ANNO ACCADEMICO 2006-2007
SISTEMI INFORMATIVI GEOGRAFICI

Sistemi Informativi Territoriali

Gli indici spaziali: R-tree

ALBERTO BELUSSI

MAGGIO 2007

Strutture d'accesso per dati spaziali

L'interrogazione di una Base di Dati geografica è influenzata dalla suddivisione dell'informazione territoriale in due parti: il dato geometrico e il dato alfanumerico. Queste due componenti hanno caratteristiche diverse sia per quanto riguarda la loro memorizzazione sia per le relazioni che esistono nei corrispondenti domini.

In particolare, i dati alfanumerici sono istanze appartenenti a domini monodimensionali, dove è presente una relazione d'ordine. Il dominio dei numeri interi e il dominio delle stringhe di caratteri di lunghezza minore di 10 sono esempi di domini alfanumerici. In essi sono presenti rispettivamente l'usuale relazione d'ordine sugli interi e l'ordinamento alfabetico sulle stringhe. I dati geometrici appartengono invece a domini multidimensionali (si considerano almeno i valori geometrici del piano cartesiano), dove non esiste una relazione d'ordine di riferimento e sono definiti altri tipi di relazioni di natura spaziale.

Le strutture d'accesso per i dati alfanumerici tradizionali non sono adatte per l'indicizzazione di dati geometrici, proprio a causa della mancanza di una relazione d'ordine di riferimento.

Strutture d'accesso per dati spaziali

Per i dati geometrici sono quindi state studiate strutture d'accesso nuove e dedicate alla trattazione di dati spaziali istanziati in spazi a due o tre dimensioni. Lo scopo di tali strutture è quello di ottimizzare le interrogazioni di selezione basate su proprietà spaziali e in particolare le "range queries".

Molti sono gli approccio proposti in letteratura per la costruzione di strutture d'accesso per dati spaziali, di seguito si elencano i principali:

- Strutture derivate come "generalizzazione" di strutture per dati alfanumerici (monodimensionali):
  - dai B-tree sono nati i **k-d-B-tree** [Robinson 1981]
  - da extendible hashing il metodo **EXCELL** [Tamminen 1982]
  - da dynamic hashing il metodo **Grid file** [Nievergelt et al. 1984]
- Mapping da R^d a R:
  z-order, curve frattali, ecc.
- Strutture disegnate *ad hoc*, per esempio:
  - **R-tree** [Guttman 1984]
  - **R^+-tree** [Sellis et al. 1987]
  - R-file [Hutflesz et al. 1990]

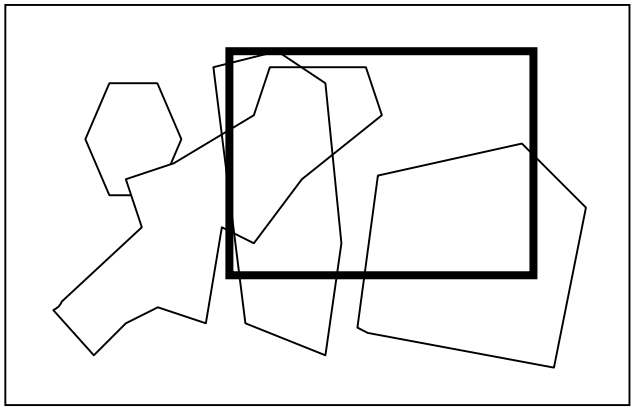
Strutture d'accesso per dati spaziali

Perché non usare strutture d'accesso tradizionali
(ad esempio il B+-tree)?

Possibile approccio per dati spaziali con estensione (poligoni):

usare il baricentro dei poligoni e costruire un B+-tree usando come chiave di ricerca la concatenazione delle coordinate X e Y.

Tale approccio, pur consentendo di indicizzare i poligoni, produce un indice che non si comporta bene nelle interrogazioni, in quanto i poligoni vicini nello spazio non è detto che risultino vicini nell'indice e ciò penalizza le interrogazioni spaziali più frequenti quali, ad esempio, le "range queries".

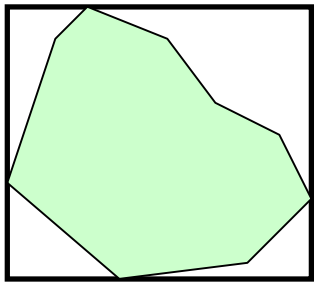


Strutture d'accesso per dati spaziali con estensione (poligoni)

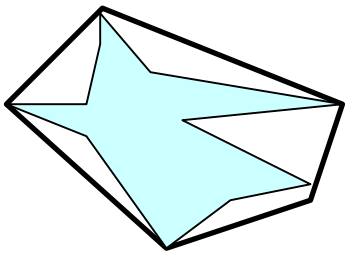
La memorizzazione e indicizzazione di dati spaziali con estensione si basa sul principio dell'**approssimazione della geometri**a.

Ciò significa che i dati spaziali vengono rappresentati sinteticamente attraverso un descrittore che solitamente approssima con un rettangolo MBR (o poligono convesso) il dato spaziale.

MBR



Poligono convesso



Strutture d'accesso per dati spaziali con estensione (poligoni)

Possibili organizzazione della struttura d'accesso basata sull'MBR dei poligoni.

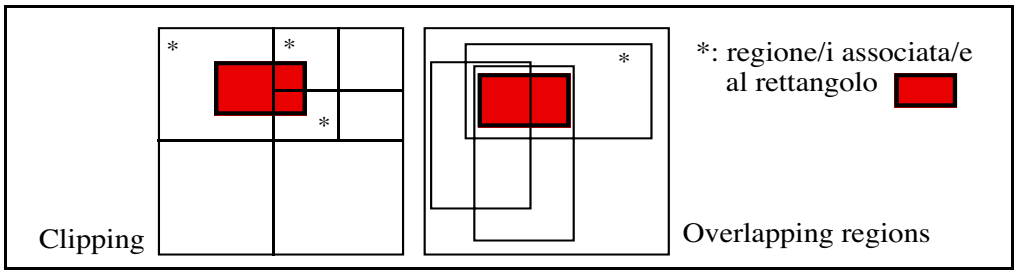
Organizzazione della struttura:
Suddivisione dello spazio in regioni rettangolari che raggruppano ciascuna un certo insieme di MRB dei poligoni dell'insieme da indicizzare.

Problema:
Dato un nuovo poligono P e il suo MBR (MBR(P)), come si assegna MBR(P) alle regioni definite nella struttura d'accesso?

Soluzioni:

Overlapping
Si associa sempre MBR(P) a una sola regione, permettendo alle regioni dell'indice di sovrapporsi (questa è la strategia adottata da: R-tree).
In generale, le regioni non coprono l'intero spazio, e possono espandersi a seguito di inserimenti

Clipping
Se tra le regioni definite nella struttura nessuna contiene interamente MBR(P), allora si esegue il *clipping* di MBR(P), vale a dire lo si associa a tutte le regioni da esso toccate (questa è la strategia adottata da: R^+-tree).
In questo caso le regioni della struttura non si sovrappongono.



Strutture d'accesso per dati spaziali: R-Tree [Guttman 1984]

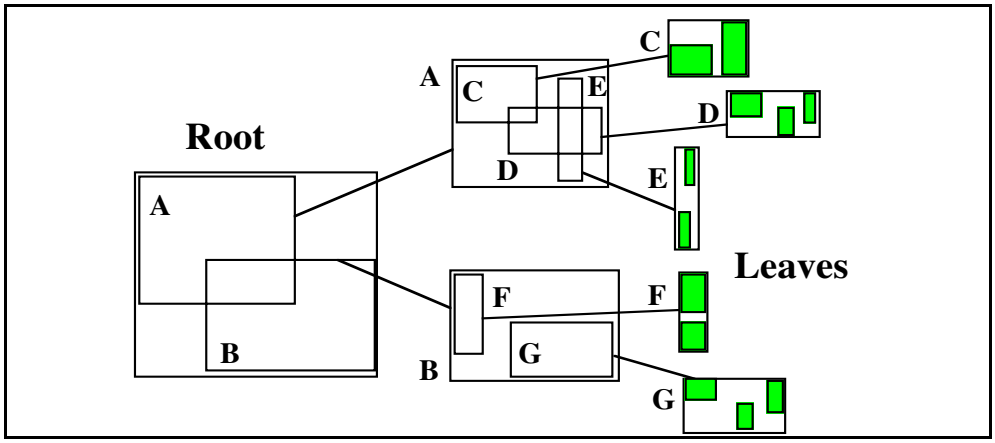
E' una struttura ad albero bilanciata e paginata, basata sull'annidamento gerarchico di ***overlapping region***s.

Caratteristiche:

- Ogni nodo (pagina indice) corrisponde ad una regione rettangolare, definita come l'MBR che contiene tutte le regioni figlie.
- L'utilizzazione di ogni nodo varia tra il 100% e un valore minimo (<=50%) che rappresenta un parametro di progetto dell'R-tree.
- Meccanismi di gestione simili a quelli del B^+ -tree, con la differenza che l'inserimento di un oggetto, e gli eventuali split che ne conseguono, possono essere gestiti con politiche diverse.

Al fine di garantire una buona selettività nella fase di filtraggio, ovvero ridurre al minimo il numero di nodi da esaminare, si sono realizzate tecniche di aggiornamento che tendono a minimizzare la sovrapposizione tra le diverse regioni.

La presenza di regioni sovrapposte crea un problema anche per le interrogazioni che richiedono il/i poligono/i che contengono un punto dato, in quanto non è noto a priori in quale regione dell'indice il rettangolo è stato inserito.



Strutture d'accesso per dati spaziali: R-Tree [Guttman 1984]

Inserimento di un nuovo valore nell'R-Tree.

Per eseguire un inserimento si deve scendere lungo l'albero e scegliere, ad ogni nodo, il nodo figlio corrispondente alla regione che necessita del **minor incremento di area per inserire il nuovo MBR** (ci possono essere anche altri criteri).

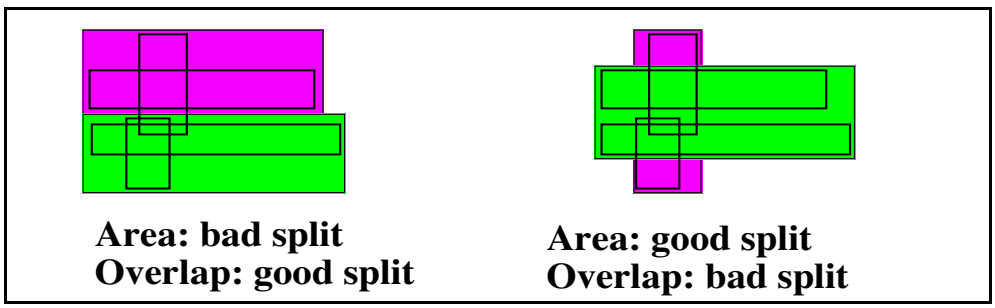
Nel caso in cui si giunga alla saturazione di una foglia dell'indice si esegue lo split della foglia e si propagano le modifiche al livello superiore.

Per la ridistribuzione degli MBR nelle due foglie (o nei nodi intermedi, se lo split si propaga) non si ha un criterio ben definito, come per i B^+-tree.

Criteri alternativi:

- **Minimizzare la somma delle aree risultanti**
- **Minimizzare l'area in sovrapposizione**

Gli algoritmi per implementare i due criteri sono di complessità NP-Hard, si implementano quindi algoritmi approssimati polinomiali.



Strutture d'accesso per dati spaziali: R$^+$-Tree [Sellis 1987]

Si tratta di una variante dell'R-Tree.

Mira a ridurre i problemi che nascono dalla presenza di sovrapposizione tra le regioni nei livelli intermedi dell'albero.

L'idea è quella di:

- eseguire il *clipping* delle regioni in modo da ottenere regioni che non si sovrappongono, e
- *replicare* l'associazione ad un MBR in ogni regione di livello superiore che lo contiene.

Un R$^+$-tree occupa normalmente più spazio di un R-tree, tende a migliorare le prestazioni in ricerca, ma richiede algoritmi più complessi per l'aggiornamento della struttura.