στις δομές των δεδομένων
  - Επέκταση υπάρχουσων
  - Δημιουργία νέων (R, R*, M, Slim δένδρα)

Μετρικοί Χώροι

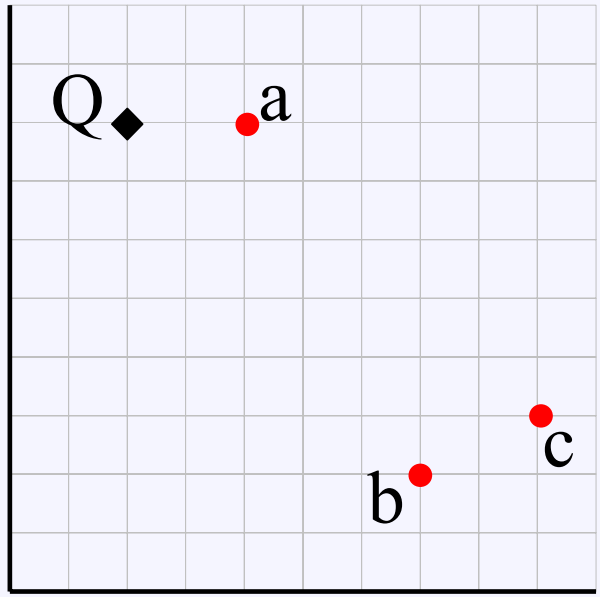
- Οι Μετρικές Δομές Οργάνωσης των Δεδομένων επιτυγχάνουν την αποδοτική δεικτοδότηση αντικειμένων που βρίσκονται σε μετρικούς χώρους
- Ένας μετρικός χώρος είναι ένα ζεύγος M=(D,d) όπου
  - D είναι το πεδίο από όπου παίρνουν τιμές τα χαρακτηριστικά των αντικειμένων
  - d είναι μία συνάρτηση απόστασης μεταξύ των αντικειμένων και πρέπει να πληρεί τις 3 ακόλουθες ιδιότητες
    - Συμμετρία, $d(O_x, O_y) = d(O_y, O_x)$
    - Θετικότητα, $d(O_x, O_y) > 0$ $(O_x \neq O_y)$ και $d(O_x, O_x) = 0$
    - Τριγωνική Ανισότητα, $d(O_x, O_y) \leq d(O_x, O_z) + d(O_z, O_y)$

Τριγωνική Ανισότητα

- Έστω ότι ψάχνουμε το κοντινότερο σημείο σε μία βάση 3 αντικειμένων σε ένα ερώτημα Q
- Έστω επίσης ότι η τριγωνική ανισότητα ισχύει και ότι οι αποστάσεις μεταξύ των αντικειμένων στη βάση έχουν υπολογιστεί
- Έστω ότι το ***a*** υπολογίζουμε ότι απέχει 2 μονάδες από το Q (και γίνεται το best so far)
- Υπολογίζοντας το d(Q,***b***) = 7.81 και συμπεραίνουμε πως δε χρειάζεται να υπολογιστεί το d(Q,***c***) διότι:

$$d(Q,\boldsymbol{b}) \le d(Q,\boldsymbol{c}) + d(\boldsymbol{b},\boldsymbol{c})$$
$$d(Q,\boldsymbol{b}) - d(b,\boldsymbol{c}) \le d(Q,\boldsymbol{c})$$
$$7.81 - 2.30 \le d(Q,\boldsymbol{c})$$
$$5.51 \le d(Q,\boldsymbol{c})$$

και έτσι το ***c*** απέχει τουλάχιστον 5.51 μονάδες από το Q ενώ το best so far απέχει μόλις 2



	a	b	c
a		6.70	7.07
b			2.30
c			

Ερωτήματα Ομοιότητας

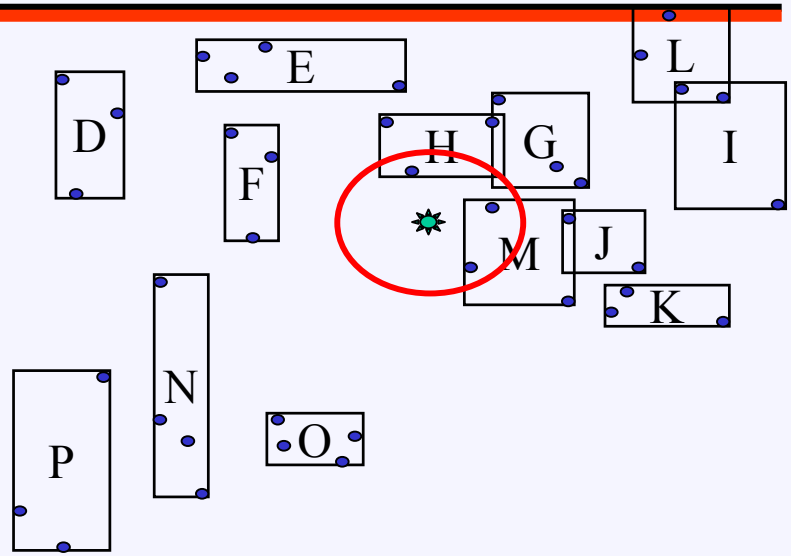
- Ερωτήματα Περιοχής
  - Δοθέντος ενός αντικειμένου ερωτήματος $Q \in D$ και μίας μέγιστης ακτίνας απόστασης $r(Q)$, το ερώτημα περιοχής range(Q, $r(Q)$) επιλέγει όλα τα αντικείμενα O_j έτσι ώστε $d(O_j, Q) \leq r(Q)$
- Ερωτήματα k πλησιέστερων γειτόνων
  - Δοθέντος ενός αντικειμένου ερωτήματος $Q \in D$ και ενός θετικού ακέραιου $k \geq 1$, το ερώτημα k πλησιέστερων γειτόνων NN(Q, k) επιλέγει τα k αντικείμενα που έχουν τη μικρότερη απόσταση από το Q.

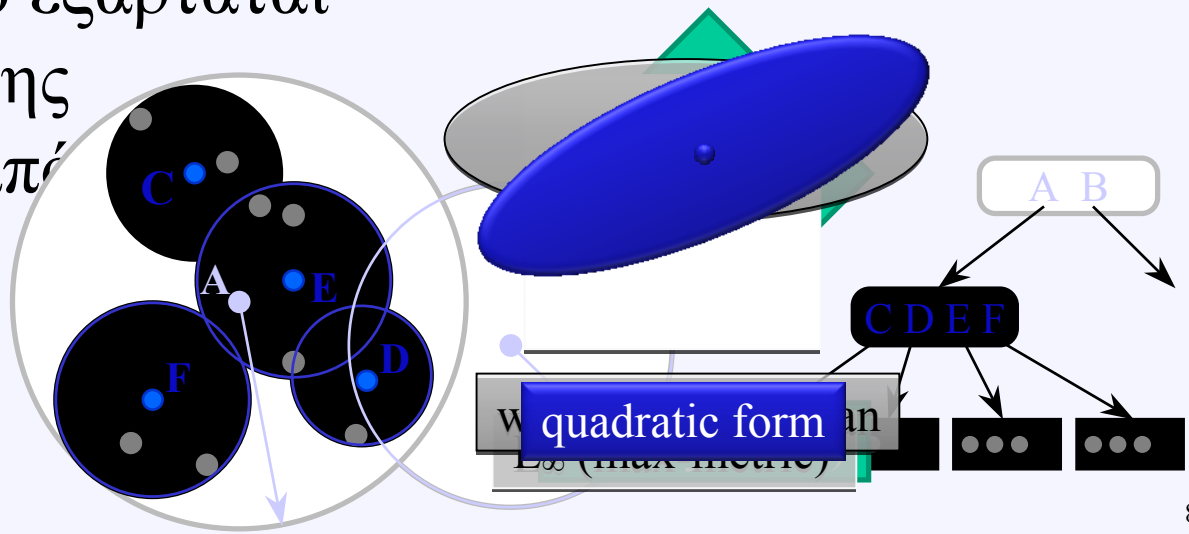
M-tree

- Είναι ένα ισοζυγισμένο δένδρο ικανό να χειριστεί δυναμικά σύνολα δεδομένων
- Βασίζεται στις σχετικές αποστάσεις μεταξύ των αντικειμένων για τον καθορισμό των κόμβων αποθήκευσης τους
- Είναι τελείως παραμετροποιήσιμο ως προς τη συνάρτηση απόστασης d (black-box)
- Η βελτιστοποίηση της απόδοσης του επικεντρώνεται
  - Σε ζητήματα CPU (υπολογισμοί απόστασης)
  - Σε θέματα I/O (πρόσβαση στο δίσκο)

M-tree

- Μορφή των δεικτοδοτούμενων περιοχών
  - Στο R-δένδρο είναι κάπως έτσι
  - Στο M-δένδρο εξαρτάται από το είδος της συνάρτησης από

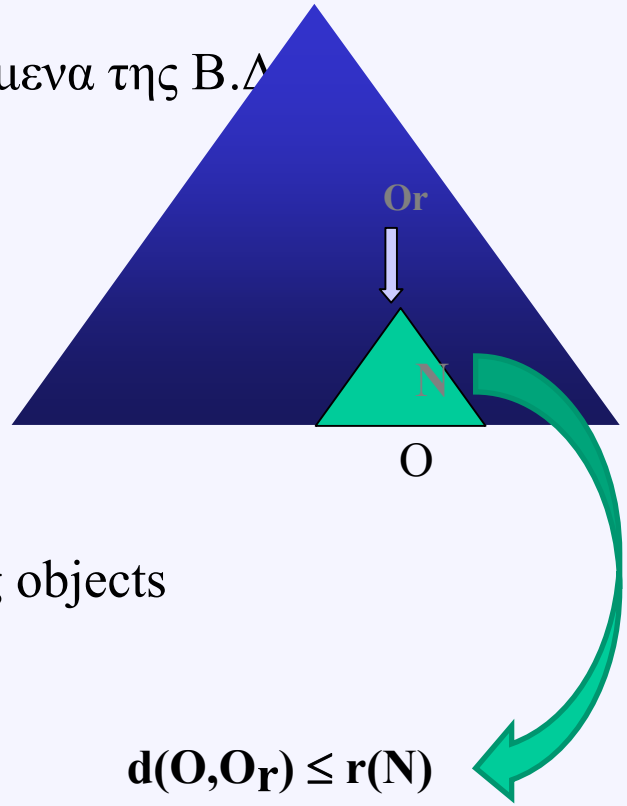




M-tree

- Δομή των κόμβων
  - Οι κόμβοι φύλλα αποθηκεύουν όλα τα αντικείμενα της Β.Δ

O_j	Τα χαρακτηριστικά του αντικειμένου O_j
$oid(O_j)$	Δείκτης προς το αντικείμενο στη Βάση Δεδομένων
$d(O_j, P(O_j))$	Απόσταση του O_j από τον πατρικό κόμβο



  - Οι εσωτερικοί κόμβοι αποθηκεύουν τα routing objects

O_r	Τα χαρακτηριστικά του routing αντικειμένου O_r
$ptr(T(O_r))$	Δείκτης στη ρίζα του υποδένδρου (covering tree) $T(O_r)$
$r(O_r)$	Ακτίνα κάλυψης του O_r
$d(O_r, P(O_r))$	Απόσταση του O_r από τον πατρικό κόμβο

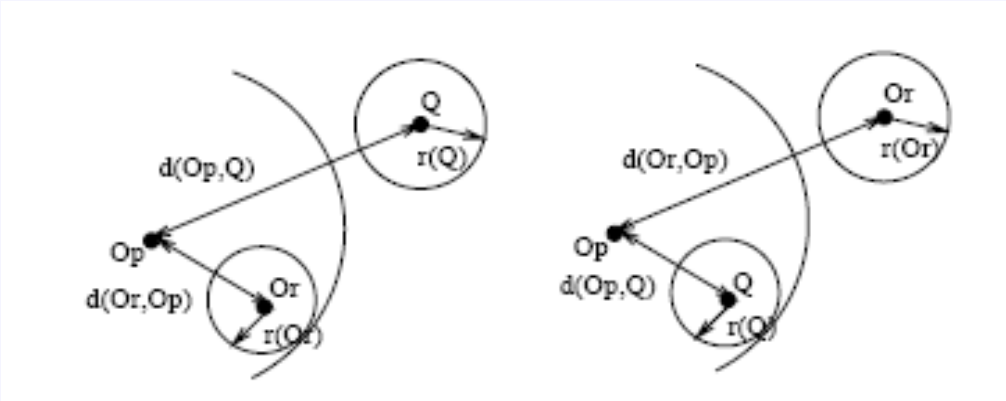
$$\mathbf{d(O,O_r) \leq r(N)}$$

M-tree

Εφαρμόζονται δύο βασικοί κανόνες για το κλάδεμα κόμβων που δεν μπορούν να συμμετέχουν στην απάντηση ενός ερωτήματος range(Q, $r(Q)$)

- Αν $d(O_r, Q) > r(Q) + r(O_r)$, τότε για κάθε αντικείμενο O_j στο $T(O_r)$ ισχύει: $d(O_j, Q) > r(Q)$. Συνεπώς το $T(O_r)$ μπορεί να κλαδευτεί με ασφάλεια
- Αν $|d(O_p, Q) – d(O_r, O_p)| > r(Q) + r(O_r)$, τότε $d(O_r, Q) > r(Q) + r(O_r)$ και ο κόμβος O_r κλαδεύεται

- Μείωση των υπολογισμών αποστάσεων κατά 40%



M-tree

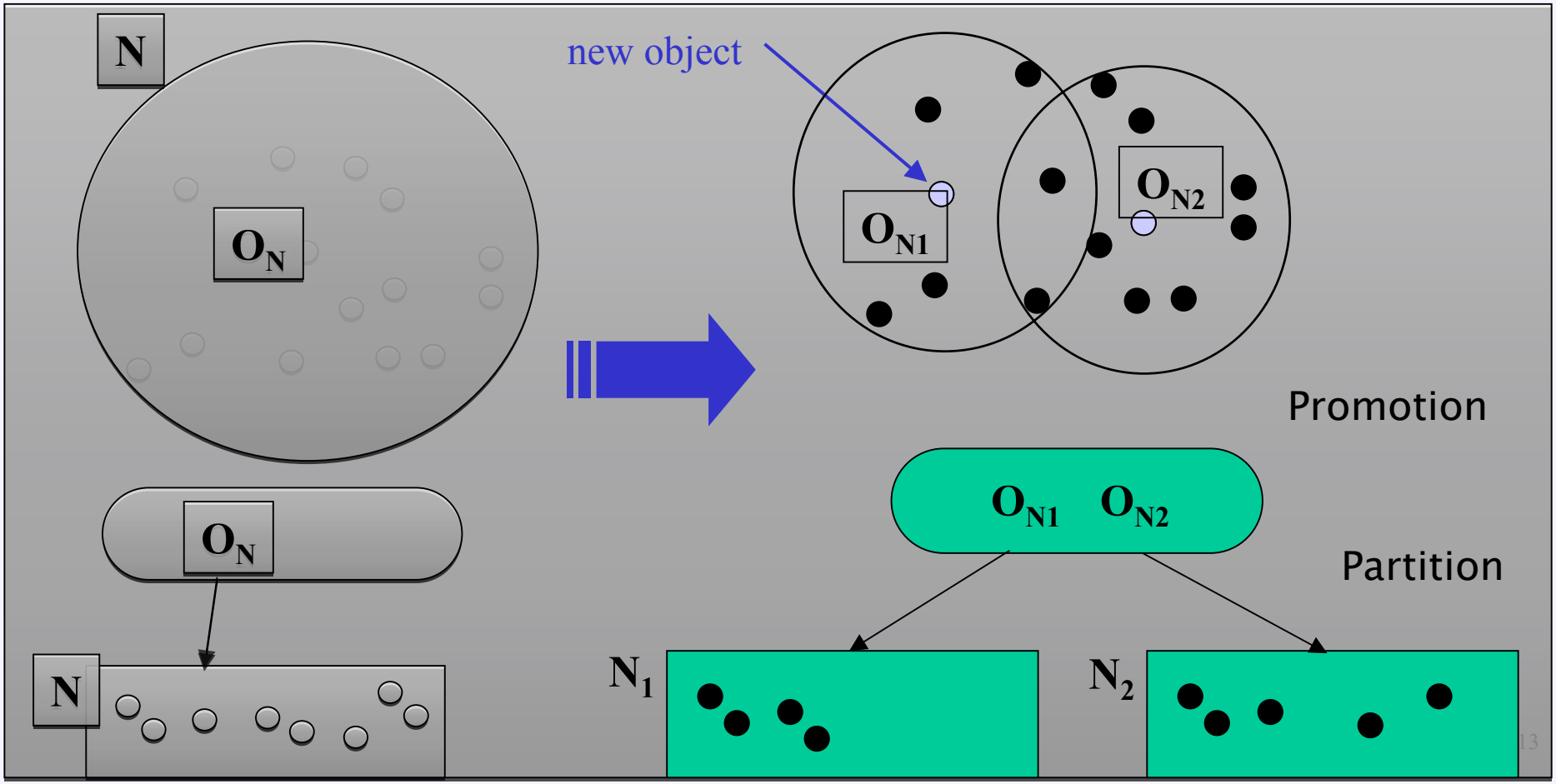
- Εκτέλεση Ερωτημάτων
  - Για την απάντηση ερωτημάτων kNN χρησιμοποιείται μία τεχνική branch-and-bound
  - Κάνει χρήση 2 καθολικών δομών
    - Μία ουρά προτεραιότητας PR
      - περιέχει δείκτες προς ενεργά υπο-δένδρα
      - και ένα ελάχιστο όριο $d_{min}(T(O_r)) = \max\{d(O_r, Q) - r(O_r), 0\}$ το οποίο χρησιμοποιείται ως ευριστικό κριτήριο για την επιλογή του επόμενου κόμβου για εξέταση (διαλέγεται εκείνος ο κόμβος που έχει το μικρότερο ελάχιστο όριο)
    - Έναν πίνακα k θέσεων NN (στο τέλος θα περιέχει το αποτέλεσμα)

M-tree

- Κατασκευή του M-δένδρου
  - Εισαγωγή νέων αντικειμένων σε οποιαδήποτε χρονική στιγμή, το M-δένδρο είναι δυναμικό
  - Πραγματοποιείται αναδρομική κατάβαση του δένδρου για να βρεθεί το πιο κατάλληλο φύλλο
    - Επιλέγεται κάθε φορά ο κόμβος που δε θα προκαλέσει αύξηση της ακτίνας κάλυψης
    - Αν υπάρχουν περισσότεροι του ενός τέτοιοι κόμβοι επιλέγεται εκείνος του οποιου το O_r είναι πλησιέστερα στο O_n
    - Αν δεν υπάρχει τέτοιος κόμβος επιλέγεται εκείνος που ελαχιστοποιεί την αύξηση της ακτίνας κάλυψης $d(O_r, O_n) – r(O_r)$

Διαχείριση Διάσπασης (split)

Εφαρμόζεται κατά την εισαγωγή ενός νέου αντικειμένου σε γεμάτο κόμβο



Διαχείριση Διάσπασης (split)

- Πολιτική Διάσπασης: καθορίζεται από τις υλοποιήσεις των μεθόδων του promoting και του partioning
- Η βέλτιστη πολιτική θα έπρεπε να πετυχαίνει
  - Τη μικρότερη δυνατή αλληλοεπικάλυψη μεταξύ των κόμβων (λιγότερα μονοπάτια προσπελαύνονται)
  - Τη μείωση του μεγέθους των κόμβων (μειώνεται το μέγεθος του δεικτοδοτημένου νεκρού χώρου)

Προαγωγή (promotion)

- Promotion
  - Δοθέντος ενός συνόλου αντικειμένων N, ο καθορισμός δύο αντικειμένων για «ανέβασμα» και αποθήκευση στον πατρικό κόμβο
    1. Το ένα από τα δύο promoted αντικείμενα που θα γίνουν είναι το αντικείμενο που περιέχεται στον πατρικό κόμβο
    2. Ο αλγόριθμος m_RAD κάνει promote τα αντικείμενα που ελαχιστοποιούν το άθροισμα των ακτινών κάλυψης $r(O_{p1}) + r(O_{p2})$ (πιο πολύπλοκος)
    3. Ο mM_RAD ελαχιστοποιεί την μέγιστη τιμή των δύο ακτίνων κάλυψης
    4. Η μέθοδος M_LB_DIST χρησιμοποιεί μόνο τις ήδη υπολογισμένες αποστάσεις
    5. Η μέθοδος RANDOM επιλέγει τυχαία τα 2 αντικείμενα
    6. Η μέθοδος SAMPLING διαλέγει τυχαία αντικείμενα, υπολογίζει το άθροισμα των ακτινών κάλυψης των αντικειμένων και επιλέγει αυτά που το ελαχιστοποιούν (η τυχαία δειγματοληψία γίνεται συνήθως μόνο για το δεύτερο αντικείμενο)

Διαμέριση (partition)

- Partition
  - Δοθέντος ενός συνόλου εγγραφών N και δύο routing αντικειμένων, το μοίρασμα των εγγραφών σε δύο υποσύνολα του N, N_1 και N_2

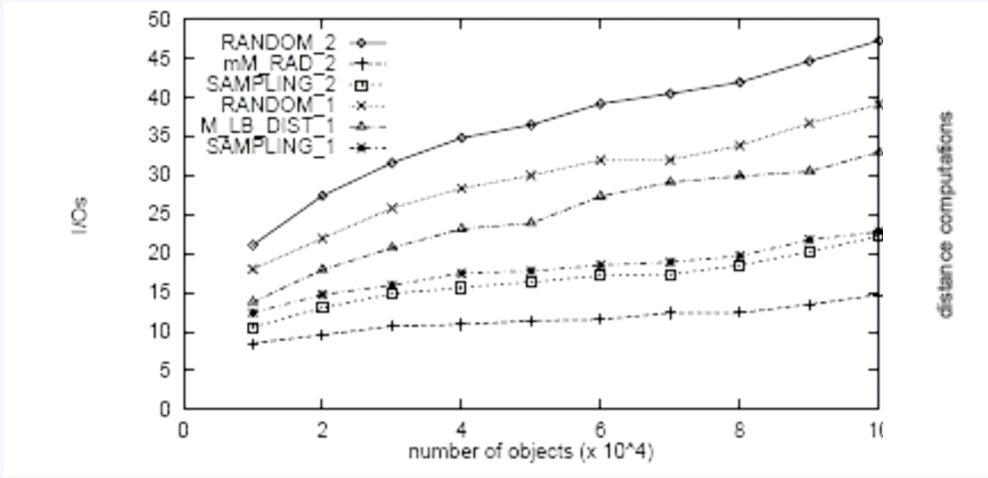
1. Σύμφωνα με τον αλγόριθμο Generalized Hyperplane, κάθε αντικείμενο $O_j \in N$ αντιστοιχίζεται στο κοντινότερο routing αντικείμενο. Αν $d(O_j, O_{p1}) \leq d(O_j, O_{p2})$, το O_j ανατίθεται στον κόμβο N_1 αλλιώς στον κόμβο N_2
2. Balanced: Αρχικά υπολογίζονται οι αποστάσεις $d(O_j, O_{p1})$ και $d(O_j, O_{p2})$ για όλα τα $O_j \in N$. Έπειτα, τα επόμενα βήματα εκτελούνται μέχρις ότου να αδειάσει το N.
   - » Τοποθέτηση στο N_1 του κοντινότερου γείτονα του O_{p1} και διαγραφή του από το N
   - » Τοποθέτηση στο N_2 του κοντινότερου γείτονα του O_{p2} και διαγραφή του από το N

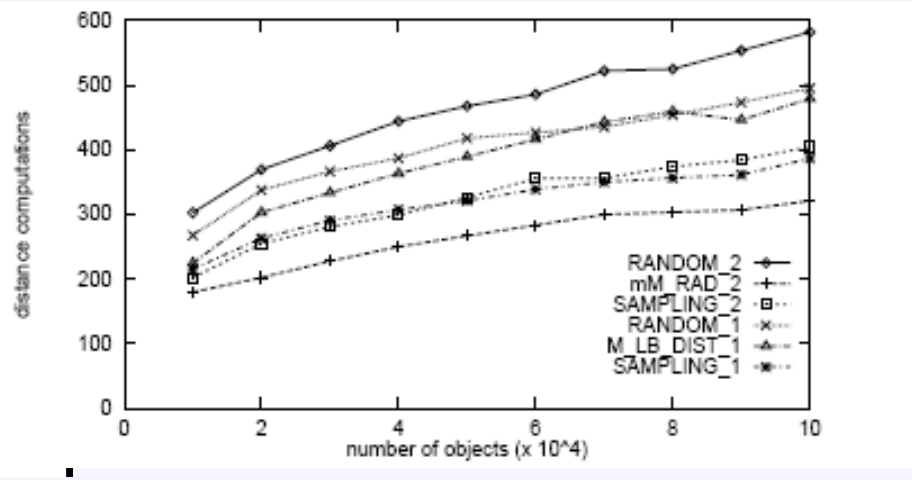
Η μέθοδος αυτή καταλήγει σε πολύ καλά ισοζυγισμένα δένδρα όπου όλοι οι κόμβοι έχουν περίπου τον ίδιο αριθμό αντικειμένων, απαιτεί όμως τον υπολογισμό μεγάλου αριθμού αποστάσεων.

Πειραματικά Αποτελέσματα

Το M-δένδρο κλιμακώνεται καλά στο μέγεθος του συνόλου δεδομένων.

Τόσο το κόστος I/O όσο και το κόστος CPU αυξάνουν λογαριθμικά.





Σύνοψη M-tree

- Τέσσερα στοιχεία καινοτομίας
  - Είναι μία paged, ισοζυγισμένη (balanced) και δυναμική δευτερεύουσα δομή δεδομένων για τη δεικτοδότηση συνόλων δεδομένων που ανήκουν σε μετρικούς χώρους
  - Ερωτήματα περιοχής και ερωτήματα πλησιέστερων γειτόνων μπορούν να εκτελεστούν με βάση ένα αντικείμενο ερωτήματος
  - Η εκτέλεση των ερωτημάτων έχει βελτιστοποιηθεί έτσι ώστε να μειωθούν τόσο ο αριθμός των σελίδων που διαβάζονται από το δίσκο, όσο και ο αριθμός των υπολογισμών των αποστάσεων που πρέπει να πραγματοποιηθούν
  - Είναι κατάλληλο για αντικείμενα που χαρακτηρίζονται από μεγάλο αριθμό διαστάσεων, έχουν δηλ. πολλά features

Slim-tree

- Αποτελεί και αυτό μία μετρική δομή οργάνωσης δεδομένων που βρίσκονται σε μετρικούς χώρους
- Μοιράζεται τη βασική δομή άλλων μετρικών δένδρων (M-δένδρο), διαφέρει όμως στα εξής:
  1. Ένας νέος αλγόριθμος διάσπασης που βασίζεται στο ελάχιστο ζευγνύον δένδρο (minimum spanning tree – MST) εισάγεται που εκτελείται πιο γρήγορα χωρίς να μειώνεται η απόδοση της ακρίβειας
  2. Ένας νέος αλγόριθμος χρησιμοποιείται για την εισαγωγή νέων αντικειμένων στους πιο κατάλληλους κόμβους
  3. Εκτελείται τέλος ο αλγόριθμος Slim-down ως ένα post-processing βήμα έτσι ώστε το δένδρο να γίνει tighter και συνεπώς γρηγορότερο κατά την αναζήτηση.

     Ο αλγόριθμος αυτός χρησιμοποιεί δύο παράγοντες (fat-factor, bloat-factor) για τη μέτρηση του βαθμού επικάλυψης μεταξύ των κόμβων

Slim-tree

Σε αντιστοιχία με το M-δένδρο οι κόμβοι διακρίνονται

Κόμβους φύλλα

Oid_i	Identifier του αντικειμένου O_i
$D(O_i, Rep(O_i))$	Απόσταση μεταξύ του αντικειμένου O_i και του αντιπροσωπευτικού αντικειμένου του κόμβου $Rep(O_i)$
O_i	Το αντικειμένου O_i

Εσωτερικούς κόμβους δεικτοδότησης

O_i	Το αντιπροσωπευτικό αντικείμενο του υπο-δένδρου του κόμβου
$Radius_i$	Η ακτίνα κάλυψης της περιοχής που καλύπτει ο κόμβος
$D(O_i, Rep(O_i))$	Απόσταση μεταξύ του αντικειμένου O_i και του αντιπροσωπευτικού αντικειμένου του κόμβου $Rep(O_i)$
$Ptr(TO_i)$	Δείκτης προς τη ρίζα του υπο-δένδρου
$NEntries(Ptr(TO_i))$	Αριθμός των εγγραφών στον κόμβου που δείχνει ο $Ptr(TO_i)$

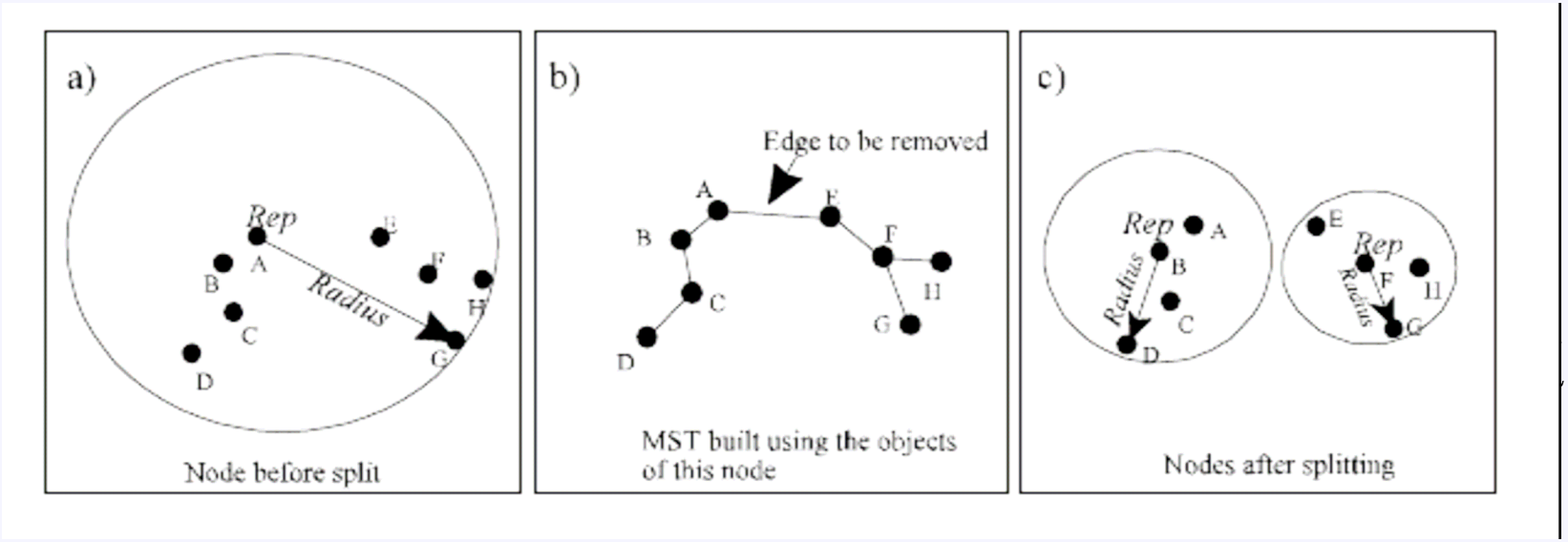
Slim-tree

Χτίσιμο του Slim-δένδρου

- Εντοπισμός ενός κόμβου που να καλύπτει το νέο αντικείμενο ξεκινώντας από τη ρίζα
- Αν δεν βρεθεί επιλέγεται εκείνος που το κέντρο του απέχει λιγότερο από το νέο αντικείμενο
- Αν υπάρχουν περισσότεροι του ενός κόμβοι εκτελείται ο αλγόριθμος ChooseSubtree
- Η παραπάνω διαδικασία εκτελείται σε όλα τα επίπεδα του δένδρου
- Επιλογές αλγορίθμου ChooseSubtree
  - random: επιλέγεται τυχαία ο κόμβος
  - mindist: επιλέγεται ο κόμβος που απέχει λιγότερο από το νέο αντικείμενο και το κέντρο του κόμβου
  - minoccup: επιλέγεται ο κόμβος που έχει την ελάχιστη εγκατοίκηση (occupancy)

Slim-tree

Διάσπαση κόμβων



– MST: Το ελάχιστο ζευγνύον δένδρο των αντικειμένων δημιουργείται, διαγράφεται η μεγαλύτερη ακμή (μεγαλύτερη απόσταση) και έτσι δημιουργούνται δύο ομάδες. Από κάθε ομάδα επιλέγεται ως αντιπροσωπευτικό αντικείμενο εκείνο που έχει την ελάχιστη μέγιστη απόσταση από τα υπόλοιπα αντικείμενα

Slim-tree

- Βελτιστοποίηση Επικάλυψης
  - Στους διανυσματικούς χώρους η επικάλυψη 2 εγγραφών αναφέρεται στην τομή των δύο περιοχών επικάλυψης
  - Στους μετρικούς χώρους οι περιοχές δεν είναι γνωστές
  - Συνεπώς επικάλυψη μεταξύ δύο εγγραφών I1 και I2 ορίζεται ο λόγος του αριθμού των αντικειμένων στα αντίστοιχα υπο-δένδρα που καλύπτονται και από τις 2 περιοχές προς τον αριθμό των αντικειμένων στα 2 υπο-δένδρα
  - Ένα δένδρο χαρακτηρίζεται από 2 αριθμητικούς παράγοντες
    - Fat-factor
    - Bloat-factor

Slim-tree

- Βελτιστοποίηση Επικάλυψης
  - Fat-factor
    - μετράει πόσο καλό είναι ένα δένδρο σε σχέση με το ποσοστό των επικαλύψεων ασχέτως αν απαιτούνται περισσότερες προσβάσεις στο δίσκο λόγω μικρότερης εγκατοίκησης των κόμβων
  - Bloat-factor
    - λαμβάνει υπόψη του τόσο το ποσοστό επικάλυψης όσο και την αποδοτική εγκατοίκηση των κόμβων
    - χρησιμοποιείται για τη σύγκριση διαφορετικών δένδρων που περιέχουν τα ίδια δεδομένα
  - Οι δύο παράγοντες κυμαίνονται από 0 μέχρι 1
    - 0 στη βέλτιστη περίπτωση
    - 1 στη χειρότερη περίπτωση

Slim-tree

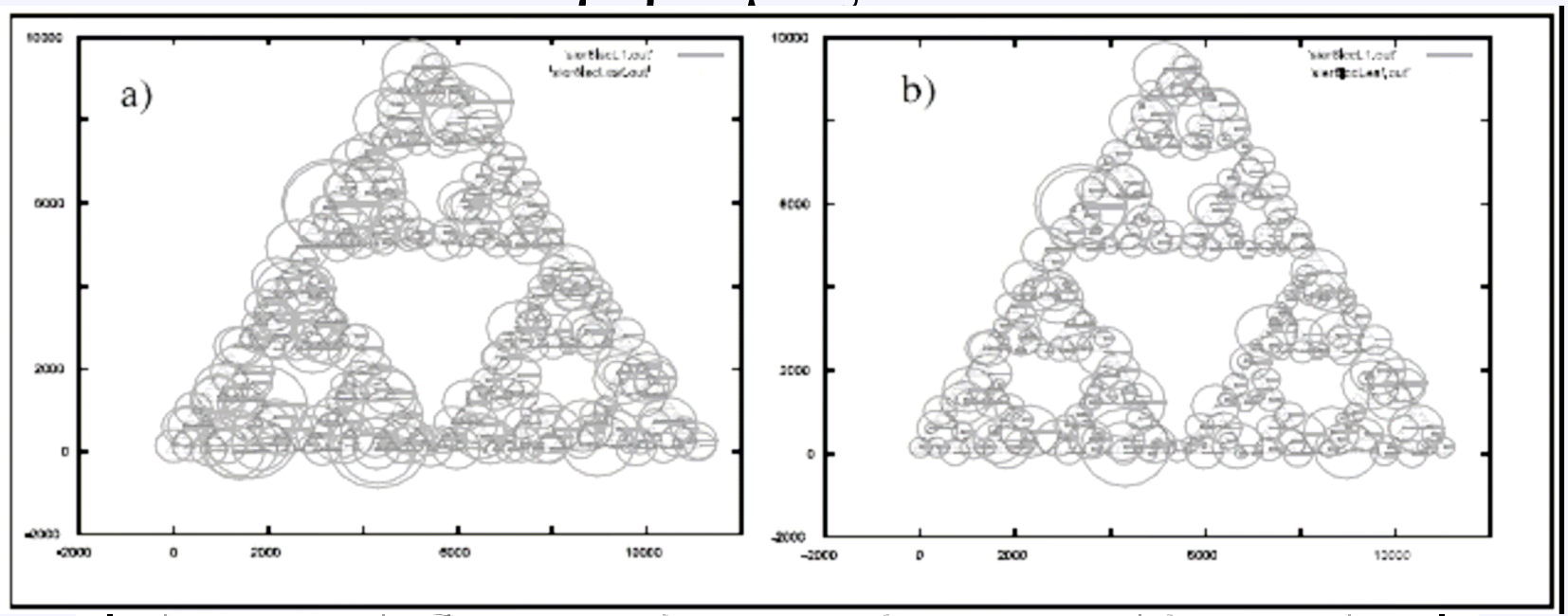
Ο Slim-down αλγόριθμος παράγει ένα καλύτερο (tighter) δένδρο

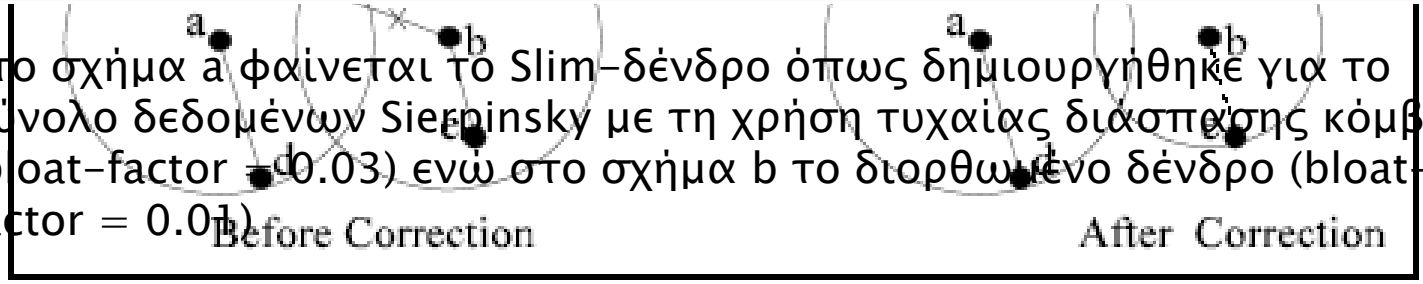
- Μειώνοντας το ποσοστό επικάλυψης των κόμβων
- Μειώνοντας τον αριθμό των κόμβων του δένδρου

1. Για κάθε κόμβο i σε ένα δοσμένο επίπεδο του δένδρου, βρίσκουμε το αντικείμενο c που απέχει περισσότερο από τον αντιπρόσωπο b
2. Βρίσκουμε έναν αδελφό κόμβο του i, έστω j, που επίσης καλύπτει το αντικείμενο c. Αν βρεθεί τέτοιος j που να μην είναι γεμάτος, βγάζουμε από τον κόμβο i το αντικείμενο c και το τοποθετούμε στον κόμβο j. Διορθώνουμε τέλος την ακτίνα του κόμβου i
3. Τα βήματα 1 και 2 εκτελούνται ακολουθιακά σε όλους τους κόμβους σε ένα δοσμένο επίπεδο του δένδρου. Αν μετά από ένα συνολικό γύρο (full round) των 2 πρώτων βημάτων, ένα αντικείμενο μετακινείται ακόμη από έναν κόμβο σε έναν άλλον, τότε πρέπει να εκτελεστεί ακόμα ένας συνολικός γύρος των βημάτων 1 και 2

Slim-tree

- Ο Slim-down αλγόριθμος

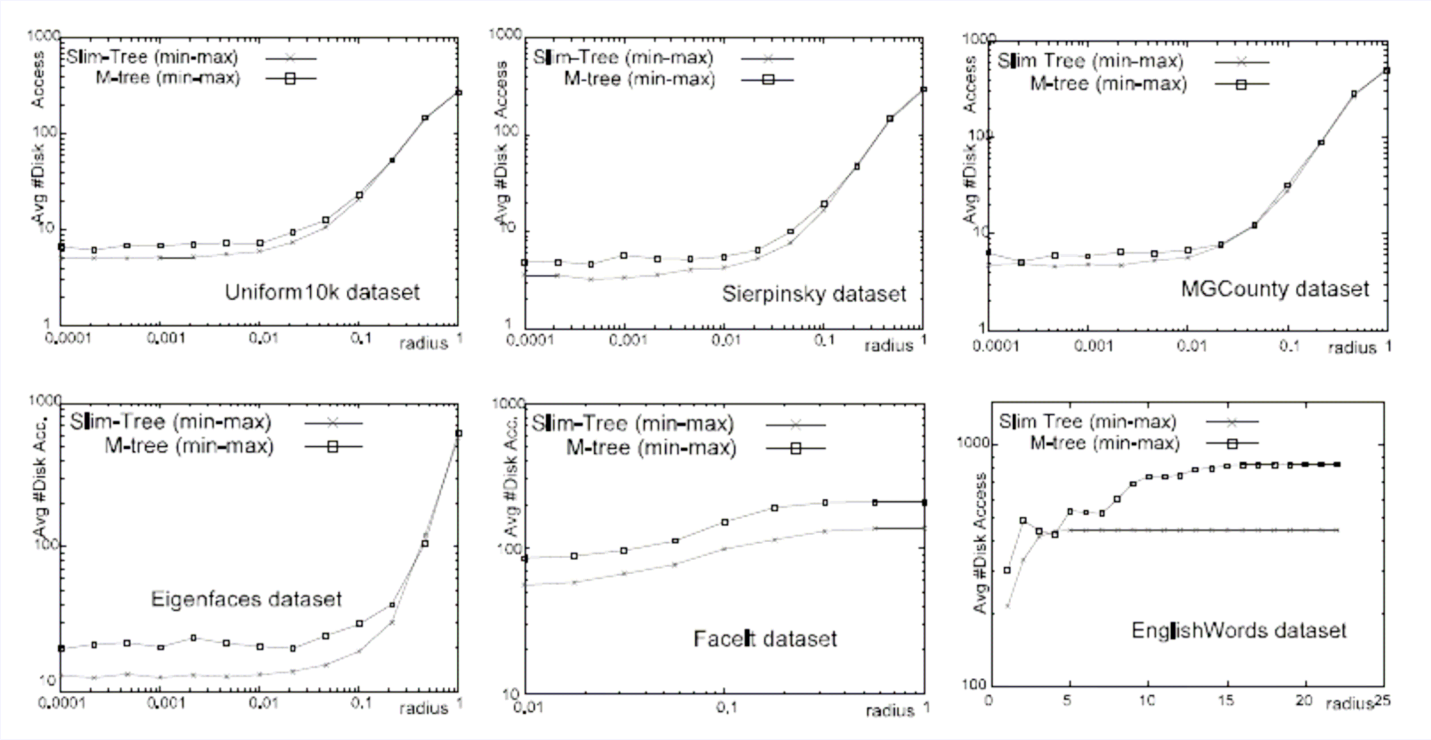




Στο σχήμα a φαίνεται το Slim-δένδρο όπως δημιουργήθηκε για το σύνολο δεδομένων Sierpinsky με τη χρήση τυχαίας διασπάσης κόμβων (bloat-factor = 0.03) ενώ στο σχήμα b το διορθωμένο δένδρο (bloat-factor = 0.01).

Slim-tree

- Πειραματικά Αποτελέσματα
  - Παρατηρείται πως το Slim-δένδρο είναι καλύτερο του M-δένδρου λόγω της μεγαλύτερης εγκατοίκησης των κόμβων



Σύνοψη Slim-tree

Τα κυριότερα στοιχεία του Slim-δένδρου είναι τα εξής:

- Ένας νέος ChooseSubtree αλγόριθμος που κατευθύνει ένα νέο αντικείμενο από έναν κόμβο σε αυτόν που έχει τη μικρότερη εγκατοίκηση σε περίπτωση που υπάρχει δυνατότητα επιλογής
- Ένας νέος αλγόριθμος διάσπασης υπερχειλισμένων κόμβων που βασίζεται στο ελάχιστο ζευγνύον δένδρο (MST)
- Ένας νέος αλγόριθμος "Slim-down" ο οποίος έχει τη δυνατότητα να βελτιώσει την απόδοση ενός δένδρου μειώνοντας την επικάλυψη, αυξάνοντας την εγκατοίκηση και μειώνοντας τον αριθμό των κόμβων
- Δύο αριθμητικοί παράγοντες χρησιμοποιούνται για τον χαρακτηρισμό της απόδοσης του δένδρου, ο fat-factor και ο bloat-factor

Προσεγγιστικά Ερωτήματα

- Προσεγγιστικά Ερωτήματα
  - Ανάγκη για αυξημένη επίδοση
  - Γρήγορη απόκριση
  - Ποιότητα στα αποτελέσματα
- Γιατί;
  - Μεγάλος όγκος δεδομένων
  - Η exact αναζήτηση υπολογιστικά/χρονικά ασύμφορη
  - Αναγκαιότητα για γρήγορη λήψη σωστών αποφάσεων

Προσεγγιστικά Ερωτήματα

- Πειράματα
  - Χρήση M-trees
  - Χρήση k-NN μεθόδου (10 k-NN)
  - Χρήση Ευκλείδειας απόστασης
  - Αρχεία πειραμάτων
    - CHV
      - 10.000 διανύσματα 45 διαστάσεων
      - Πραγματικά δεδομένα
    - UV
      - Συνθετικά δεδομένα
      - Διανύσματα που κατανένομονται ομοιόμορφα
    - CV
      - Συνθετικά δεδομένα
      - Χρήση cluster

Προσεγγιστικά Ερωτήματα

- Μέτρα απόδοσης
  - Improvement in efficiency (IE), που σχετίζει το κόστος της ακριβής αναζήτησης με αυτό της προσεγγιστικής. $IE = \frac{cost(O_N^k)}{cost(O_A^k)}$,
  - Precision of approximation (P)

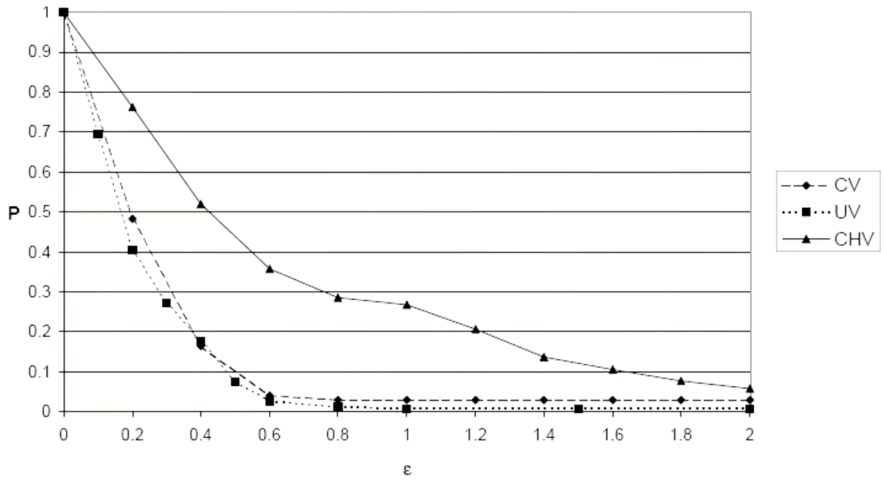
$$P = \frac{\sum_{i=1}^{k} P_i}{k} = \frac{\sum_{i=1}^{k} \frac{i}{\#range(Q, d(Q, O_A^i))}}{k}.$$

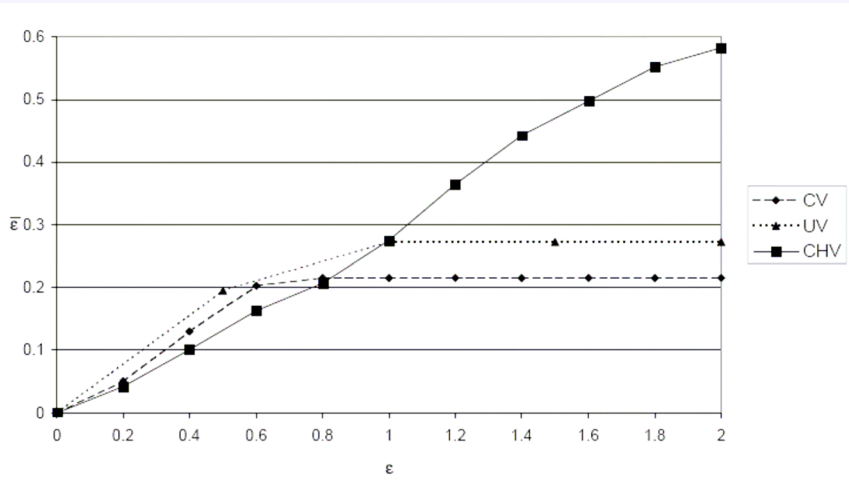
  - Relative distance error (ε) $\overline{\epsilon} = \frac{\sum_{i=1}^{k} \overline{\epsilon_i}}{k} = \frac{\sum_{i=1}^{k} \frac{d(Q, O_A^i)}{d(Q, O_N^i)}}{k} - 1.$

    - όταν το ε = 0 τότε η προσεγγιστική μέθοδος δίνει τα ίδια αποτελέσματα με την ακριβή/εξαντλητική μέθοδο

Προσεγγιστικά Ερωτήματα

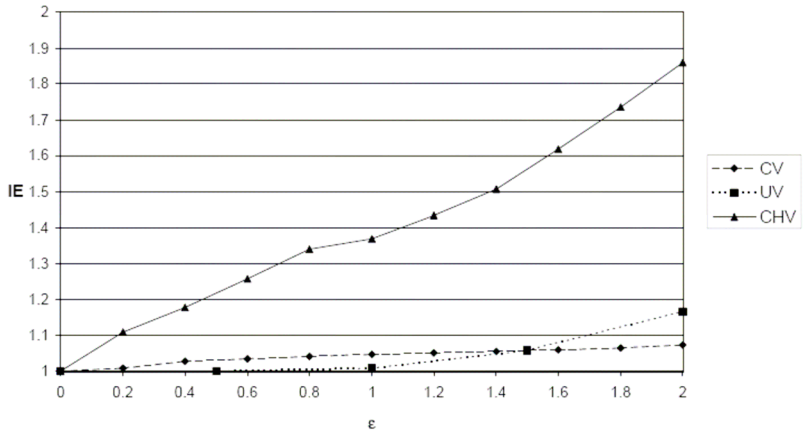
- Approximation through relative distance errors





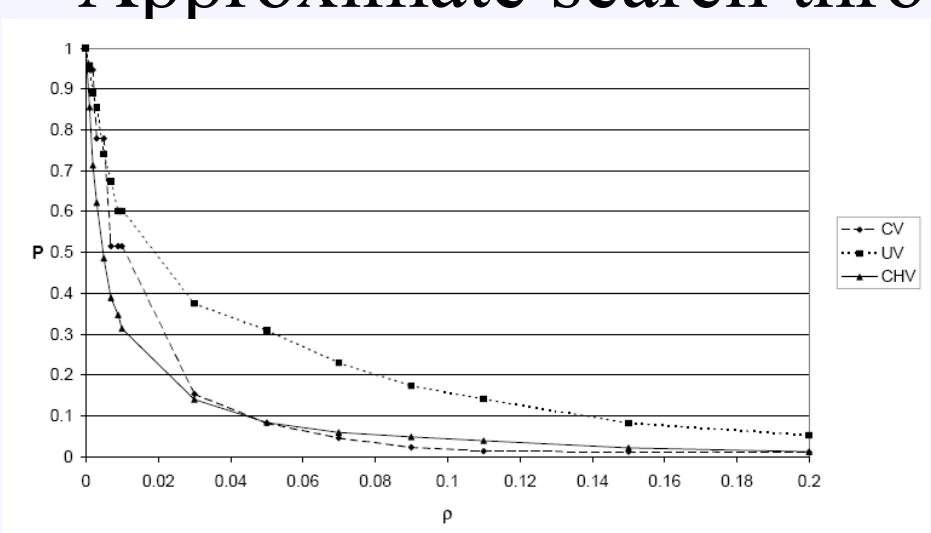
Σημαντικό ρόλο στην απόδοση διαδραματίζει η πυκνότητα του χώρου αναζήτησης και όχι η κατανομή των δεδομένων.

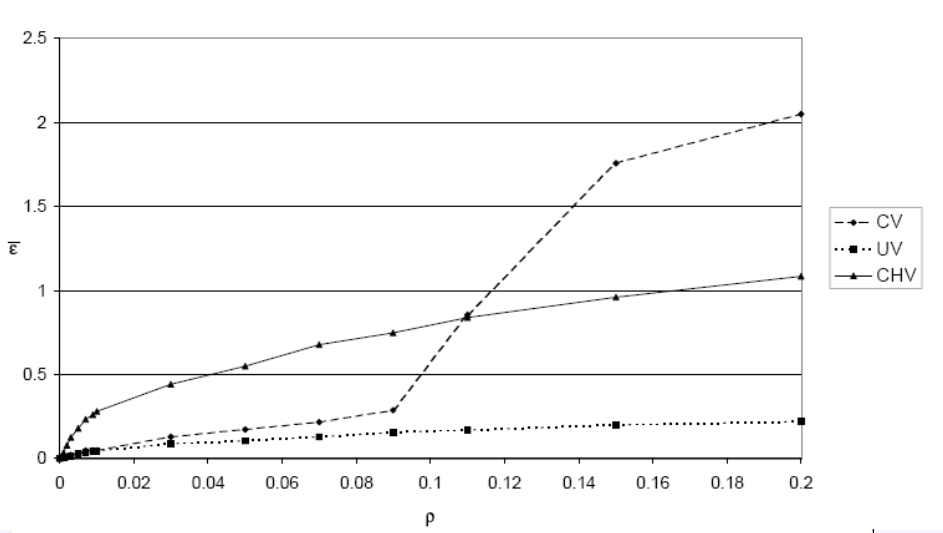
Λιγότερο πυκνοί χώροι παρέχουν καλύτερη ακρίβεια και υψηλότερη βελτίωση της απόδοσης αλλά μπορεί να δώσουν και προσεγγίσεις με μεγαλύτερο σχετικό σφάλμα.



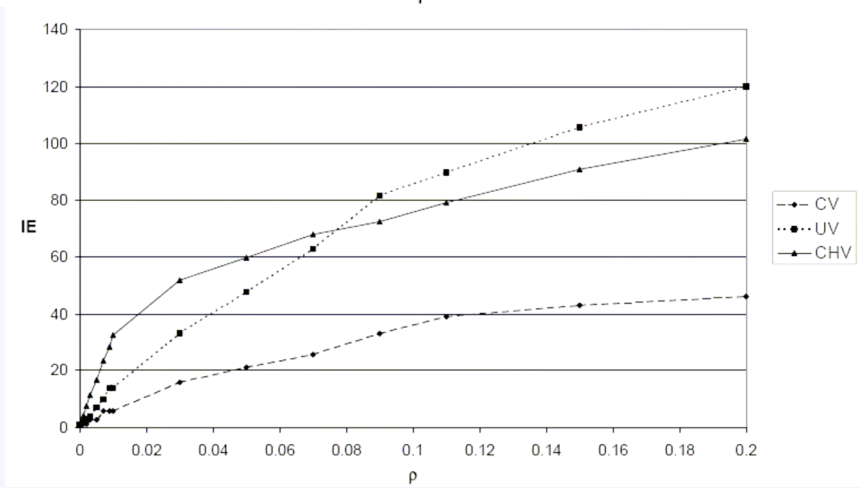
Προσεγγιστικά Ερωτήματα

- Approximate search through distance distributions



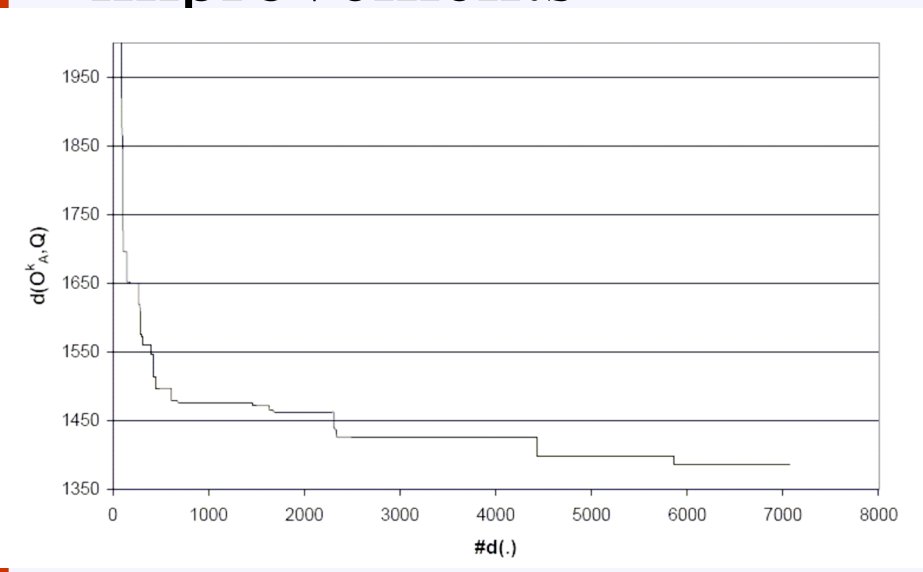


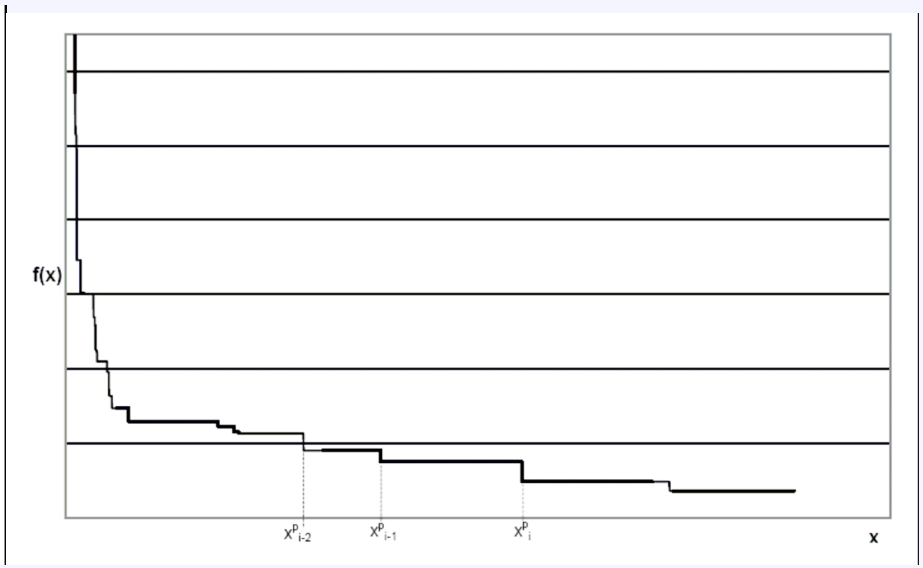
Η μέθοδος αυτή εκμεταλλεύεται χαρακτηριστικά της κατανομής της απόστασης για να ελέγξει την αναζήτηηση προσεγγιστικής ομοιότητας.
Με αυτή τη μέθοδο μπορούμε να βρούμε τιμές του ρ για τις οποίες να πετύχουμε υψηλή απόδοση, υψηλή ακρίβεια και μικρό σχετικό λάθος (πχ για ρ=0,01).



Προσεγγιστικά Ερωτήματα

- Approximation through the slowdown of distance improvements

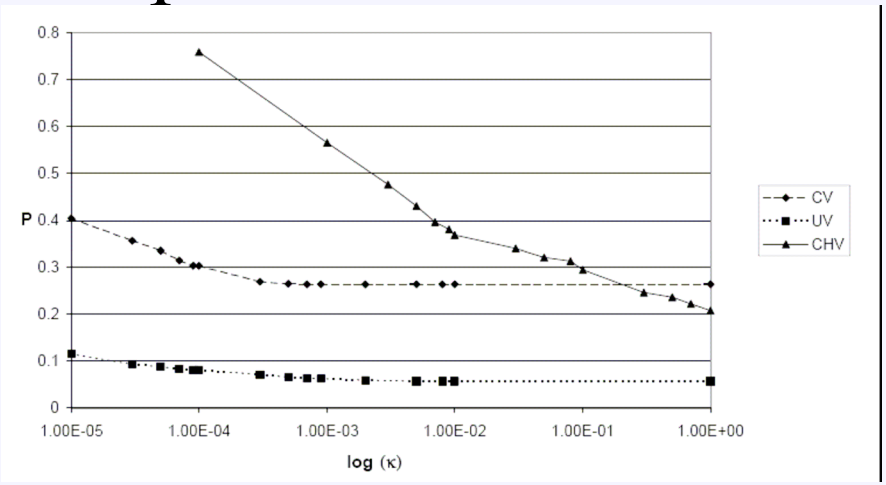


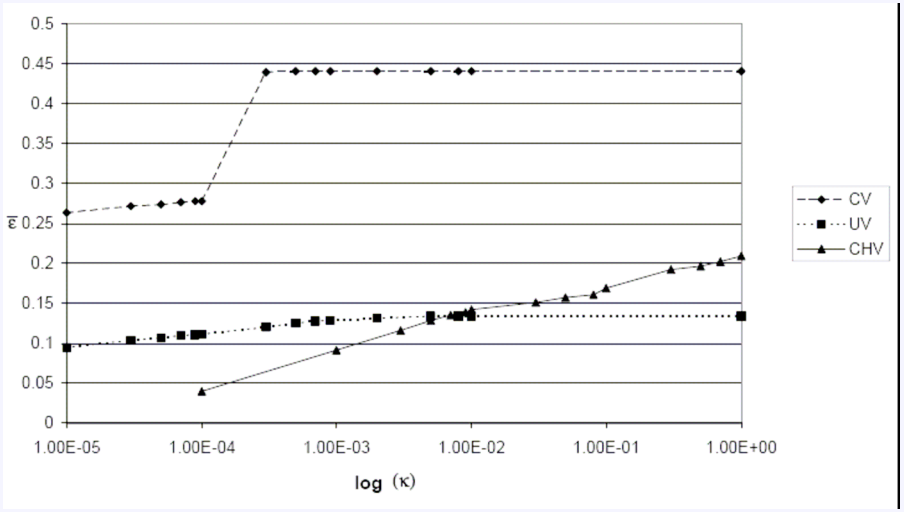


Η μέθοδος αυτή βασίζεται σε μια πραγματική παρατήρηση ότι η ακριβής απάντηση ανακτάται μέσα από πολλαπλά στάδια αναζήτησης τα οποία βελτιώνουν την ακρίβεια (precision) των προηγούμενων προσεγγιστικών αποτελεσμάτων.

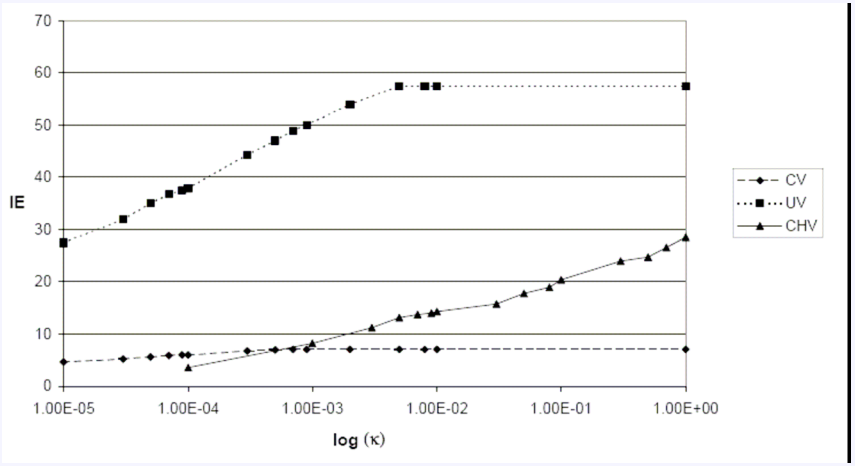
Προσεγγιστικά Ερωτήματα

- Approximation through the slowdown of distance improvements





Γενικά η μέθοδος αυτή πέτυχε υψηλή βελτίωση στην απόδοση και διατήρησε καλή ποιότητα στα αποτελέσματα.

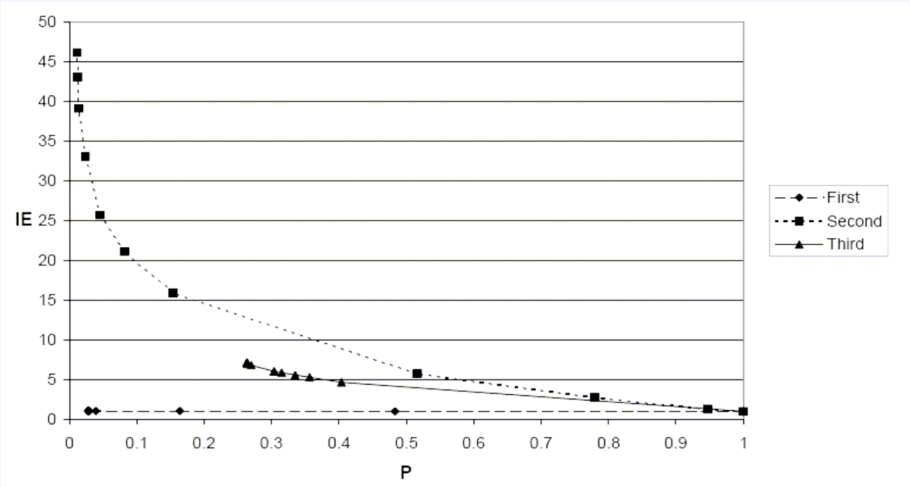


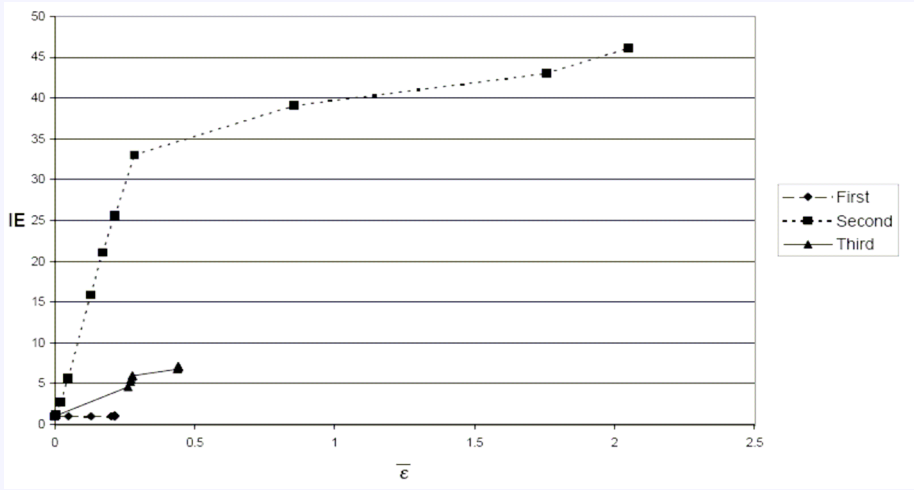
Προσεγγιστικά Ερωτήματα

- Τα πειράματα έδειξαν πως η πρώτη μέθοδος δεν είναι πολύ αποδοτική και καλύτερη όλων είναι η δεύτερη μέθοδος που έχει και υψηλή απόδοση και καλό precision.
- Γενικά αποδεικτικέ πως το να βρω 10 αντικείμενα από τα 100 καλύτερα είναι 100 φορές πιο γρήγορο από το να βρω τους πραγματικούς 10 πλησιέστερους γειτόνους.
- Γενικά και οι 3 μέθοδοι είναι εύκολο να υλοποιηθούν με δυσκολότερη υλοποίηση να έχει η μέθοδος προσέγγισης μέσω κατανομών απόστασης.
- Η δυσκολία έγκειται στην ανάγκη υπολογισμού και διατήρησης της κατανομής της απόστασης για όλα τα αντικείμενα της συλλογής.
- Σε τέτοιες περιπτώσεις καλύτερα να επιλέγεται η τρίτη μέθοδος.

Προσεγγιστικά Ερωτήματα

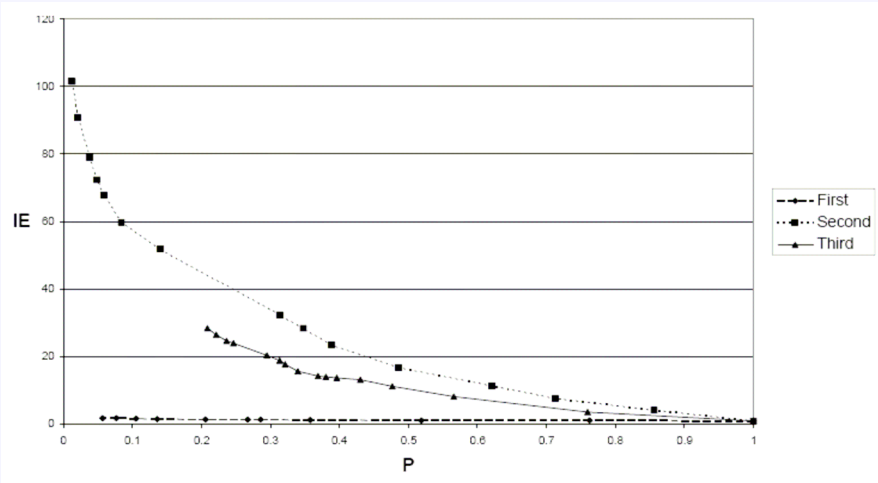
- CV αρχεία

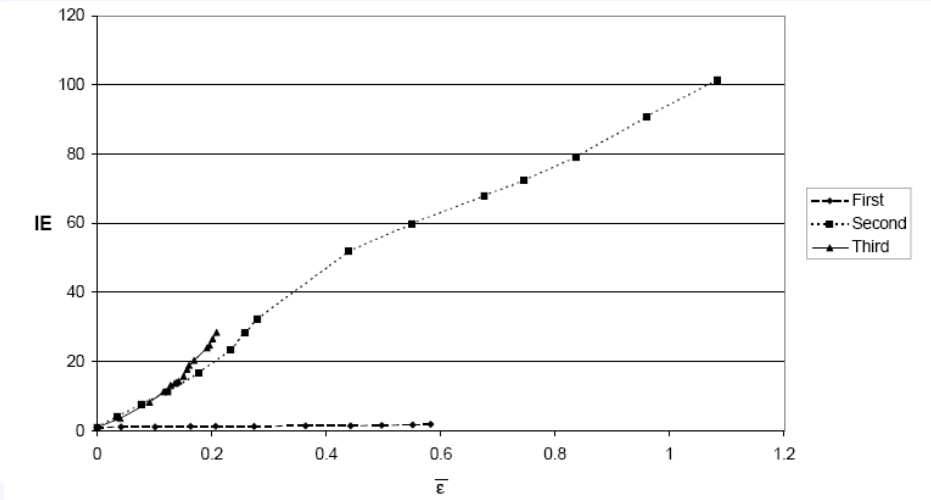




Προσεγγιστικά Ερωτήματα

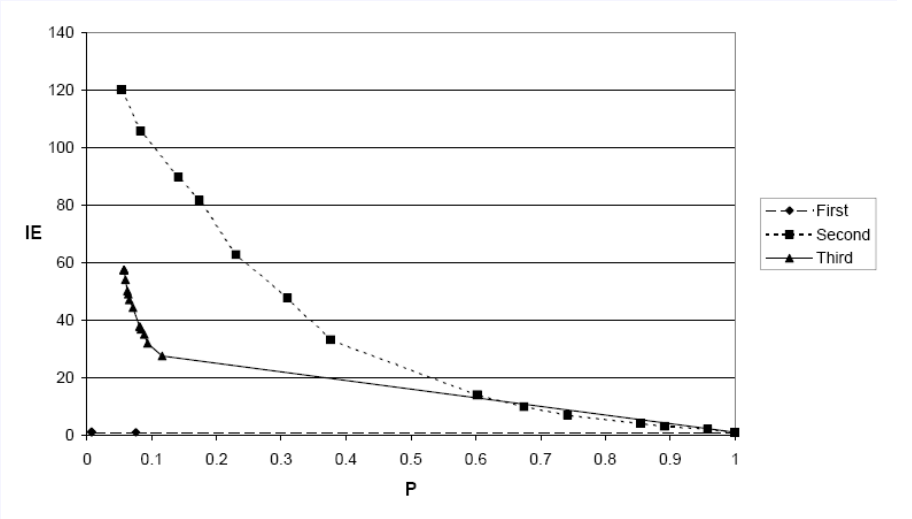
- CHV αρχεία

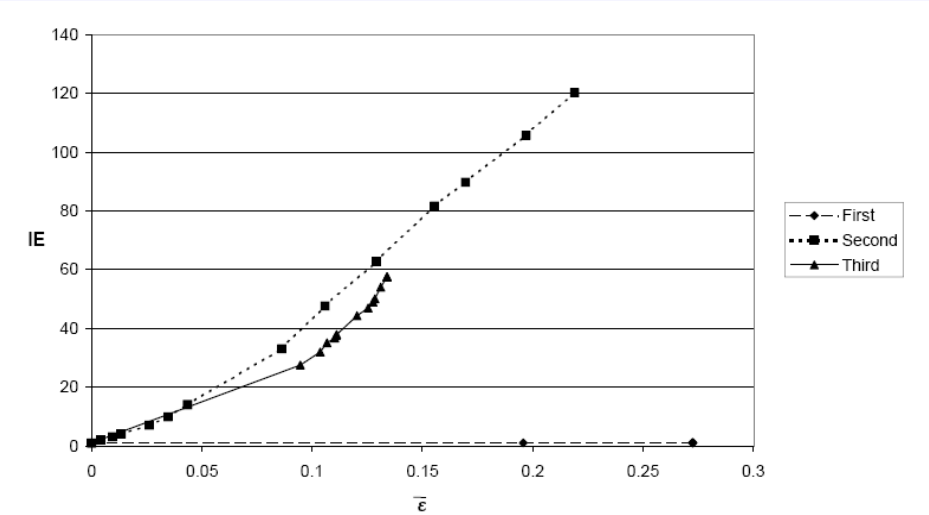




Προσεγγιστικά Ερωτήματα

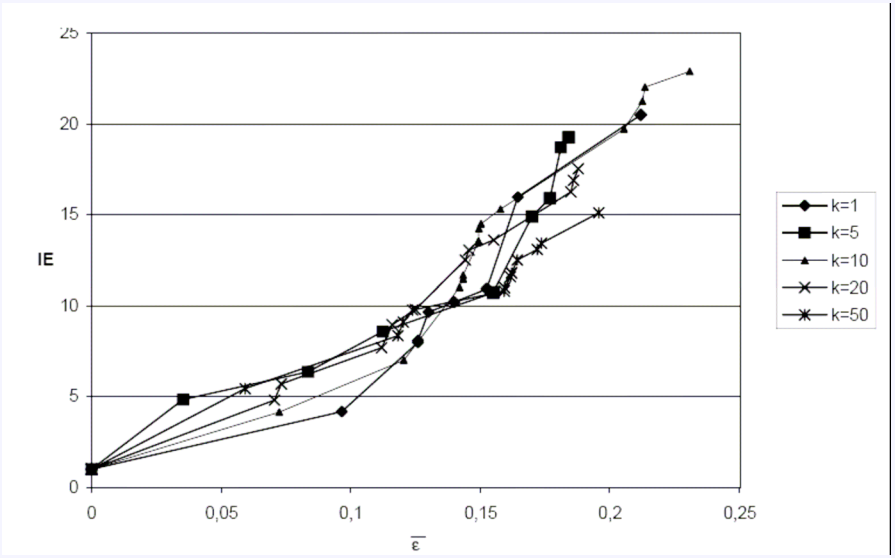
- UV αρχεία

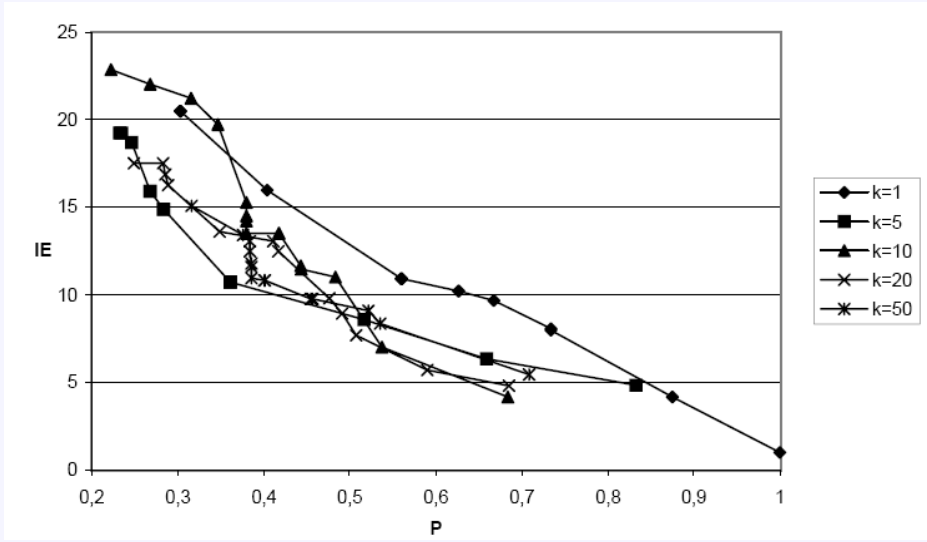




Προσεγγιστικά Ερωτήματα

- Ερωτήματα βάσει της τρίτης μεθόδου και για κ μεγαλύτερο του 10





Προσεγγιστικά Ερωτήματα

- Οι διάφορες μέθοδοι επεξεργασίας προσεγγιστικών ερωτημάτων μπορούν να ταξινομηθούν με βάση τα παρακάτω κριτήρια:
  - Ο τύπος των δεδομένων στα οποία θα εφαρμοστεί η μέθοδος
  - Οι μετρικές που θα χρησιμοποιηθούν για να υπολογίσουμε τα λάθη από την προσέγγιση
  - Τι εγγυήσεις υπάρχουν για την ποιότητα των αποτελεσμάτων
  - Ο βαθμός της αλληλεπίδρασης του χρήστη με την μέθοδο

Προσεγγιστικά Ερωτήματα

- **Τύποι Δεδομένων**
  - *MS (metric spaces)*
  - *VS (vector spaces)*
  - *VS_{Lp} (vector spaces, L_p distance)*

Προσεγγιστικά Ερωτήματα

- **Μετρικές Λάθους**
  - *CS (changing space)*
  - *RC (reducing comparisons)*

Προσεγγιστικά Ερωτήματα

- **Εγγυήσεις Ποιότητας**
  - *NG (no guarantees)*
  - *DG (deterministic guarantees)*
  - *PG (probabilistic guarantees)*
    - *PGpar (probabilistic guarantees, parametric)*
    - *PGnpar (probabilistic guarantees, non-parametric)*

Προσεγγιστικά Ερωτήματα

- **Αλληλεπίδραση με τον Χρήστη**
  - ***SA (static approach)***
  - ***IA (interactive approach)***

Προσεγγιστικά Ερωτήματα

- Το παραπάνω σχήμα ταξινόμησης των προσεγγιστικών μεθόδων για ερωτήματα ομοιότητας μπορεί να αποδεικτεί αρκετά χρήσιμο γιατί μέσω αυτού μπορούμε να εντοπίσουμε συσχετίσεις και ομοιότητες μεταξύ των μεθόδων που με μια πρώτη ματιά μπορεί να μην είναι προφανείς.