

Πανεπιστήμιο Πειραιώς Τμήμα Πληροφορικής



Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (Data Warehousing & Data Mining)

Εξόρυξη Γνώσης από Χωρικά Δεδομένα (spatial data mining)

Γιάννης Θεοδωρίδης, Νίκος Πελέκης

Ομάδα Διαχείρισης Δεδομένων
Εργαστήριο Πληροφοριακών Συστημάτων

<http://isl.cs.unipi.gr/db>

Περιεχόμενα



- Χωρικά δεδομένα και εφαρμογές εξόρυξης γνώσης
- Γενίκευση / Εξειδίκευση (προοδευτική βελτίωση)
- Ανακάλυψη κανόνων χωρικών συσχετίσεων
- Χωρική κατηγοριοποίηση
- Χωρική συσταδοποίηση

Περιεχόμενα



- Χωρικά δεδομένα και εφαρμογές εξόρυξης γνώσης
- Γενίκευση / Εξειδίκευση (προοδευτική βελτίωση)
- Ανακάλυψη κανόνων χωρικών συσχετίσεων
- Χωρική κατηγοριοποίηση
- Χωρική συσταδοποίηση

3

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Χωρικό Αντικείμενο



- Περιγράφεται από χωρικά και μη χωρικά γνωρίσματα.
- Σε αυτά μπορεί να περιλαμβάνεται κάποιος τύπος σχετικός με θέση:
 - Γεωγραφικό Μήκος και Πλάτος
 - Ταχυδρομικός Κωδικός
 - Διεύθυνση
- Η ανάκτηση του αντικειμένου θα πρέπει να είναι δυνατή με χρήση χωρικών ή/και μη χωρικών γνωρισμάτων

4

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Εφαρμογές εξόρυξης γνώσης από χωρικά δεδομένα



- Γεωλογία
- Συστήματα Γεωγραφικών Πληροφοριών (GIS)
- Περιβαλλοντολογική Επιστήμη
- Γεωργία
- Ιατρική
- Ρομποτική
- Οπουδήποτε συνδυάζεται η χρονική με τη χωρική διάσταση

Απόσταση μεταξύ Αντικειμένων



- Σημειακά αντικείμενα: Ευκλείδεια, Manhattan, ...
- Μη σημειακά αντικείμενα:

- minimum.

$$dis(A, B) = \min_{(x_a, y_a) \in A, (x_b, y_b) \in B} dis((x_a, y_a), (x_b, y_b))$$

- maximum.

$$dis(A, B) = \max_{(x_a, y_a) \in A, (x_b, y_b) \in B} dis((x_a, y_a), (x_b, y_b))$$

- average.

$$dis(A, B) = average_{(x_a, y_a) \in A, (x_b, y_b) \in B} dis((x_a, y_a), (x_b, y_b))$$

- center.

$$dis(A, B) = dis((x_{ca}, y_{ca}), (x_{cb}, y_{cb}))$$

where (x_{ca}, y_{ca}) is a center point for object A .

Περιεχόμενα



- Χωρικά δεδομένα και εφαρμογές εξόρυξης γνώσης
- Γενίκευση / Εξειδίκευση (προοδευτική βελτίωση)
- Ανακάλυψη κανόνων χωρικών συσχετίσεων
- Χωρική κατηγοριοποίηση
- Χωρική συσταδοποίηση

7

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Προοδευτική Βελτίωση

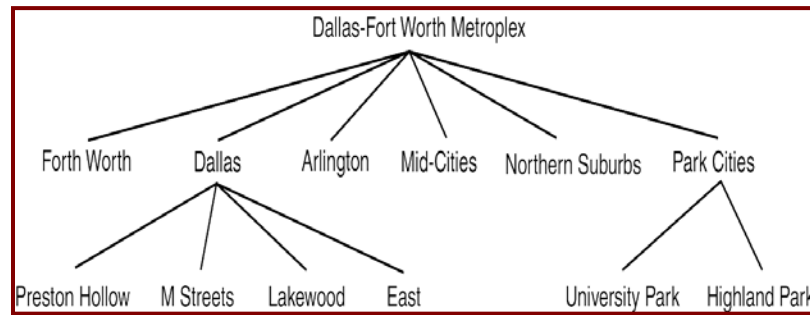


- Δίνει προσεγγιστικές απαντήσεις, προτού αναζητηθούν πιο ακριβείς.
- Φιλτράρει τα δεδομένα που δεν είναι εφαρμόσιμα σε ένα πρόβλημα.
- Τα ιεραρχικά επίπεδα βασίζονται σε χωρικές συσχετίσεις.
- Τα κατηγορήματα που εξάγονται βελτιώνονται αναδρομικά έως ότου βρεθούν οι ακριβείς απαντήσεις.

8

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Προοδευτική Βελτίωση



9

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

STING

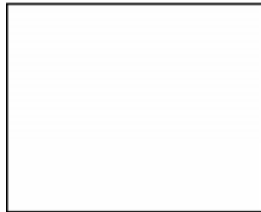


- S**T**atistical Information Grid-based
- Χρησιμοποιεί μια ιεραρχική τεχνική για τη διαίρεση των χωρικών περιοχών σε ορθογώνια κελιά
- Κάθε κόμβος στη δομή πλέγματος συνοψίζει την πληροφορία για τα στοιχεία εντός της.
- Μπορεί να θεωρηθεί ως τεχνική ιεραρχικής συσταδοποίησης

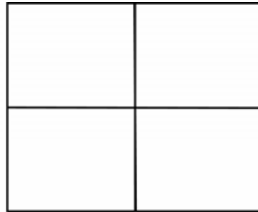
10

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

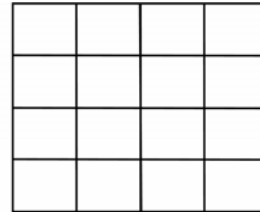
STING



a) Level 1



b) Level 2



c) Level 3

11

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Ο Αλγόριθμος STING Build



```

Input:
     $D$            //Data to be placed in the hierarchical structure.
     $k$            //Number of desired cells at the lowest level.
Output:
     $T$            // Tree.
STING BUILD Algorithm
    // Create empty tree from top down.
     $T =$  root node with data values initialized; // Initially only root node
     $i = 1$ ;
    repeat
        for each node in level  $i$  do
            create 4 children nodes with initial values;
             $i = i + 1$ ;
    until  $4^i = k$ ;
    // Populate tree from bottom up.
    for each item in  $D$  do
        determine leaf node  $j$  associated with the position of  $D$ ;
        update values of  $j$  based on attribute values in item;
     $i := \log_4(k)$ ;
    repeat
         $i := i - 1$ ;
        for each node  $j$  in level  $i$  do
            update values of  $j$  based on attribute values in its 4 children;
    until  $i = 1$ ;
  
```

12

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Ο Αλγόριθμος STING



Input:

T // Tree.
 q // Query.

Output:

R // Regions of relevant cells.

STING Algorithm

$i = 1$;

repeat

for each node in level i do

determine if this cell is relevant to q and mark as such;

$i = i + 1$;

until all layers in the tree have been visited;

identify neighboring cells of relevant cells to create regions of cells;

13

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Περιεχόμενα



- Χωρικά δεδομένα και εφαρμογές εξόρυξης γνώσης
- Γενίκευση / Εξειδίκευση (προοδευτική βελτίωση)
- Ανακάλυψη κανόνων χωρικών συσχετίσεων
- Χωρική κατηγοριοποίηση
- Χωρική συσταδοποίηση

14

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Παραδείγματα



- **Κανόνας χωρικού χαρακτηριστικού**

Στο Dallas το μέσο οικογενειακό εισόδημα είναι \$50,000.

- **Κανόνας χωρικού διαχωρισμού**

Στο Dallas το μέσο οικογενειακό εισόδημα είναι \$50,000, ενώ στο Plano είναι \$75,000.

- **Κανόνας χωρικής συσχέτισης**

Στο Dallas το μέσο οικογενειακό εισόδημα για οικογένειες που ζουν κοντά στη White Rock Lake είναι \$100,000.

15

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Κανόνες Χωρικών Συσχετίσεων



- Είτε το πρότερο είτε το απότοκο του κανόνα πρέπει να περιέχει κάποια χωρικά κατηγορήματα (π.χ. κοντά)
- Η βάση που εξετάζεται θεωρείται ως ένα σύνολο από χωρικά αντικείμενα.
- Οι κανόνες μπορούν να δημιουργηθούν χρησιμοποιώντας τεχνικές προοδευτικής βελτίωσης.

16

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Αλγόριθμος ανακάλυψης κανόνων χωρικών συσχετίσεων



Input:

D //Data, including spatial and nonspatial attributes
 C //Concept hierarchies
 s //Minimum support for levels.
 α //Confidence
 q //Query to retrieve interested objects
 P //Topological predicate(s) of interest

Output:

R // Spatial Association Rules.

SPATIAL ASSOCIATION RULE Algorithm

$D' = q(D)$;
 CP is built by applying the coarse predicate version of P to D' ;
 // CP consists of the set of coarse predicates satisfied by pairs of objects in D' .
 determine the set of frequent coarse predicates FCP
 by finding the coarse predicates which satisfy s ;
 find the set of frequent fine predicates FFP from FCP ;
 find R by finding all frequent fine predicates and then generating rules;

17

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Περιεχόμενα



- Χωρικά δεδομένα και εφαρμογές εξόρυξης γνώσης
- Γενίκευση / Εξειδίκευση (προοδευτική βελτίωση)
- Ανακάλυψη κανόνων χωρικών συσχετίσεων
- Χωρική κατηγοριοποίηση
- Χωρική συσταδοποίηση

18

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Χωρική Κατηγοριοποίηση



- Στοχεύει στη διαμέριση συνόλων χωρικών αντικειμένων
- Μπορεί να γίνει κατηγοριοποίηση με χρήση μη χωρικών ή/και χωρικών γνωρισμάτων.
- Τεχνικές γενίκευσης και προοδευτικής βελτίωσης μπορούν να χρησιμοποιηθούν.

19

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Επέκταση του ID3



- **Γράφος Γειτνίασης**
 - Κόμβοι – αντικείμενα
 - Ακμές – συνδέουν γείτονες
- Ο ορισμός του «γείτονα» ποικίλει.
 - Μπορεί να οριστεί βάσει οποιασδήποτε συσχέτισης μεταξύ των χωρικών αντικειμένων.
- Ο ID3 θεωρεί για σκοπούς κατηγοριοποίησης τα μη χωρικά γνωρίσματα όχι μόνο του αντικείμενου-στόχου αλλά και των γειτονικών αντικειμένων.

20

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Δένδρο Χωρικής Απόφασης



- Παρόμοια προσέγγιση με αυτή που χρησιμοποιείται στους κανόνες χωρικών συσχετίσεων.
- Βασίζεται στο ότι τα χωρικά αντικείμενα μπορούν να περιγραφούν βάση των αντικειμένων που είναι κοντά σε αυτά – **Ενδιάμεση Ζώνη**.
- Περιγραφή των κλάσεων βασισμένη σε μια συνάθροιση των πιο σχετικών κατηγορημάτων για κοντινά αντικείμενα.

21

ΠΑ.ΠΕΙ – Γιάννης Θεοδορίδης

Ο Αλγόριθμος του Δένδρου Χωρικής Απόφασης



Input:

D //Data, including spatial and nonspatial attributes
 C //Concept hierarchies

Output:

T // Binary Decision Tree.

SPATIAL DECISION TREE Algorithm

find a sample S of data from D with known classification;
 identify the best predicates p to use for classification;
 determine the best buffer size and shape;
 using p and C , generalize the predicates for each buffer;
 build binary T using the generalized predicates and ID3;

22

ΠΑ.ΠΕΙ – Γιάννης Θεοδορίδης

Περιεχόμενα



- Χωρικά δεδομένα και εφαρμογές εξόρυξης γνώσης
- Γενίκευση / Εξειδίκευση (προοδευτική βελτίωση)
- Ανακάλυψη κανόνων χωρικών συσχετίσεων
- Χωρική κατηγοριοποίηση
- Χωρική συσταδοποίηση

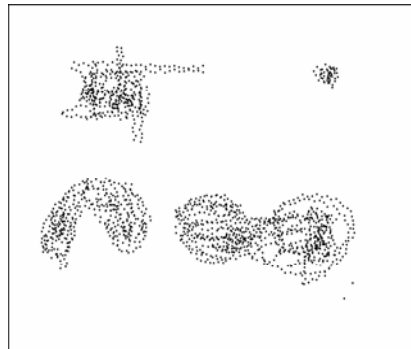
23

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Χωρική Συσταδοποίηση



- Εντοπίζουν συστάδες από διαφορετικά σχήματα.
- Ένας αλγόριθμος που δουλεύει χρησιμοποιώντας κέντρα βάρους και απλές μετρήσεις απόστασης πιθανόν δεν θα είναι σε θέση να αναγνωρίζει ασυνήθιστα σχήματα.
- Οι συστάδες πρέπει να προκύπτουν ανεξάρτητα της σειράς με την οποία εξετάστηκαν τα σημεία στο χώρο.



24

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

DBCLASD



- Επέκταση του αλγορίθμου DBSCAN
- Distribution Based Clustering of LARge Spatial Databases (συσταδοποίηση μεγάλων βάσεων χωρικών δεδομένων βασισμένη σε κατανομές)
- Υποθέτει ότι τα στοιχεία εντός μιας συστάδας είναι ομοιόμορφα κατανομημένα.
- Επιχειρεί να προσδιορίσει την κατανομή που ικανοποιείται από τις αποστάσεις μεταξύ πλησιέστερων γειτόνων.
- Στοιχεία προστίθενται στη συστάδα, όσο το σύνολο των πλησιέστερων – βάσει της απόστασης – γειτόνων ικανοποιεί την υπόθεση της ομοιόμορφης κατανομής.

25

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Ο Αλγόριθμος DBCLASD



Input:

D //Spatial objects to be clustered.

Output:

K // Set of Clusters.

DBCLASD Algorithm

$k = 0$; // Initially there are no clusters.

$c = \emptyset$; // Initialize the set of candidates to be empty.

for each point p in D do

if p is not in a cluster then

create a new cluster C and put p in C ;

add neighboring points of p to C ;

for each point q in C do

add the points in the neighborhood of q which have not been processed to c ;

expand C ;

26

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης