Bases de Datos Multimedia

Capítulo 5

Índices Métricos

Este material se basa en el curso de Base de Datos Multimedia del DCC de la Universidad de Chile (Prof. Benjamín Bustos).

5.1 Conceptos básicos

- Para ciertas aplicaciones, es posible definir una función de distancia entre objetos multimedia, pero no pueden ser descritos en forma razonable como vectores
- También es posible que calcular la similitud directamente entre objetos sea más simple que transformarlos en vectores
- Índices métricos: función de distancia es una métrica

5.1 Conceptos básicos

- Definición de espacio métrico
  - Universo de objetos válid $\mathbb{X}$:
  - Función de distanci $\delta : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}^{+}$
  - ($\mathbb{X}$, δ) representa un espacio métrico ssi δ cumple con las siguientes propiedades:
    - t Positividad estricta $\forall x, y \in \mathbb{X},\ x \neq y \Rightarrow \delta(x, y) > 0$
    - Simetría $\forall x, y \in \mathbb{X},\ \delta(x, y) = \delta(y, x)$
    - Reflexividad $\forall x \in \mathbb{X},\ \delta(x, x) = 0$
    - Desigualdad triangu $\forall x, y, z \in \mathbb{X},\ \delta(x, z) \leq \delta(x, y) + \delta(y, z)$

5.1 Conceptos básicos

- Objetos de $\mathbb{X}$on directamente comparados utilizando δ
- δ indica el grado de disimilitud entre dos objetos
- Ejemplo: strings + distancia de edición
  - String: secuencia de caracteres
  - Distancia de edición: mínimo número de inserciones, eliminaciones o sustituciones para transformar un string en otro

5.2 Búsqueda en espacios métricos

- Base de datos: $\mathbb{U} \subset \mathbb{X}$
- Algoritmo ingenuo: búsqueda secuencial
- En general, en espacios métricos se considera que la función de distancia es computacionalmente costosa de calcular
- Tiempo total para evaluar una consulta

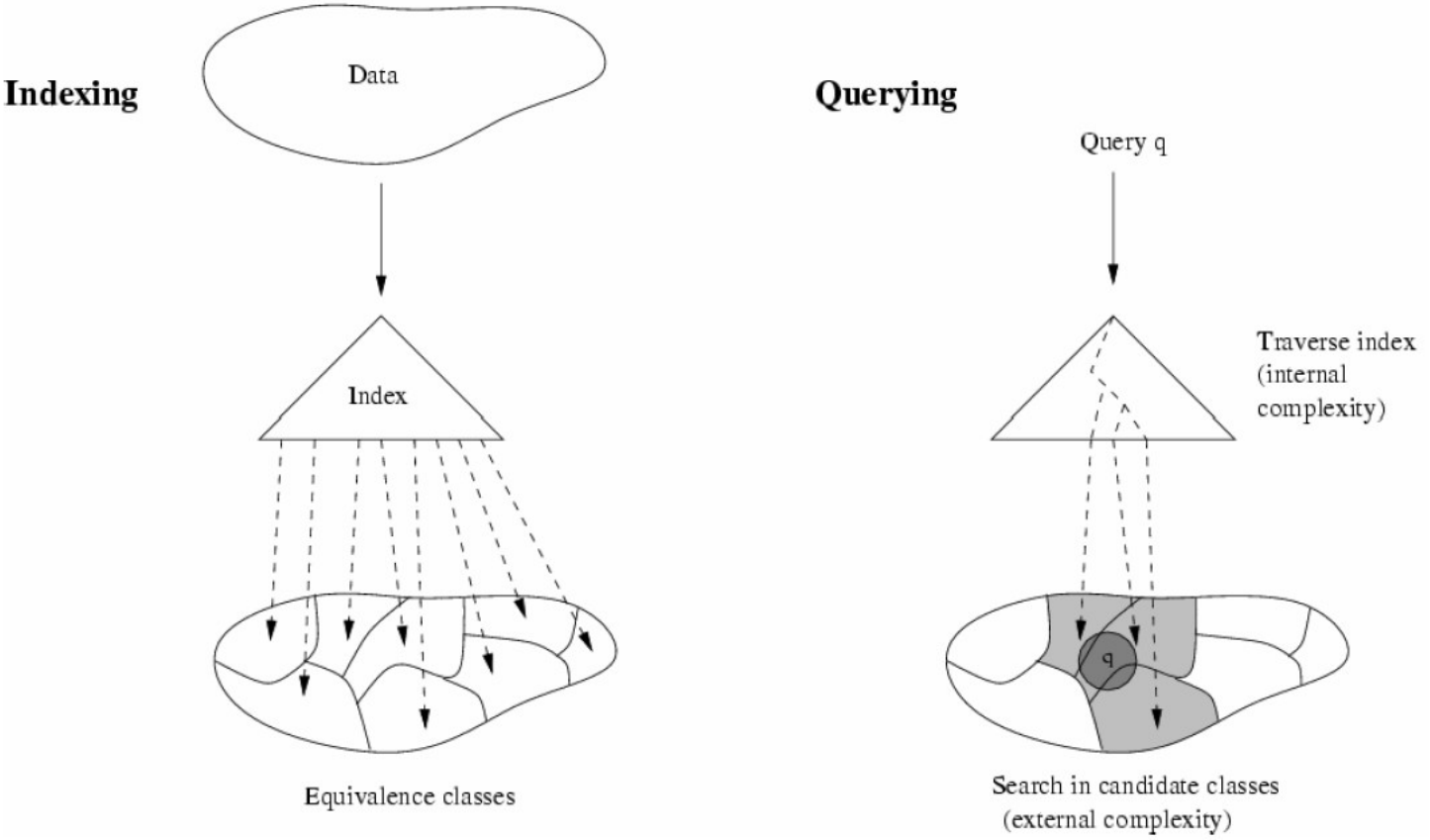
$$T = \text{cómputos de distancia} \cdot \text{complejidad de } \delta + \text{tiempo de CPU} + \text{tiempo de E/S}$$

5.2 Búsqueda en espacios métricos

- Índices métricos tratan de minimizar el número de cómputos de distancia necesarios para responder búsquedas por similitud
- Otros componentes del costo pueden ser obviados (depende de la aplicación)
- Índice métrico particiona el espacio en clases de equivalencia
  - Durante la búsqueda se descartan clases
  - Clases no descartadas deben ser examinadas

5.2 Búsqueda en espacios métricos

- Indexamiento y búsqueda

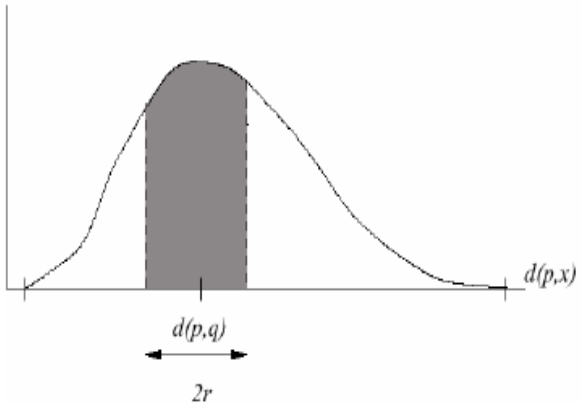


5.2 Búsqueda en espacios métricos

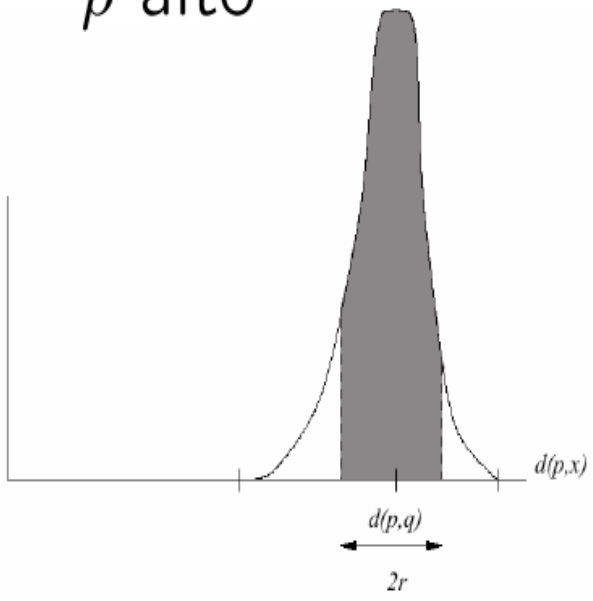
- Efectos en espacios métricos

Dimensión intrínseca $\rho = \frac{\mu^2}{2\sigma^2}$

ρ bajo



ρ alto



5.3 Índices basados en pivotes

- Un pivote es un objeto distinguido de la BD
- Puede ser utilizado para descartar objetos durante una búsqueda
- Dada una consulta por rango *(q,r)* y un pi

$$\delta(p,u) \leq \delta(p,q)+\delta(q,u) \text{ y } \delta(p,q) \leq \delta(p,u)+\delta(u,q)$$

- Cota inf

$$|\delta(p,q) - \delta(p,u)| \leq \delta(q,u)$$

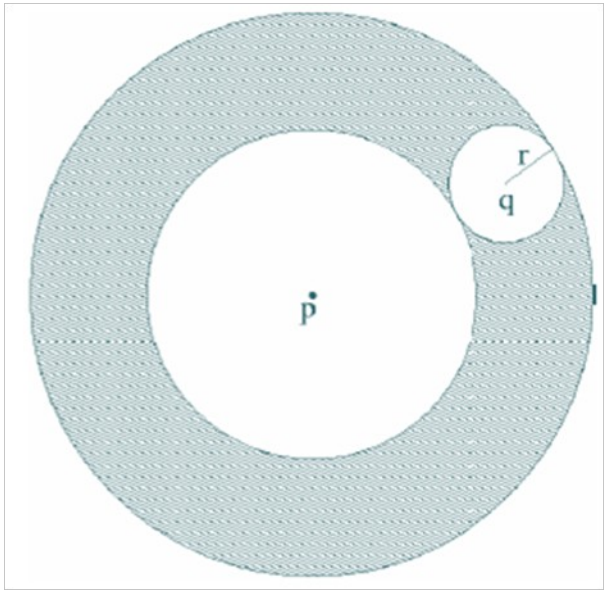
q y *u*

5.3 Índices basados en pivotes

- Criterio de exclusión $|\delta(p,q) - \delta(p,u)| > r$

  $r < |\delta(p,q) - \delta(p,u)| \leq \delta(q,u) \Rightarrow u$ no puede ser relevante

- Gráficamente



5.3 Índices basados en pivotes

- Algoritmo de búsqueda canónico basado en pivotes
  - Elegir *k* pivotes entre los objetos de la BD
  - Índice consiste en *kn* distancias precalculadas entre pivotes y objetos de la BD

	p_1	...	p_k
u_1	$\delta(p_1, u_1)$	...	$\delta(p_k, u_1)$
...	...	...	...
u_n	$\delta(p_1, u_n)$	...	$\delta(p_k, u_n)$

5.3 Índices basados en pivotes

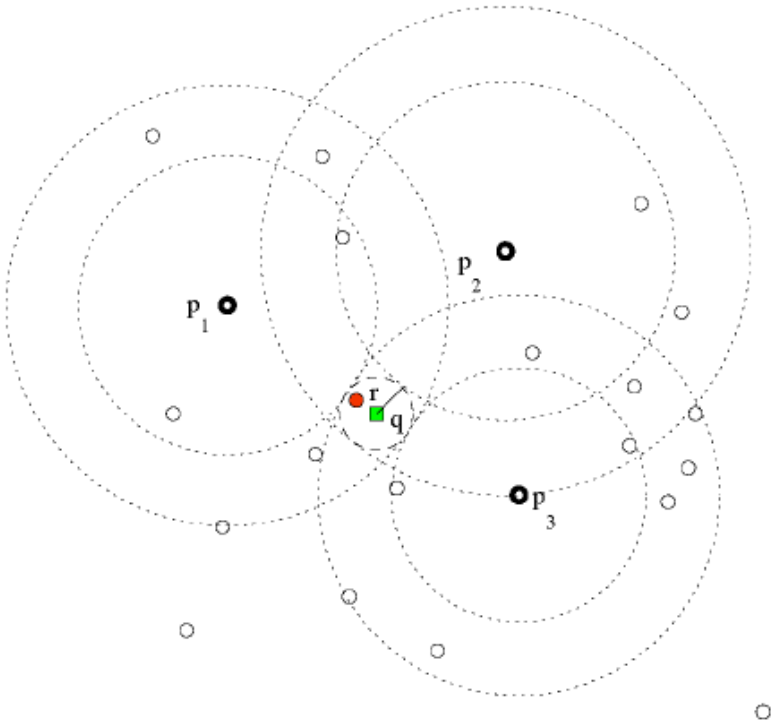
- Consultas por rango
  - Calcular *k* distancias entre *q* y pivotes
  - Descartar objetos usando el criterio de exclusión (basta que con un pivote se d

$$|\delta(p_i, q) - \delta(p_i, u)| > r \text{ para algún } p_i$$

  - Lista de objetos candidatos (los que no se pudieron descartar) deben ser comparados directamente con *q*

5.3 Índices basados en pivotes

- Ejemplo



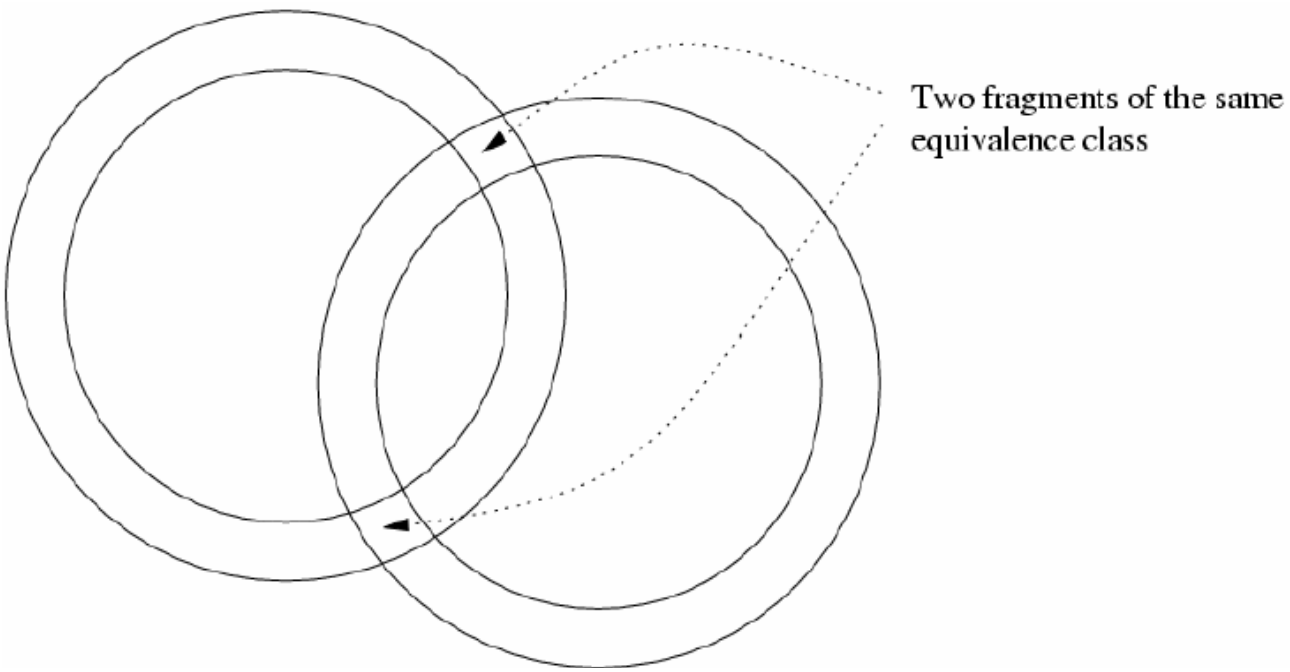
$$\begin{bmatrix} \delta(p_1,u_1) & \delta(p_2,u_1) & \delta(p_3,u_1) \\ \delta(p_1,u_2) & \delta(p_2,u_2) & \delta(p_3,u_2) \\ \vdots & \vdots & \vdots \\ \delta(p_1,u_n) & \delta(p_2,u_n) & \delta(p_3,u_n) \end{bmatrix}$$

Criterio de exclusión:

$$|\delta(p_i,q) - \delta(p_i,u)| > r$$

5.3 Índices basados en pivotes

- Clase de equivalencia con respecto a un conjunto de pivotes

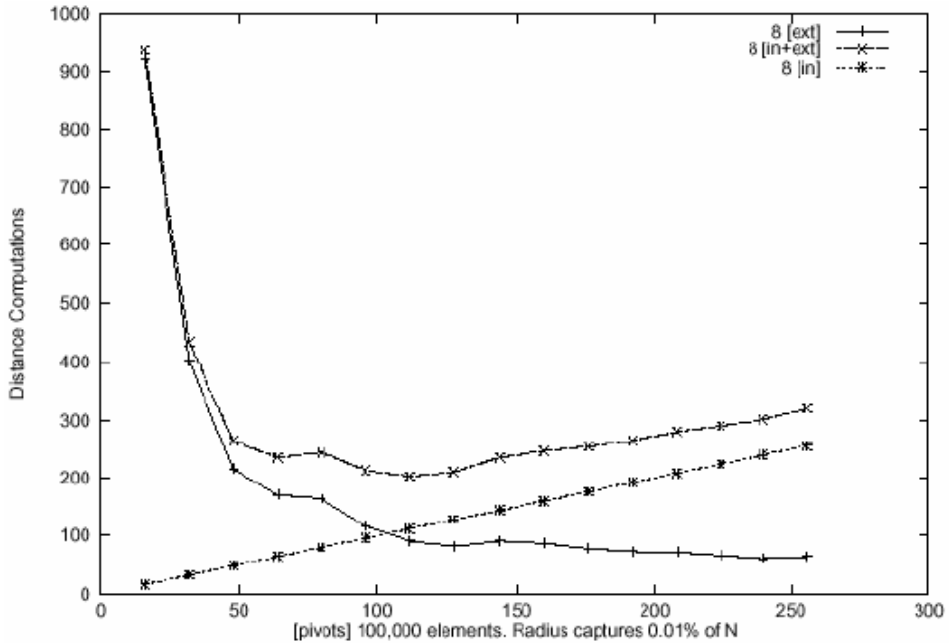


5.3 Índices basados en pivotes

- Complejidad de la búsqueda
  - Complejidad interna (k): cómputos de distancia entre pivotes y q
  - Complejidad externa (m): cómputos de distancia entre objetos no descartados y q
  - Complejidad total: $k+m$
  - Dado que la complejidad interna crece con el número de pivotes y la complejidad externa decrece con el número de pivotes: existe un número óptimo de pivotes a utilizar

5.3 Índices basados en pivotes

- Complejidad de la búsqueda vs. # pivotes



  - *k** muy alto, se utiliza toda la memoria disponible

5.3.1 Vantage Point Tree

- VPT es un árbol binario
- Algoritmo de construcción
  - Usar cualquier objeto *p* como raíz
  - Calcular la mediana de todas las distancias a *p*

$$M = \text{mediana}\{\delta(p,u),\ u \in \mathbb{U}\}$$

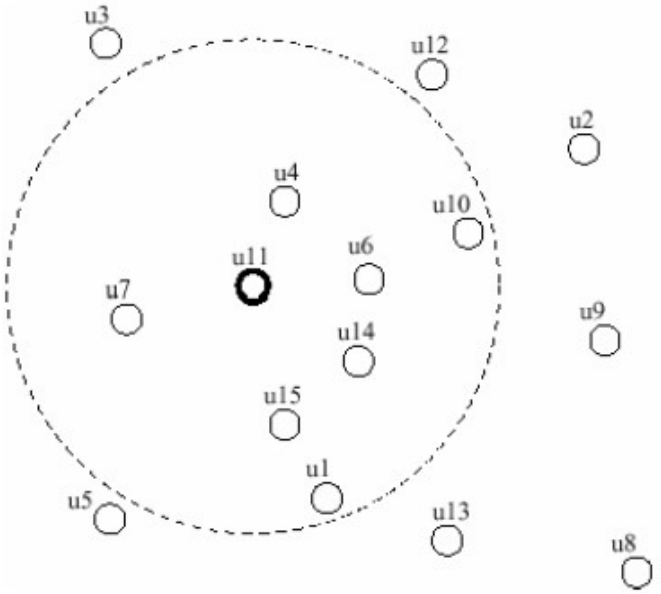
  - Subárbol izquierdo: $\delta(p,u) \leq M$
  - Subárbol derecho: $\delta(p,u) > M$
  - Proceder recursivam[illegible]e sólo haya un objeto en subárbol

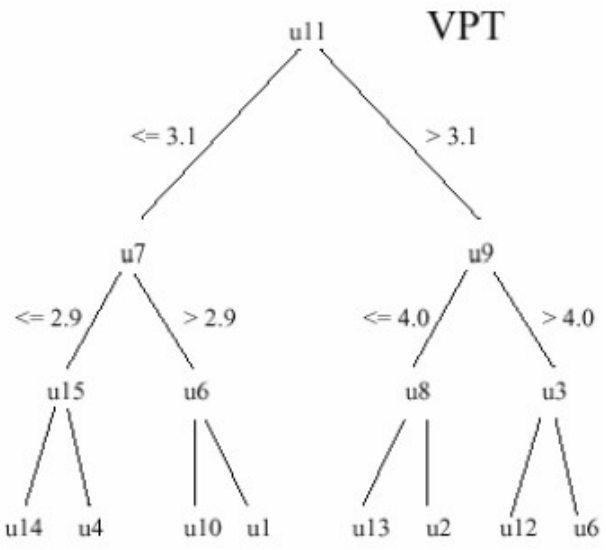
5.3.1 Vantage Point Tree

- Heurística propuesta para elegir *p*: escoger objetos “lejos” del resto
- Algoritmo de búsqueda, consultas por rango
  - Calcular
  - Si $d \leq r$, añadir *p* al resultado
  - Si $d - r \leq M$, buscar recursivamente en subárbol izquierdo
  - Si $d + r > M$, buscar recursivamente en subárbol derecho

5.3.1 Vantage Point Tree

- Ejemplo de un VPT





5.3.2 Técnicas de selección de pivotes

- Selección de los pivotes afecta el rendimiento del algoritmo de búsqueda
  - 2 pivotes muy cercanos no aumentan mucho el poder discriminativo del índice
- En general, pivotes se eligen aleatoriamente
- Existen técnicas de selección de pivotes

5.3.2 Técnicas de selección de pivotes

- Espacio de pivotes
  - Espacio *k*-dimensional (cada coordenada es la distancia entre *i*-ésimo pivote y el objeto)

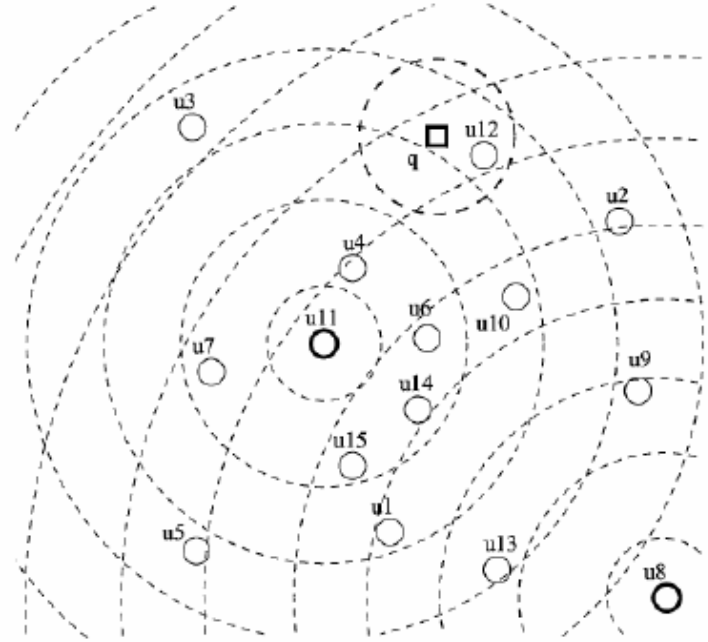
$$[x] = (\delta(x, p_1), \ldots, \delta(x, p_k)) \in \mathbb{R}^k$$

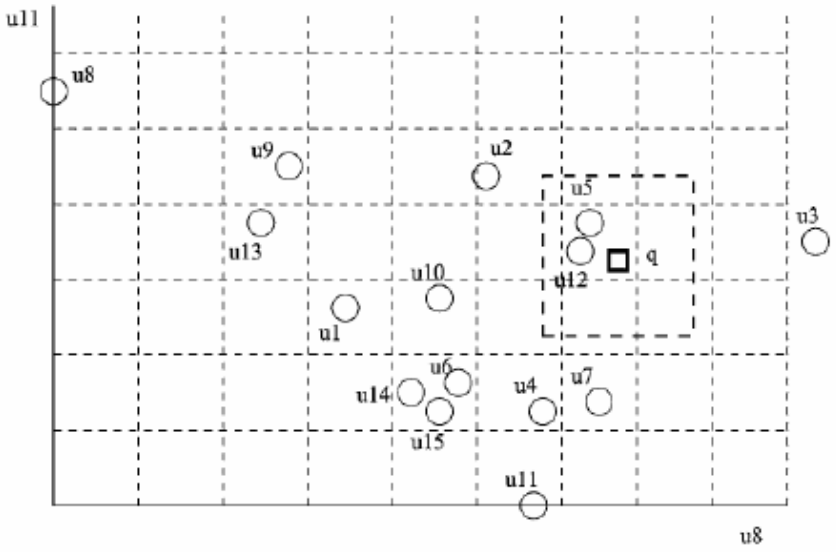
  - Distancia utilizada: distancia del máximo

$$\Delta_{p_1,\ldots,p_k}([x],[y]) = \max_{i=1}^{k}(|\delta(x,p_i) - \delta(y,p_i)|)$$

5.3.2 Técnicas de selección de pivotes

- Mapeando objetos al espacio de pivotes



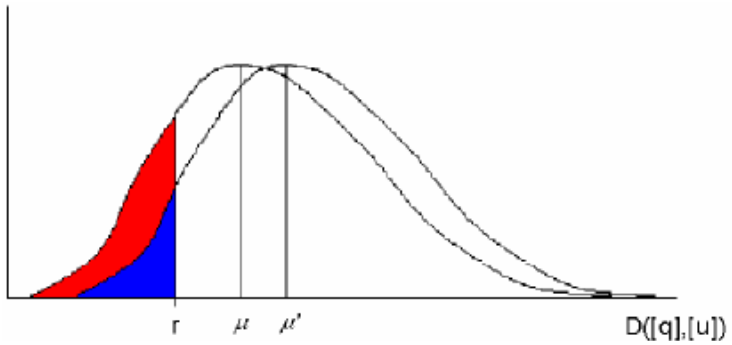


5.3.2 Técnicas de selección de pivotes

- Condición de exclusión en espacio de pivotes

$$|\delta(p_i, u) - \delta(p_i, q)| > r \Rightarrow \Delta_{p_1,\ldots,p_k}([q],[u]) > r$$

Distribución de Δ



Criterio de eficiencia para comparar dos conjuntos de pivotes:

$$\mu'_{\Delta_{p'_1,\ldots,p'_k}} > \mu_{\Delta_{p_1,\ldots,p_k}}$$

5.3.2 Técnicas de selección de pivotes

- Estimación de μ_Δ
  - Elegir un conjunto A de pares de puntos
  - Calcular μ_Δ para el *i*-ésimo par
  - Media de distancias Δ_i se estima como

$$\mu_\Delta = \frac{1}{A}\sum_{i=1}^{A}\Delta_i$$

  - Costo para estimar μ_Δ : $2kA$ cálculos de δ

5.3.2 Técnicas de selección de pivotes

- Selección de N grupos aleatorios
  - Elegir *N* grupos de *k* pivotes aleatorios
  - Estimar μ_Δ para para cada grupo
  - Escoger el grupo que maximiza μ_Δ
  - Costo de optimización : $2kAN$ cálculos de δ

5.3.2 Técnicas de selección de pivotes

- Selección incremental
  - Elegir un pivote de una muestra de *N* objetos aleatorios, que maximice la media de distancias Δ e
  - Elegir un segundo pivote de una muestra de N objetos aleatorios, tal que el conjunto maximice la media de dis ncias de Δ
  - Repetir hasta que se elijan *k* pivotes
  - Costo de optimización: *2kAN* cálculos de δ
- Ventajas
  - Permite agregar pivotes sin tener que rehacer todo el trabajo de optimización

5.3.2 Técnicas de selección de pivotes

- Selección de óptimo local
  - Elegir *k* pivotes aleatorios
  - Calcular matriz *M*

$$M_{A\times k} = \Delta_{p_j}([a_r],[a'_r]),\ 1 \le r \le A,\ 1 \le j \le k$$

  - Se sigue que

$$\Delta([a_r],[a'_r]) = \max_{j=1}^{k} M[r,j]$$

5.3.2 Técnicas de selección de pivotes

- Selección de óptimo local
  - Contribución del pivote p_j : suma sobre las A filas de cuánto ayuda p_j a incrementar el valor de Δ
    - r_{max} : índice del pivote con el valor máximo en esa fila
      - r_{max2} : índice del pivote con el segundo valor

$$\text{contribución} = \begin{cases} M[r, r_{max}] - M[r, r_{max2}] & \text{si } j = r_{max} \\ 0 & \text{si no} \end{cases}$$

5.3.2 Técnicas de selección de pivotes

- Selección de óptimo local
  - Pivote con contribución mínima es la víctima
  - Se reemplaza, si es posible, por un pivote elegido de una muestra de *X* objetos
  - Proceso se repite *N'* veces
  - Costo de optimización:
    - Construcción de M: *2Ak* cálculos de δ
    - Calcular víctima: 0
    - Reemplazar víctima: *2AX* cálculos de δ
    - Costo total: *2A(k + N'X)* cálculos de δ

5.3.2 Técnicas de selección de pivotes

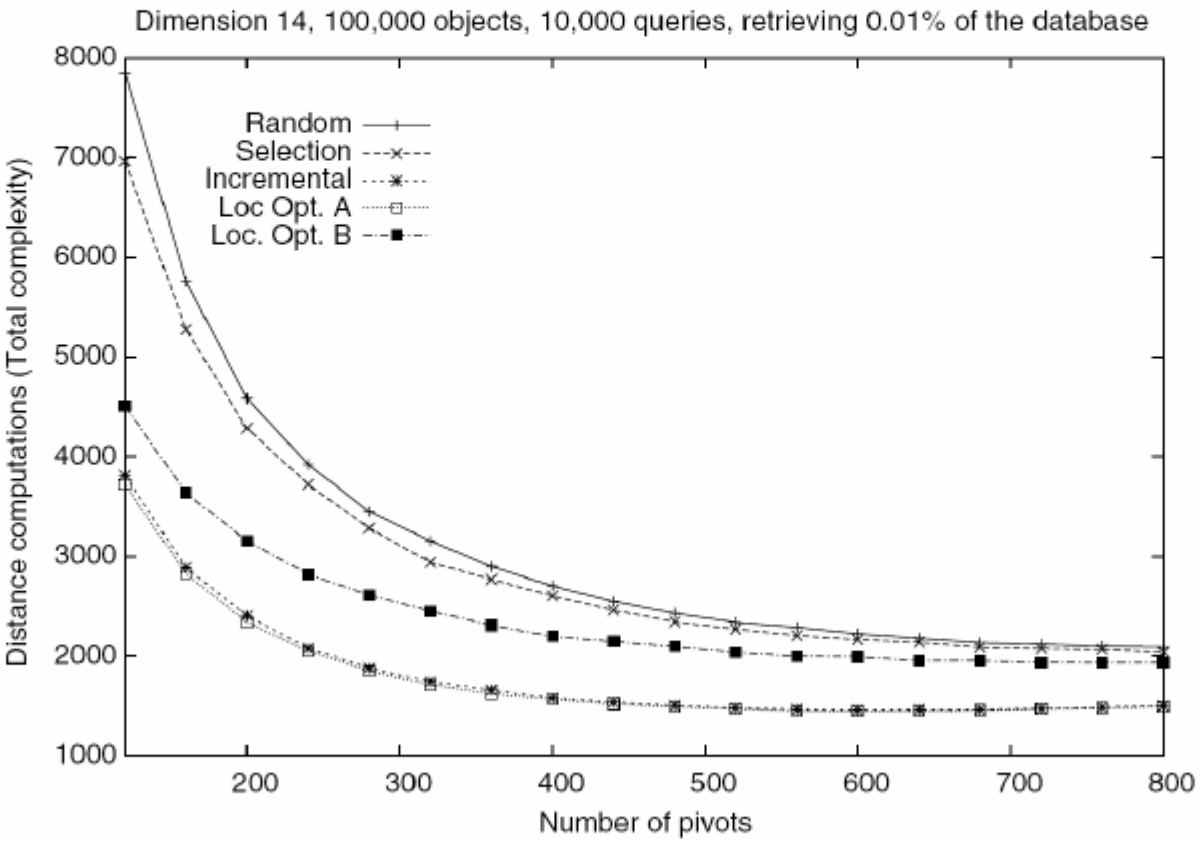
- Selección de óptimo local
  - Si $kN = k + N'X$ (es decir, $N'X=k(N-1)$), costo = $2AKN$ cálculos de δ
  - Notar que se pueden intercambiar los valores de N' y X manteniendo el costo total de optimización
    - Óptimo local A: $N'=k$ $X= N - 1$
    - Óptimo local B: $N'=N - 1$ $X=k$

5.3.2 Técnicas de selección de pivotes

- Experimentalmente se encontró que
  - Es mejor tener buena estimación de μ_Δ (*A* debe ser muy grande)
  - Muestras pequeñas de objetos son suficientes para obtener buenos resultados (*N* puede ser pequeño)
- Pruebas experimentales
  - Con datos sintéticos
  - Métodos de selección encuentran conjuntos de “buenos pivotes”

5.3.2 Técnicas de selección de pivotes

- Resultados

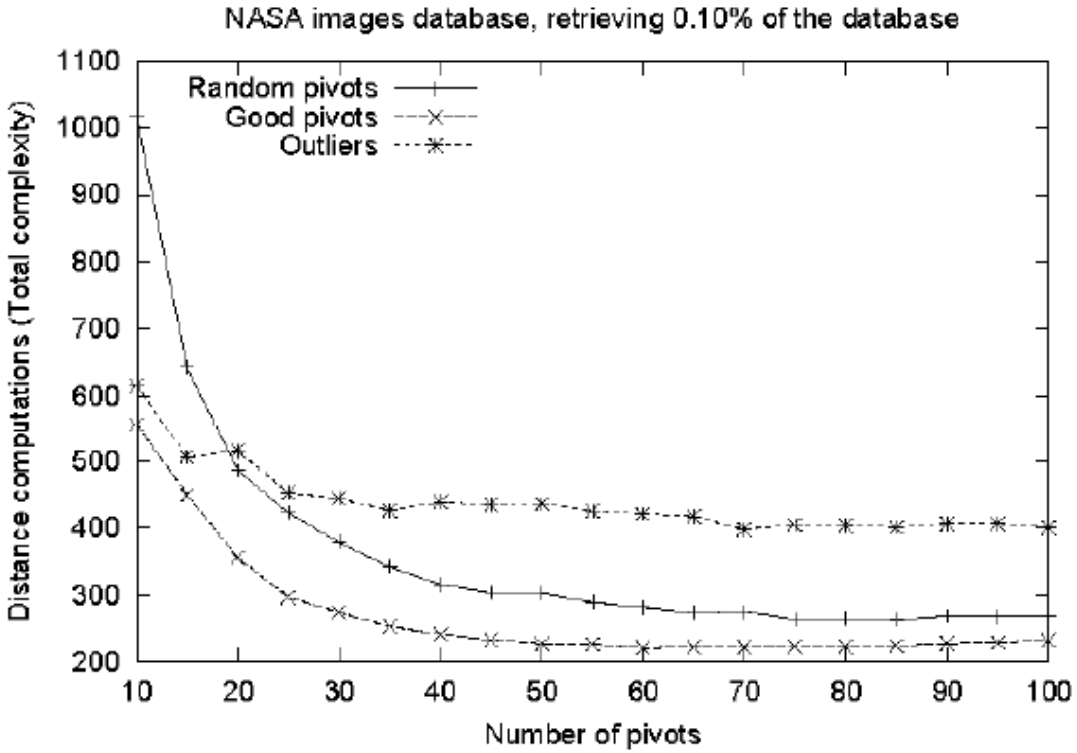


5.3.2 Técnicas de selección de pivotes

- Características de los “buenos pivotes”
  - Son objetos alejados del resto de la BD
  - Son objetos alejados entre sí
- Objetos con estas características se denominan *outliers*
  - Método alternativo de selección: elegir outliers
  - Producen mejores resultados de “buenos pivotes” en espacios sintéticos (distribución uniforme), funcionan mal con datos reales

5.3.2 Técnicas de selección de pivotes

- Resultados experimentales



Observaciones

- "Buenos pivotes" elegidos considerablemente mejores que pivotes aleatorios
- "Buenos pivotes" alcanzan la eficiencia óptima de los pivotes aleatorios con muchos menos pivotes
- Outliers usados como pivotes obtienen peores resultados que pivotes aleatorios

5.3.2 Técnicas de selección de pivotes

- Conclusiones
  - “Buenos pivotes” son outliers
  - No todo outlier es un buen pivote
  - Existen conjuntos de “buenos pivotes”
    - Se necesita una buena estimación de la media de distancias de 
    - No es necesario tener una muestra muy grande para escoger los pivotes
  - Tema de investigación en la actualidad

5.3.3 Otros índices basados en pivotes

- Árboles métricos basados en pivotes
  - Burkhard-Keller Tree
  - Fixed Queries Tree
  - Fixed-Height Queries Tree
  - Multi-Vantage Point Tree
- Representación del árbol en arreglos
  - Spaghettis
  - Fixed Queries Array
- Otras estructuras
  - Approximating and Eliminating Search Algorithm (AESA)
  - Linear AESA

5.4 Índices basados en particiones compactas

- Dividen el espacio en zonas o particiones lo más compactas posible
- Cada zona tiene un punto representativo *c* denominado “centro”
- Cada zona puede ser particionada recursivamente en más zonas, formando una jerarquía de búsqueda
- Hay dos criterios generales para descartar zonas:
  - Partición de Voronoi
  - Radio cobertor

5.4 Índices basados en particiones compactas

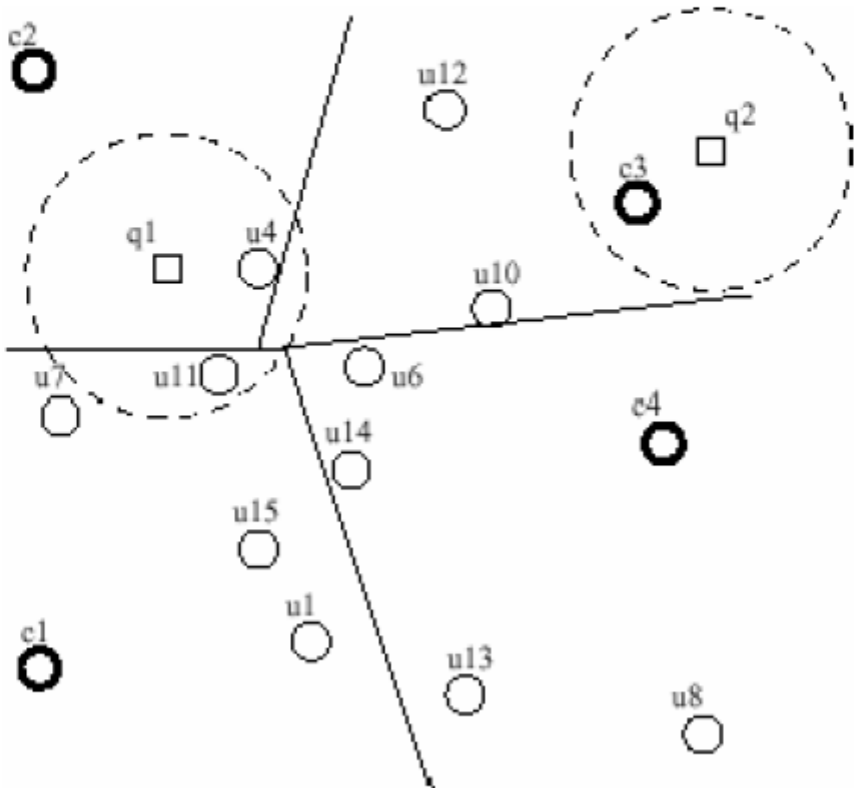
- Partición de Voronoi
  - Se eligen *m* centros (objetos de la BD)
  - El resto de los objetos se asigna a su centro más cercano (partición de Voronoi)
  - Dada una consulta por rango *(q,r)*, se calculan las distancias entre *q* y los *m* centros
  - Sea *c* el centro más cercano a *q*. Se pueden descartarlas zonas *i* cuyo centro cumpla con

$$\delta(q, c_i) > \delta(q, c) + 2r$$

  - Asegura que si hay intersección, entonces no se puede descartar

5.4 Índices basados en particiones compactas

- Ejemplo de partición de Voronoi



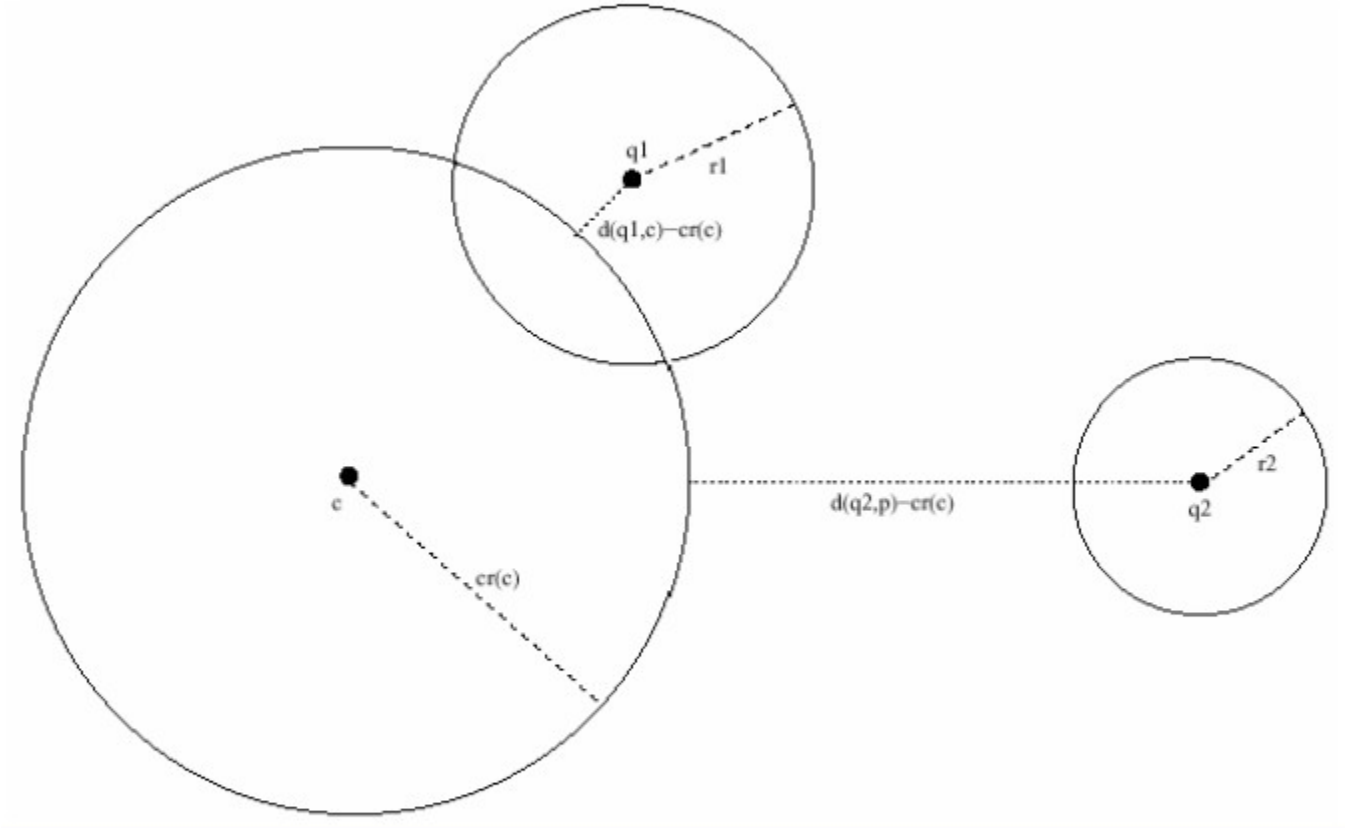
5.4 Índices basados en particiones compactas

- Criterio de radio cobertor
  - El radio cobertor *cr(c)* es la distancia máxima entre un centro *c* y algún objeto de su zona
  - Dada la consulta *(q,r)*, la zona de centro *c* no puede tener intersección con la bola de consulta si

$$\delta(q, c) - cr(c) > r$$

5.4 Índices basados en particiones compactas

- Ejemplo de radio cobertor



5.4 Índices basados en particiones compactas

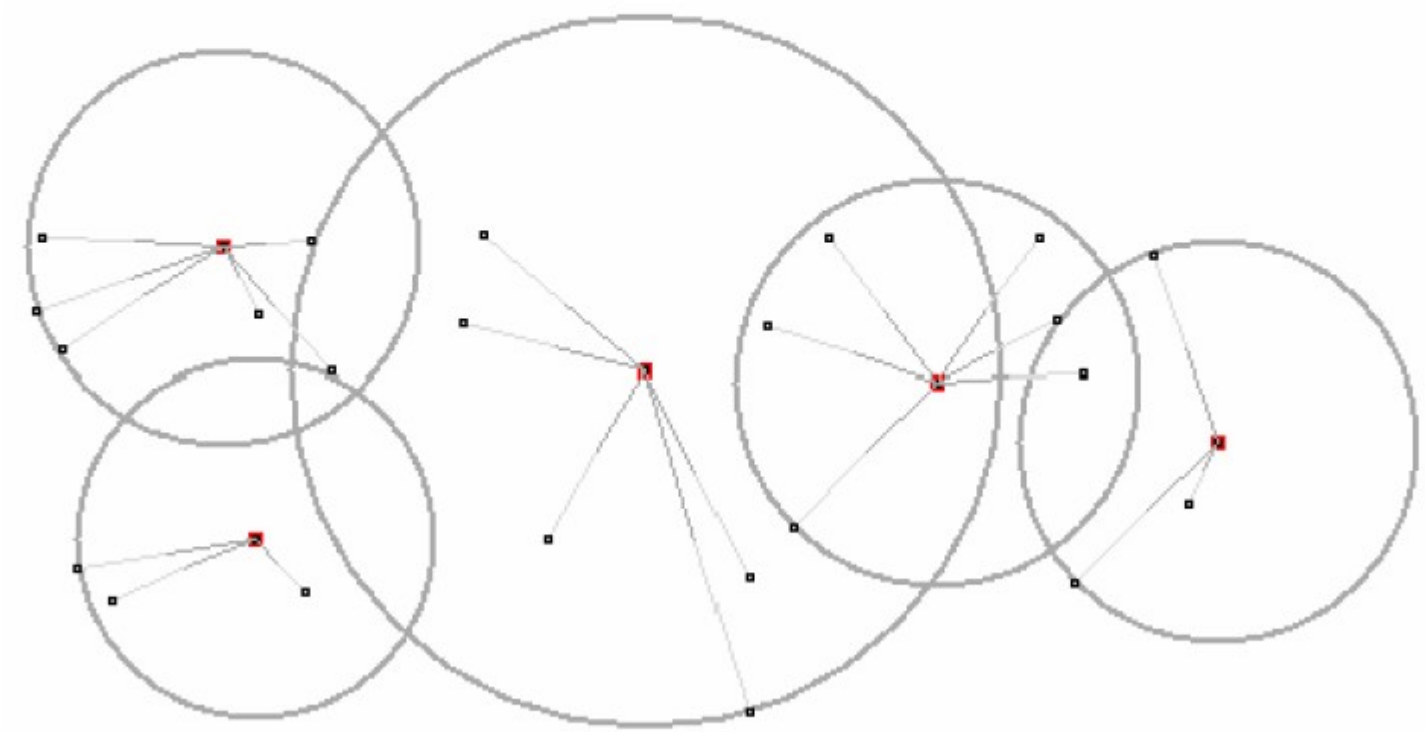
- Índices basados en particiones compactas
  - Partición de Voronoi
    - Generalized-Hyperplane Tree
  - Radio cobertor
    - Bisector Tree
    - Voronoi Tree
    - Monotonous BST
    - M-Tree
    - List of Clusters
  - Ambos criterios
    - Geometric Near-neighbor Access Tree
    - Spatial Approximation Tree

5.4.1 M-tree

- M-tree
  - Índice basado en el criterio de radio cobertor
  - Árbol balanceado
  - Objetivos
    - Índice dinámico (permite inserciones, eliminaciones y actualizaciones)
      - Buen desempeño de E/S
      - Pocos cómputos de distancia por consulta
      - Funcionamiento en memoria secundaria

5.4.1 M-tree

- Ejemplo de M-tree



5.4.1 M-tree

- Estructura del M-tree
  - Hojas (nodos externos)
    - Almacenan los objetos indexados
  - Nodos internos
    - Almacenan los *routing objects* (centros de esferas)
    - Por cada routing object O_r
      - *ptr(T(O_r))* apunta a la raíz del subárbol
        - *T(O_r)* árbol cubridor de O_r
        - *r(O_r)* radio cobertor de O_r
        - *P(O_r)* padre de O_r (indefinido para la raíz)

5.4.1 M-tree

- Nodos del M-tree

Nodos internos:

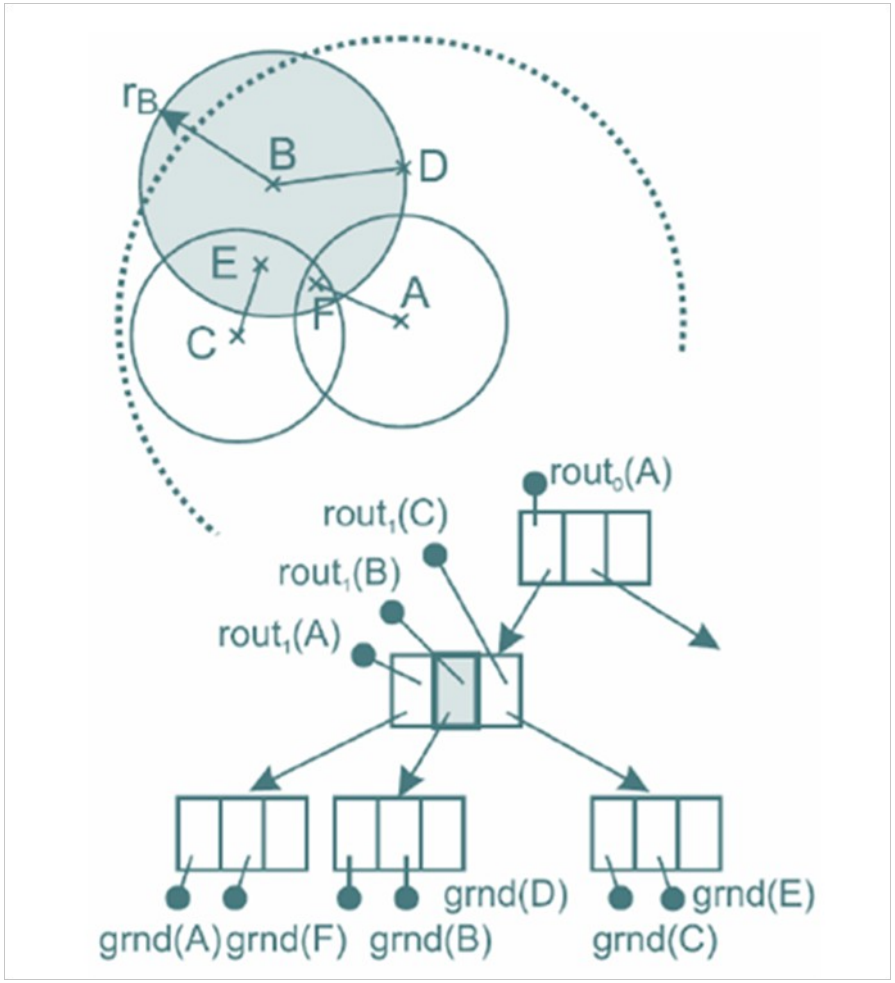
O_r	(feature value of the) routing object
$ptr(T(O_r))$	pointer to the root of $T(O_r)$
$r(O_r)$	covering radius of O_r
$d(O_r, P(O_r))$	distance of O_r from its parent

Nodos externos:

O_j	(feature value of the) DB object
$oid(O_j)$	object identifier
$d(O_j, P(O_j))$	distance of O_j from its parent

5.4.1 M-tree

- Ejemplo



5.4.1 M-tree

- Búsquedas: consulta por rango

```
RS(N:node, Q:query_object, r(Q):search_radius)
{ let O_p be the parent object of node N;
  if N is not a leaf
  then { ∀ O_r in N do:
         if |d(O_p,Q) − d(O_r,O_p)| ≤ r(Q) + r(O_r)
         then { Compute d(O_r,Q);
                if d(O_r,Q) ≤ r(Q) + r(O_r)
                then RS(*ptr(T(O_r)),Q,r(Q)); }}
  else { ∀ O_j in N do:
         if |d(O_p,Q) − d(O_j,O_p)| ≤ r(Q)
         then { Compute d(O_j,Q);
                if d(O_j,Q) ≤ r(Q)
                then add oid(O_j) to the result;}}}
```

5.4.1 M-tree

- Consulta por *k* vecinos más cercanos
  - Hjaltason y Samet
  - Dos colas de prioridad
    - *k* mejores candidatos hasta el momento
      - *Active Page List*
        - Nodos cuyo padre ya fu visitado pero ellos no han sido visitados aún
          - Ordenados por cota inferior de la distancia al objeto de consulta

5.4.1 M-tree

- Consulta por *k* vecinos más cercanos
  - Inicialmente: lista de candidatos vacía, APL sólo contiene la raíz del árbol
  - Se saca un nodo de la APL
    - Medir distancia al centro y agregarlo a la lista de candidatos si es necesario (menos de *k* candidatos o distancia al centro menor que al *k*-ésimo candidato)
      - Insertar todos los hijos a la APL
  - Terminar si APL vacía o mínima cota inferior de un nodo mayor que distancia a *k*-ésimo candidato, sino iterar

5.4.1 M-tree

- Algoritmo de inserción

```
Insert(N:node, entry(O_n):M-tree_entry)
{ let 𝒩 be the set of entries in node N;
  if N is not a leaf then
  { let 𝒩_in = entries such that d(O_r, O_n) ≤ r(O_r);
    if 𝒩_in ≠ ∅
    then let entry(O*_r) ∈ 𝒩_in : d(O*_r, O_n) is minimum;
    else { let entry(O*_r) ∈ 𝒩:
                        d(O*_r, O_n) − r(O*_r) is minimum;
           let r(O*_r) = d(O*_r, O_n); }
    Insert(*ptr(T(O*_r)), entry(O_n)); }
  else /* N is a leaf */
  { if N is not full
    then store entry(O_n) in N
    else Split(N, entry(O_n)); }}
```

5.4.1 M-tree

- Split de nodos

```
Split(N:node; E:M-tree_entry)
{ let 𝒩 = entries of node N ∪ {E};
  if N is not the root then
    let O_p be the parent of N, stored in N_p node;
  Allocate a new node N';
  Promote(𝒩, O_p1, O_p2);
  Partition(𝒩, O_p1, O_p2, 𝒩_1, 𝒩_2);
  Store 𝒩_1's entries in N and 𝒩_2's entries in N';
  if N is the current root
  then { Allocate a new root node, N_p;
         Store entry(O_p1) and entry(O_p2) in N_p; }
  else { Replace entry(O_p) with entry(O_p1) in N_p;
         if node N_p is full
         then Split(N_p, entry(O_p2))
         else store entry(O_p2) in N_p; }}
```

Promote escoge dos routing objects para ser insertados en el Nodo padre

Partition divide las entradas en el nodo que hizo overflow en dos subconjuntos disjuntos

5.4.1 M-tree

- Nuevos radios cobertores después de un split
  - Si es una hoja

$$r(O_{p_1}) = max\{d(O_j, O_{p_1}) | O_j \in \mathcal{N}_1\}$$

    - Distancia entre el centro y su objeto mas lejano
  - Si es un nodo interno

$$r(O_{p_1}) = \max\{d(O_r, O_{p_1}) + r(O_r) | O_r \in \mathcal{N}_1\}$$

    - D hijos mas radio del cobertor del hijo

5.4.1 M-tree

- Métodos *Promote* y *Partition*
  - Definen una política de split
  - Objetivos
    - Minimizar “volumen”
    - Minimizar “solapamiento”
  - *Promote*
    - Elegir el padre del nodo como uno de los routing objects *(confirmed split policy)*
    - Aleatoriamente
    - Minimizar suma de radios cobertores resultantes

5.4.1 M-tree

- Métodos *Promote* y *Partition*
  - *Partition*
    - t *Generalized Hyperplane*: asignar cada objeto al routing object más cercano
      - *Balanceado:* repartir la mitad de los objetos a cada routing object (criterio: distancia a los routing objects)
  - Generalized Hyperplane funciona mejor en la práctica para asegurar solapamiento mínimo, aunque no se garantiza CERO solapamiento

5.4.1 M-tree

- Conclusiones
  - Índice métrico dinámico
  - Performance depende de la política de split utilizada
  - Si se conocen los objetos de la BD de antemano: *bulk loading*
  - Se sigue realizando investigación para extender capacidades de este índice

5.4.2 List of Clusters

- Basado en el criterio de radio cobertor
- Árbol binario desbalanceado
  - Hijo izquierdo siempre contiene datos
  - Hijo derecho contiene árbol con resto de los objetos
  - Último nodo sólo tiene hijo izquierdo

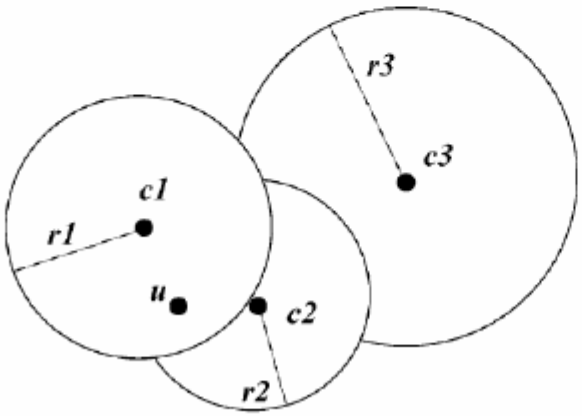
5.4.2 List of Clusters

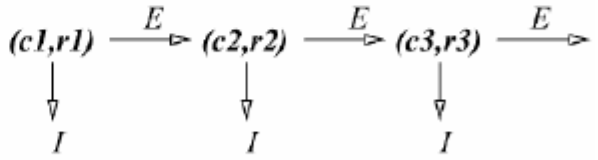
- Estructura
  - Se elige un objeto como centro y un radio cobertor *r*
  - Se inserta en la región de *c* todos los objetos a distancia menor o igual que *r* de *c*
    - Objetos en la región: forman nodo izquierdo
      - Resto del espacio: nodo derecho
  - Se elige otro centro y se repite el proceso hasta que no queden más objetos
  - Orden de las regiones en la lista es importante
    - Si hay traslape, el objeto pertenece al primer nodo de la lista que cubra la posición donde está

5.4.2 List of Clusters

- Algoritmo de construcción

```
Build (𝕌)
    if 𝕌 = ∅ then return an empty list
    Select c ∈ 𝕌
    Select a radius r_c
    I  ⟵  {u ∈ 𝕌 − {c},  d(c,u) ≤ r_c}
    E  ⟵  𝕌 − I
    return (c, r_c, I):Build(E)
```

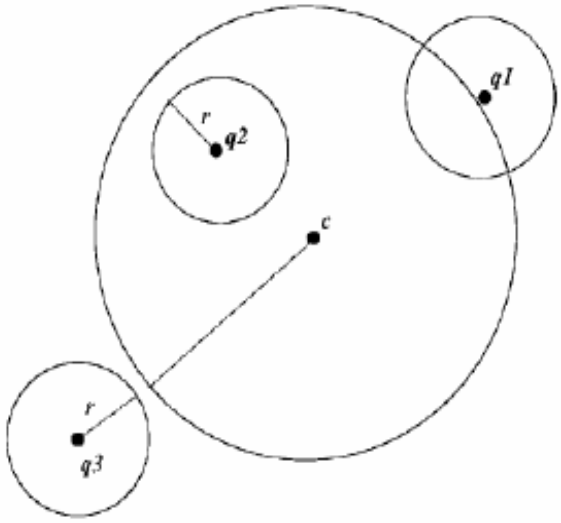




5.4.2 List of Clusters

- Algoritmo de búsqueda por rango

```
Search (L,q,r)
    if L is empty then return
    Let L = (c, r_c, I) : E
    Compute d(c,q)
    if d(c,q) ≤ r then add c to the list of results
    if d(c,q) ≤ r_c + r then search I exhaustively
    if d(c,q) > r_c − r then Search (E,q,r)
```



5.4.2 List of Clusters

- Selección del centro
  - Existen distintas opciones
  - Mejores resultados empíricos: objeto que maximice la suma de distancias a los centros previos
- Selección del radio
  - Radio fijo o número de objetos por región fijo
  - Mejores resultados: número de objetos fijo

5.4.2 List of Clusters

- Conclusiones
  - Índice simple, desbalanceado
  - En la práctica funciona bien en espacios métricos con dimensión intrínseca alta
    - Orden sublineal