BUSQUEDA DE LOS M-VECINOS MAS PROXIMOS EN EL ARBOL-BD

O. SANTANA, M. DIAZ, O. MAYOR, J. GONZALEZ
E.U. Informática, Universidad Politécnica de Canarias
Aptdo. 550, Las Palmas de Gran Canaria, España

El presente estudio está dedicado a desarrollar un esquema para las recuperaciones de los m-vecinos más próximos en la estructura de Árbol-BD. La primera parte del trabajo lleva a cabo el planteamiento algorítmico para este tipo de recuperaciones. En ella, se han desarrollado una serie de tests específicos para la estructura, sobre todo en cuanto a los tests de solapamiento. El clásico test de solapamiento con la esfera ha sido desdoblado en dos, uno de solapamiento con el interior de una zona y otro con el exterior. Así mismo, ha sido necesario el desarrollo de otro test adicional para detectar los solapamientos con el exterior en los descensos alternativos (exterior de interior).

La segunda parte del artículo está dedicada al estudio experimental de la estructura ante el esquema de búsqueda de los m-vecinos más próximos. Previamente se justifica experimentalmente el uso del test de solapamiento exterior de interior. Otra disyuntiva se plantea en el uso de dos tipos alternativos de distancias. Se opta por la utilización paralela de ambas, a lo largo de todo el proceso de estudio, mostrándose una de ellas superior en todo momento, para este tipo de estructura. Los parámetros fueron medidos frente a variaciones del tamaño de celda, la dimensionalidad y el número de vecinos a localizar.

BUSQUEDA DE LOS M-VECINOS MAS PROXIMOS EN EL ARBOL-BD

O. SANTANA, M. DIAZ, O. MAYOR, J. GONZALEZ
E.U. Informática, Universidad Politécnica de Canarias
Aptdo. 550, Las Palmas de Gran Canaria, España

Resumen:

En este trabajo se presenta la petición de los m_vecinos más próximos en la estructura de árbol_BD. Se introduce un tipo de solapamiento denominado solapamiento exterior en los descensos alternativos (exterior de interior), que mejora el tiempo de respuesta. Además se muestra experimentalmente como influye la dimensionalidad, el tamaño de la celda, y el número de vecinos más próximos en dicho tiempo.

0) INTRODUCCION

A menudo, es necesario recuperar información de una base de datos que está localizada en un espacio nd_dimensional. El campo de aplicación de estas estructuras de ficheros no se queda solo en las bases de datos. La capacidad de recuperación multiclave es especialmente requerida por aplicaciones que abarcan áreas como la inteligencia artificial, procesamiento de imágenes, robótica, cartografía, reconocimiento de formas, estadística, y recuperación de información, entre otras.

Una de los tipos de recuperación de información que tiene más posibilidades de generar campos de aplicación, es la búsqueda de los m_vecinos más próximos. Este artículo está dedicado al planteamiento algorítmico de este tipo de búsqueda [3] y a la realización del estudio experimental del mismo.

La estructura de árbol_BD utilizada en el estudio procede básicamente de la propuesta por [4] y optimizada por [1,2,5]. Para ella, se ha desarrollado el esquema de búsqueda de los m_vecinos más próximos que se presenta en la sección 1.

La sección 2 presenta el estudio experimental. Plantea previamente unas comparaciones a fin de escoger el esquema de búsqueda adecuado, y a continuación se realizan las medidas paramétricas necesarias para mostrar la respuesta de la estructura a este tipo de requerimientos. Las conclusiones de este artículo aparecen en la sección 3.

1) BUSQUEDA DE LOS M-VECINOS MAS PROXIMOS

Este tipo de búsqueda consiste básicamente en, dados un espacio nd_dimensional, un punto del mismo, y una función distancia, recuperar los m_registros más próximos al punto dado.

Para realizar esta búsqueda, se parte del nodo cabeza y se recorre el árbol de forma recursiva hasta alcanzar una celda, análogamente al proceso de búsqueda exacta. Si el punto es localizado en la celda, finaliza el proceso. En caso contrario, se insertan en una lista los m_elementos de la celda más próximos al punto, siempre que en la celda existan más de m_elementos, si no se insertan todos. A continuación se recorre el árbol retrocediendo recursivamente en el camino de bajada inicial, probando a la vez otras alternativas de descenso, hasta obtener una lista con los m_vecinos más próximos de todo el árbol.

1.1 Lista de los m-vecinos más próximos

En el recorrido inicial por el árbol se llega a una celda (zona donde pudiese estar el punto), y se testea si algún registro de la misma coincide con

la petición.

Una vez que se ha comprobado que el punto no existe, se comienza a crear la lista de forma ordenada con los registros existentes en la celda. Sus elementos están clasificados en el sentido de mayor a menor distancia al punto. La función que evalúa dicha distancia es la siguiente:

```
función DISTANCIA (ec)
{ec elemento de celda}
sum=0
para j = 1 hasta nd hacer
     sum=sum + dist (p(j),ec(j))
fin para
DISTANCIA=F(sum)
retornar
```

Las funciones dist y F dependen de la métrica utilizada.

Los m primeros puntos probados se insertan en la lista de forma ordenada sin ninguna restricción. Posteriormente, cualquier otro punto sondeado pasará a formar parte de la lista si su distancia al punto es menor que la correspondiente del primer elemento, eliminando a este último e insertándose en la posición adecuada según la ordenación.

Por lo tanto, la inclusión o no de un punto en la lista, viene determinada por el radio de la esfera, considerado como la distancia del punto al vecino más alejado que pertenezca a la lista. El acceso a las restantes celdas se realiza retrocediendo recursivamente por el recorrido de bajada inicial, probando otras alternativas de descenso utilizando el criterio de los solapamientos de la esfera.

1.2 Solapamientos de la esfera

Cuando se retrocede recursivamente en el camino de bajada inicial se prueba la contención de la esfera actual con la zona, siempre que se llegue a ella por el interior. Si se verifica la contención, finaliza el proceso, si no, se prueba el solapamiento de la esfera con el exterior de dicha zona y si se verifica, se desciende recursivamente; en caso contrario se sigue retrocediendo. Si se llega a la zona por el exterior, se prueba el solapamiento con el interior. Si se verifica, se desciende recursivamente y si no se sigue retrocediendo.

1.2.1 Solapamientos de la esfera con el interior de una zona

En el estudio del solapamiento de la esfera con el interior de una zona se distinguen dos casos:

a) El punto p, está situado fuera de la zona.

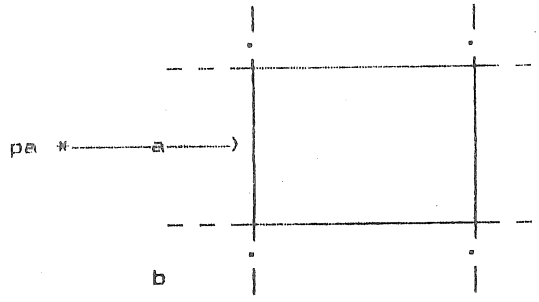

Figura 1

Según se muestra en la figura 1, para posiciones análogas a pa se determina la distancia a y para posiciones análogas a pb se determina la distancia b. En cualquier caso, si el radio de la esfera es superior a dichas distancias hay solapamiento. La función que lo verifica es la siguiente:

```
función SOLESFIN_BD (liz(j),lsz(j):j=i..nd)
{liz,lsz son los límites de la zona sobre la que se quiere sondear el
 solapamiento de la esfera}
sum=0
para k=1 hasta nd hacer
     si p(k) ( liz(k) entonces
        sum=sum + dist (p(k),liz(k))
        si F(sum) ) radio entonces SOLESFIN_BD=falso; retornar fin si
                    si no
        si p(k) ) lsz(k) entonces
           sum=sum + dist (p(k),lsz(k))
           si F(sum) ) radio entonces SOLESFIN_BD=falso; retornar fin si
        fin si
     fin si
fin para
SOLESFIN_BD=verdadero; retornar
```

b) El punto es interior a la zona (figura 2).

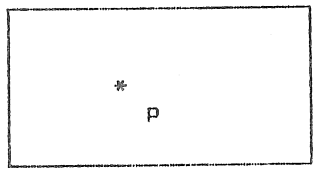

Figura 2

El solapamiento es obvio en este caso. El problema consiste en determinar si la zona contiene a la esfera. Para ello se calculan las distancias del punto a los límites de la zona. Si alguna de ellas es inferior al radio no hay contención y en caso contrario sí la hay. La función que lo verifica es la siguiente:

```
función CONESF_BD (liz(j),lsz(j),j=1..nd)
{liz, lsz son los límites de una zona que incluye a la petición}
para k=1 hasta nd hacer
     si liz(k) <> domin(k) entonces
        si dist (p(k),liz(k)) < radio entonces
           CONESF_BD=falso; retornar
        fin si
     fin si
     si lsz(k) <> domax(k) entonces
        si dist (p(k),lsz(k)) < radio entonces
           CONESF_BD=falso; retornar
        fin si
     fin si
fin para
CONESF_BD=verdadero; retornar
```

1.2.2 Solapamientos de la esfera con el exterior de una zona

Una clasificación de este tipo de solapamiento debe considerar tres zonas: la zona del punto, Zp, aquella sobre la que se determinó el solapamiento interior, Z, y la zona más pequeña que contenga a esta última, CZ. Todas las posibles situaciones que se pueden presentar se estudian a continuación.

a) Zp está incluida en Z, y ésta a su vez, lo está en CZ.

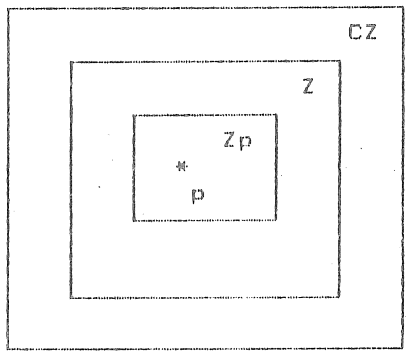

Figura 3a

Respecto a la situación mostrada en la figura 3a, siempre que Z no contenga a la esfera, hay solapamiento con el exterior de Z, si Z y CZ no tienen límites comunes. Si hay algún límite común, como en la figura 3b, es necesario medir las distancias a los límites no comunes. Si alguna de ellas es inferior al radio de la esfera hay solapamiento. La función que realiza el proceso se presenta a continuación.

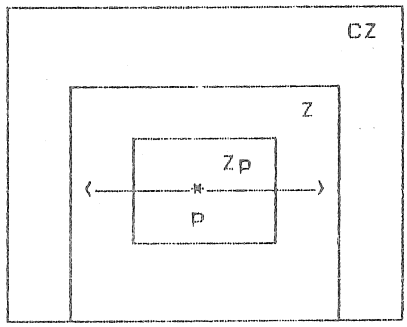

Figura 3b

```
función SOLESFEX_BD (liz(j),lsz(j),cliz(j),clsz(j):j=1..nd)
{liz, lsz son los límites de la zona Z}
{cliz, clsz son los límites de la zona CZ}
para k=1 hasta nd hacer
     si liz(k) <> cliz(k) entonces
        si dist (p(k),liz(k)) < radio entonces
           SOLESFEX_BD=verdadero; retornar
        fin si
     fin si
     si lsz(k) <> clsz(k) entonces
        si dist (p(k),lsz(k)) < radio entonces
           SOLESFEX_BD=verdadero; retornar
        fin si
     fin si
fin para
SOLESFEX_BD=falso; retornar
```

b.1) CZ incluye a Z, y Zp es una zona exterior a CZ.

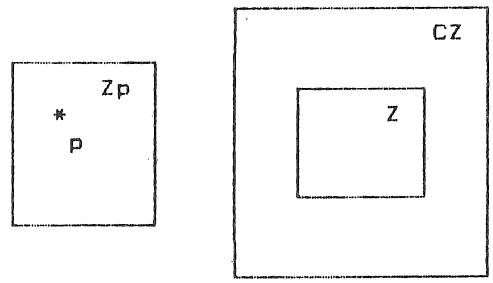

Figura 4a

b.2) CZ incluye a Z, y Zp es mimplemente el exterior de CZ.

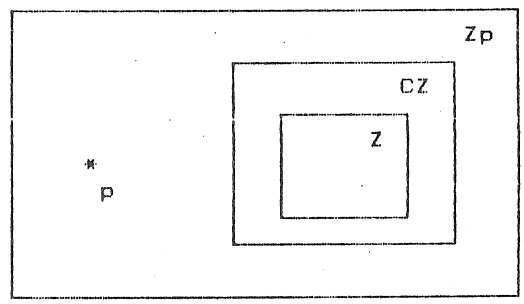

Figura 4b

Para determinar el solapamiento de la esfera con la parte exterior de la zona hay que tener en cuenta la posibilidad de que existan límites comunes entre las zonas Z y CZ.

En el caso de la figura 4a, siempre que Z y CZ no tengan ningún límite común, o en el caso de tenerlo, si éste no está hacia la parte exterior en la que se encuentra el punto, como se muestra en la siguiente figura 4c, el solapamiento está garantizado.

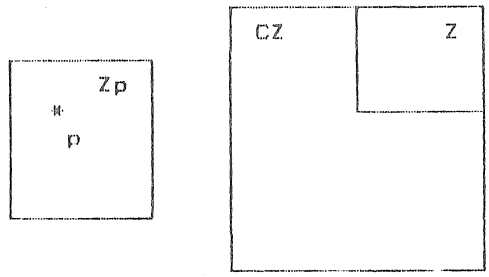

Figura 4c

En cualquier otro caso se presentan situaciones análogas a las de la figura 4d.

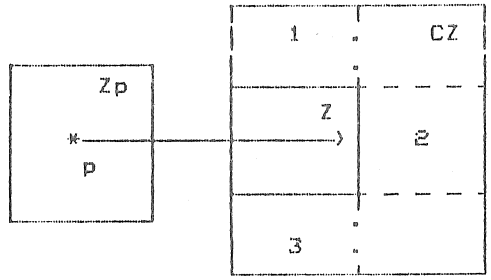

Figura 4d

La existencia de solapamiento se verifica dividiendo la parte exterior de Z en subzonas como se indica en la figura 4d, y estudiando el solapamiento interior a cada una de ellas a través de la función SOLESFIN_BD. Alguna de las zonas puede no existir, debido a que Z y CZ tengan límites comunes.

El estudio del solapamiento en el caso de la figura 4b es idéntico al de la figura 2a. La situación para límites comunes se presenta en la figura 4e.

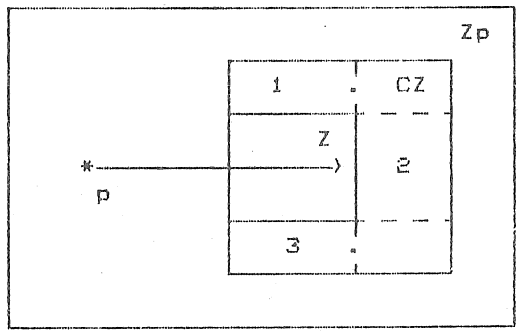

Figura 4e

c) Zp y Z no se contienen y están ambas incluídas en CZ.

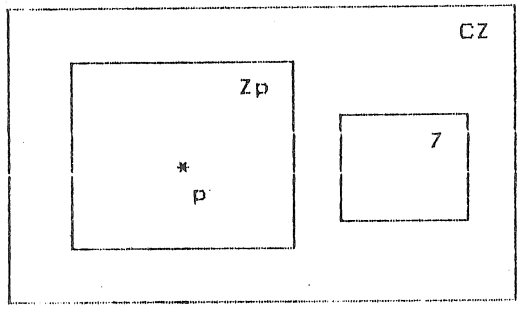

Figura 5a

En el caso de la figura 5a, siempre que Z no tenga límites comunes con Zp, el solapamiento está garantizado; si esto no ocurre, se presentan situaciones análogas a las de la figura 5b.

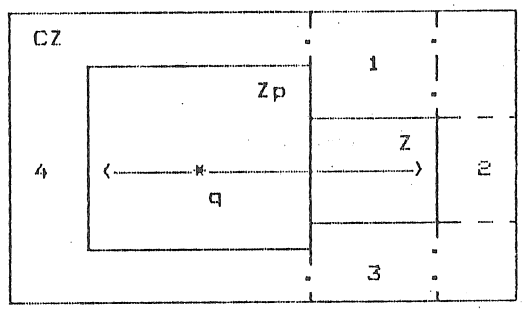

Figura 5b

Una estrategia para verificar la existencia de solapamiento consiste en dividir la parte exterior de Z en subzonas como se indica en la figura 3b, y sondear el solapamiento interior con 1, 2 y 3, a través de la función SOLESFIN_BD, y el solapamiento exterior en 4, a través de la función SOLESFEX_BD.

1.3 Alternativas de descenso

Se realiza el recorrido de los subárboles alternativos al camino de bajada inicial. El descenso en cada subárbol se hace de forma recursiva, probando en cada nivel el solapamiento con el interior. Si se verifica se desciende por la rama interior y cuando retorna se prueba el solapamiento con el exterior de ese interior, para decidir si se desciende por la rama exterior o si se retorna al nivel de recursividad anterior. En caso contrario, el recorrido continúa por el exterior sin necesidad de probar el solapamiento correspondiente, ya que el procedimiento fue invocado, la primera vez, al existir solapamiento con la parte del árbol que se está rastreando.

```
procedimiento DA_BD (nodo,lizI(j),lszI(j),cliz(j),clsz(j):j=1..nd)
{lizI, lszI son los limites de ZI}
{cliz, clsz son los limites de la zona a que se desciende}
si nodo es celda entonces
   LVMP (nodo)    {realiza el mantenimiento de la lista}
                  si no
   liz(j)=cliz(j) j=1..nd
   lsz(j)=clsz(j) j=1..nd
   CALCULO_LIMITES_DE_ZONA (liz(j),lsz(j):j=1..nd)
   si SOLESFIN_BD (liz(j),lsz(j):j=1..nd) entonces
      DA_BD (hijoizdo de nodo,lizI(j),lszI(j),liz(j),lsz(j):j=1..nd)
     si SOLESFEXIN_BD (liz(j),lsz(j),cliz(j),clsz(j):j=1..nd) entonces
        DA_BD (hijodcho de nodo,lizI(j),lszI(j),cliz(j),clsz(j):j=1..nd)
      fin si
                                              si no
      DA_BD (hijodcho de nodo,lizI(j),lszI(j),cliz(j),clsz(j):j=1..nd)
 fin si
fin si
retornar
```

1.4 Procedimiento de búsqueda de los m-vecinos más próximos

Este procedimiento realiza el recorrido inicial descendiendo recursivamente por el árbol hasta alcanzar la celda donde pudiese estar el punto. En cada retorno al nivel de recursividad anterior, se prueba la posibilidad de descenso alternativo a través de los solapamientos con el exterior o el interior, según la rama opuesta al descenso inicial, con el fin de actualizar la lista de los m_vecinos más próximos.

```
procedimiento BUSVMP_BD (nodo,zi(j),zs(j):j=1..nd)
{nodo  nodo interno testeado actualmente}
{zi, zs son los limites de la zona, dentro de la cual está la petición}
nzi(j)=zi(j) j=1..nd
nzs(j)=zs(j) j=1..nd
CALCULO_DE_LOS_LIMITES_DE_ZONA (nzi(j),nzs(j):j=1..nd)
si la petición es interior a la zona de nodo entonces
   si hijoizdo de nodo es celda entonces
      LVMP (hijoizdo de nodo)   {realiza el mantenimiento de la lista}
                                si no
      BUSVMP_BD (hijoizdo de nodo,nzi(j),nzs(j):j=1..nd)
   fin si
   si CONESF_BD (nzi(j),nzs(j):j=1..nd) entonces
      retornar al programa principal
                                          si no
      si SOLESFEX_BD (nzi(j),nzs(j),zi(j),zs(j):j=1..nd) entonces
         DA_BD (hijodcho de nodo,nzi(j),nzs(j),zi(j),zs(j):j=1..nd)
      fin si
   fin si
                                                si no
   si hijodcho de nodo es celda entonces
      LVMP (hijodcho de nodo)
                                si no
      BUSVMP_BD (hijodcho de nodo,zi(j),zs(j):j=1..nd)
   fin si
   si SOLESFIN_BD (nzi(j),nzs(j):j=1..nd) entonces
      DA_BD (hijoizdo de nodo,zi(j),zs(j),nzi(j),nzs(j):j=1..nd)
   fin si
fin si
retornar
```

2) ESTUDIO EXPERIMENTAL DE LA PETICION DE LOS M-VECINOS MAS PROXIMOS

El tiempo de respuesta a una petición de los m_vecinos más próximos se mide en función de los siguientes parámetros:

a) número de nodos internos accedidos, NIA.
g) número de celdas visitadas, CV.
j) número de distancias calculadas, DISTC.
k) número de solapamientos y contenciones estudiados.

Las pruebas se realizaron con distintos árboles de 10.000 registros cada uno, teniendo en cuenta las variaciones de dimensionalidad, nd, y tamaño de la celda, tmc.

Las peticiones se generaron de forma aleatoria, con la condición de que no existiesen en el árbol. Inicialmente, para cada árbol se realizaron búsquedas de 5_vecinos más proximos, con las distancias euclídea e infinito y se estudiaron posteriormente los resultados variando este número de vecinos, según los valores 10, 15, 20 y 25. En cada caso se realizaron 2000 pruebas y todos los resultados se dan como promedio.

Una de las cuestiones planteadas era la utilidad de un test de solapamiento exterior de interior. El test decide si se debe bajar por el exterior y por lo tanto ahorra recorrido por el árbol, pero a costa de cálculos adicionales en el nodo. En la figura 6 se muestran los resultados de los tiempos promedios por petición de 5_vecinos más próximos para la comparación de la búsqueda con, los test de solapamiento interior, exterior y exterior de interior (I+E+EI), y sin el test exterior de interior (I+E), con distancia euclídea, frente a la variación de tmc. Obteniéndose un mejor resultado para la primera, debido al ahorro progresivo de distancias calculadas y a la disminución de la diferencia en tiempos de solapamientos y contenciones, a medida que aumenta tmc, como se observa en las figura 7.

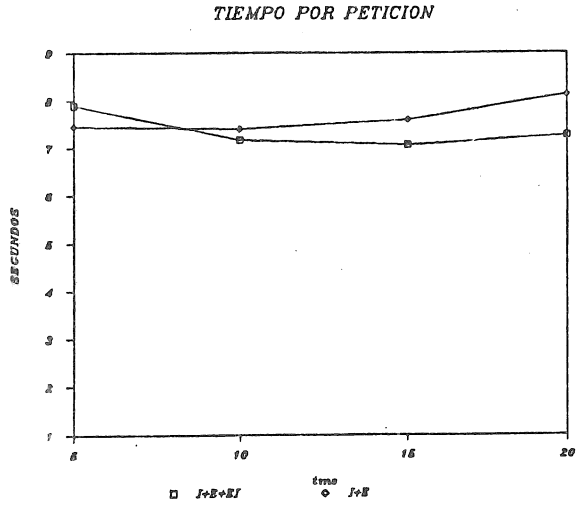

Figura 6

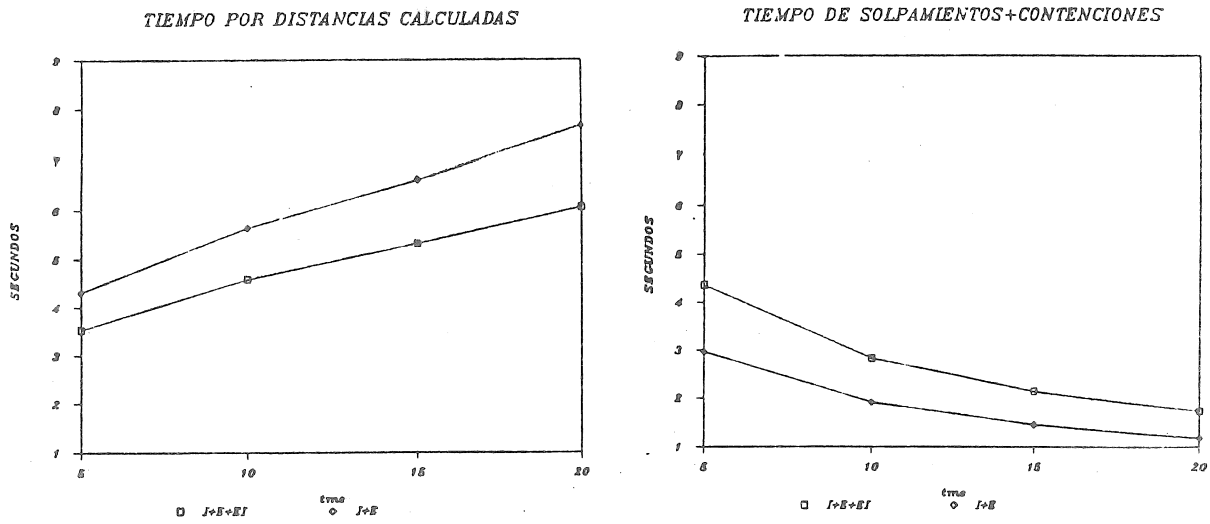

Figura 7

En un espacio nd_dimensional con la distancia euclídea, el conjunto de puntos que distan de uno fijo menos que un radio dado, geométricamente representan una hiperesfera; si la distancia considerada es la infinita se tendría un hipercubo. A igual radio la hiperesfera está inscrita en el hipercubo.

- En la búsqueda de los vecinos más próximos se obtienen obviamente hipercubos que corresponden a un radio menor o igual que el de las hiperesferas. Por ejemplo para dimensión dos se comprueba en la figura 8 que la relación que existe entre los radios es: reuclídea/raiz(2) (= rinfinito (= reuclídea. Los resultados experimentales, figura 9, muestran una mayor proximidad de rinfinito a reuclídea/raiz(2).

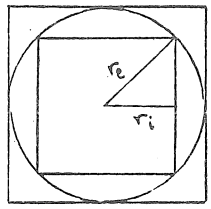

Figura 8

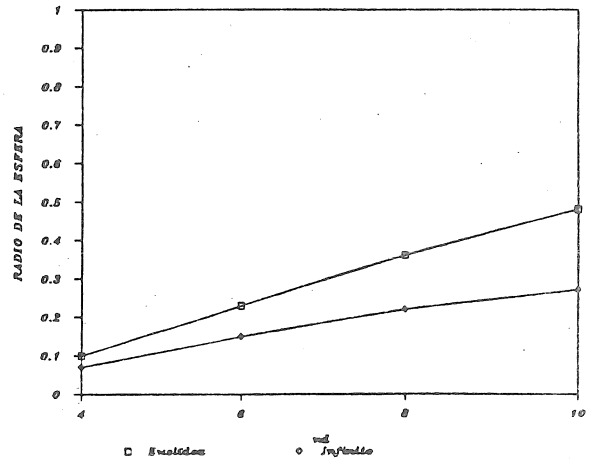

Figura 9

Esta relación entre los radios indica que el volumen del hipercubo es menor que el de la hiperesfera. Por otra parte, el hipercubo es más adaptable a la división del espacio en zonas, que se realiza en el árbol_BD. Esto explica el mejor comportamiento de la distancia infinito frente a la euclídea.

- Dado tmc, para cada métrica, el incremento de la dimensionalidad implica un incremento, figuras 10 y 11, en el número de nodos internos accedidos y en el de celdas visitadas, ya que para cada una de ellas aumenta el número de colindantes.

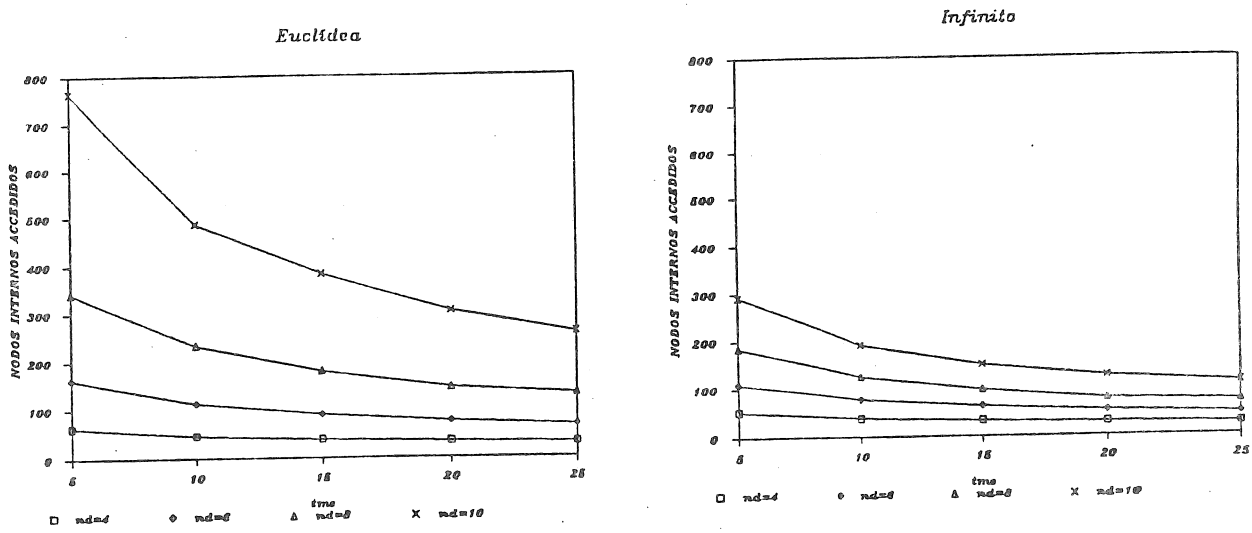

Figura 10

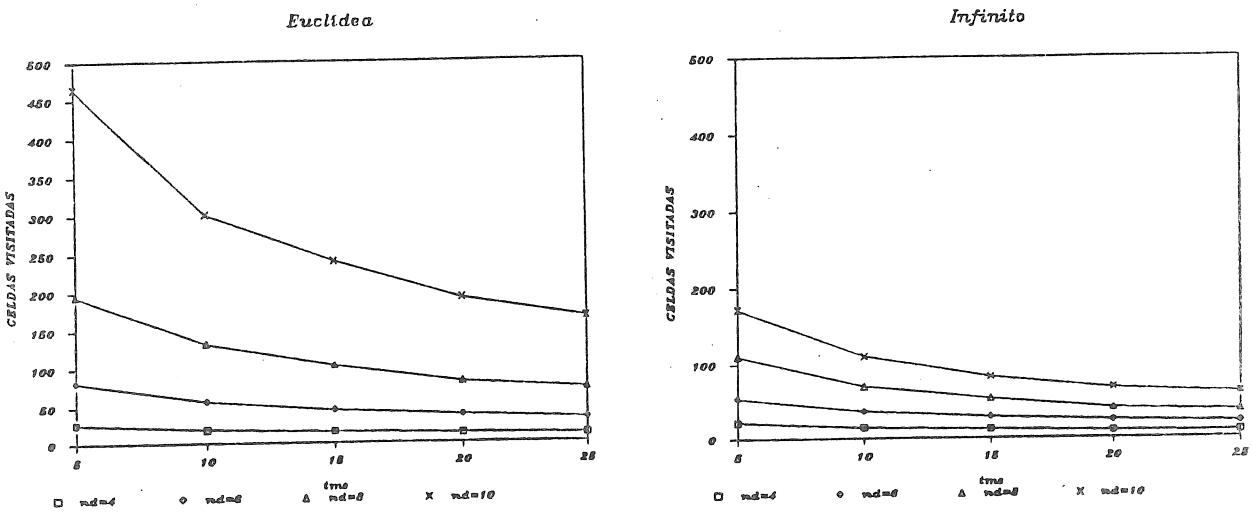

Figura 11

- El número de celdas visitadas decrece al aumentar tmc, debido a que las celdas contienen más puntos, por consiguiente el número de nodos internos accedidos también decrece.

- El número de distancias calculadas está en correlación con el tmc y nd, como se observa en la figura 12.

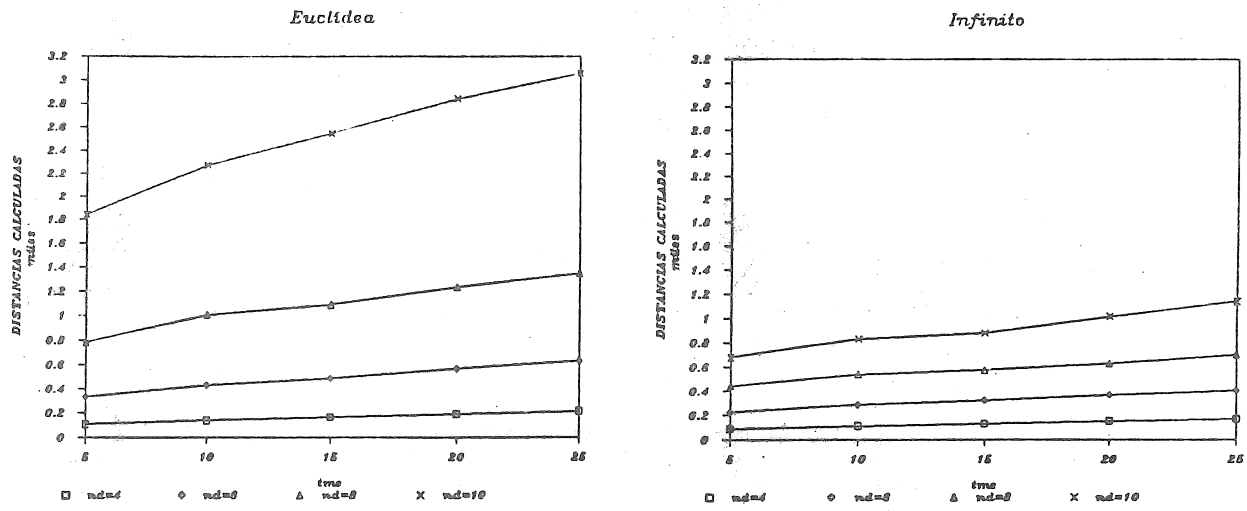

Figura 12

- El número de solapamientos con el interior, de las zonas representadas por los nodos internos, obviamente, tiene el mismo comportamiento que el número de nodos internos accedidos al variar nd y tmc para cada distancia, respectivamente. Las diferencias son mínimas y se deben a los nodos del camino de bajada inicial en los que se comprueba la contención de la esfera y a aquellos otros que están situados en niveles superiores al nodo en el que se verifica la contención.

- El número de solapamientos exteriores de interior tiene un comportamiento análogo al de los solapamientos con el interior, la diferencia está en que a los nodos del camino de bajada inicial no se les sondea este tipo de solapamiento y tampoco a los nodos de las alternativas de descenso para los que el solapamiento con el interior es falso.

- El número de solapamientos exteriores coincide con el numero de contenciones falsas de la esfera y los resultados obtenidos, figura 13, están en correlación con las alturas promedio de los árboles considerados al variar tmc. El aumento de la dimensionalidad retrasa la contención de la esfera y por tanto favorece el incremento ligero de los solapamientos exteriores.

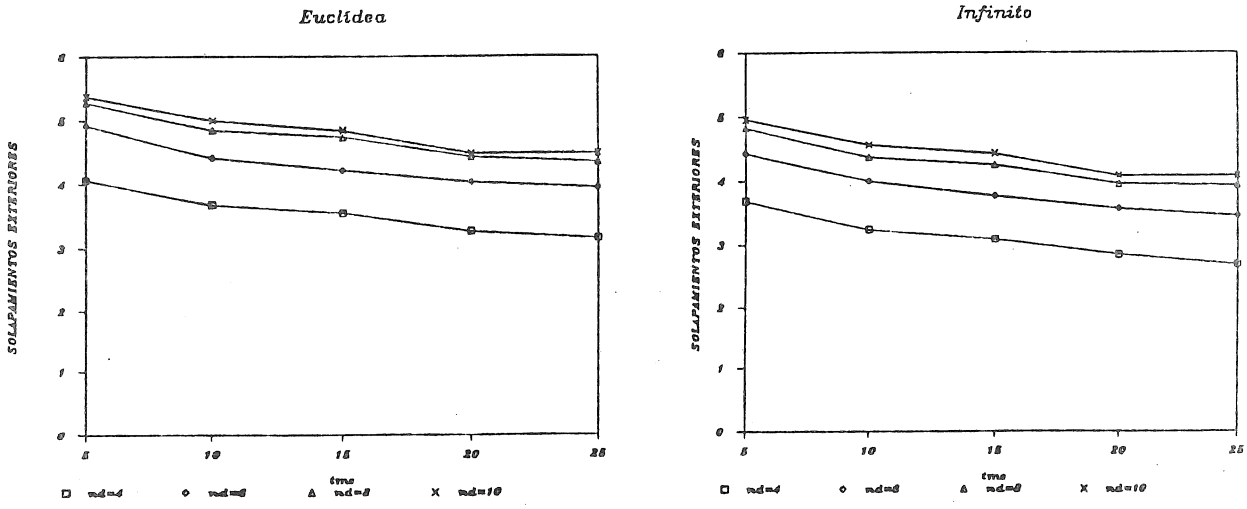

Figura 13

El incremento del número de **vecinos más próximos** provoca un aumento de todos los parámetros medidos, como se observa en la figura 14 con nd=8. La variación de la **dimensionalidad** provoca un incremento de los parámetros de acuerdo con los resultados comentados para cinco **vecinos más próximos.**

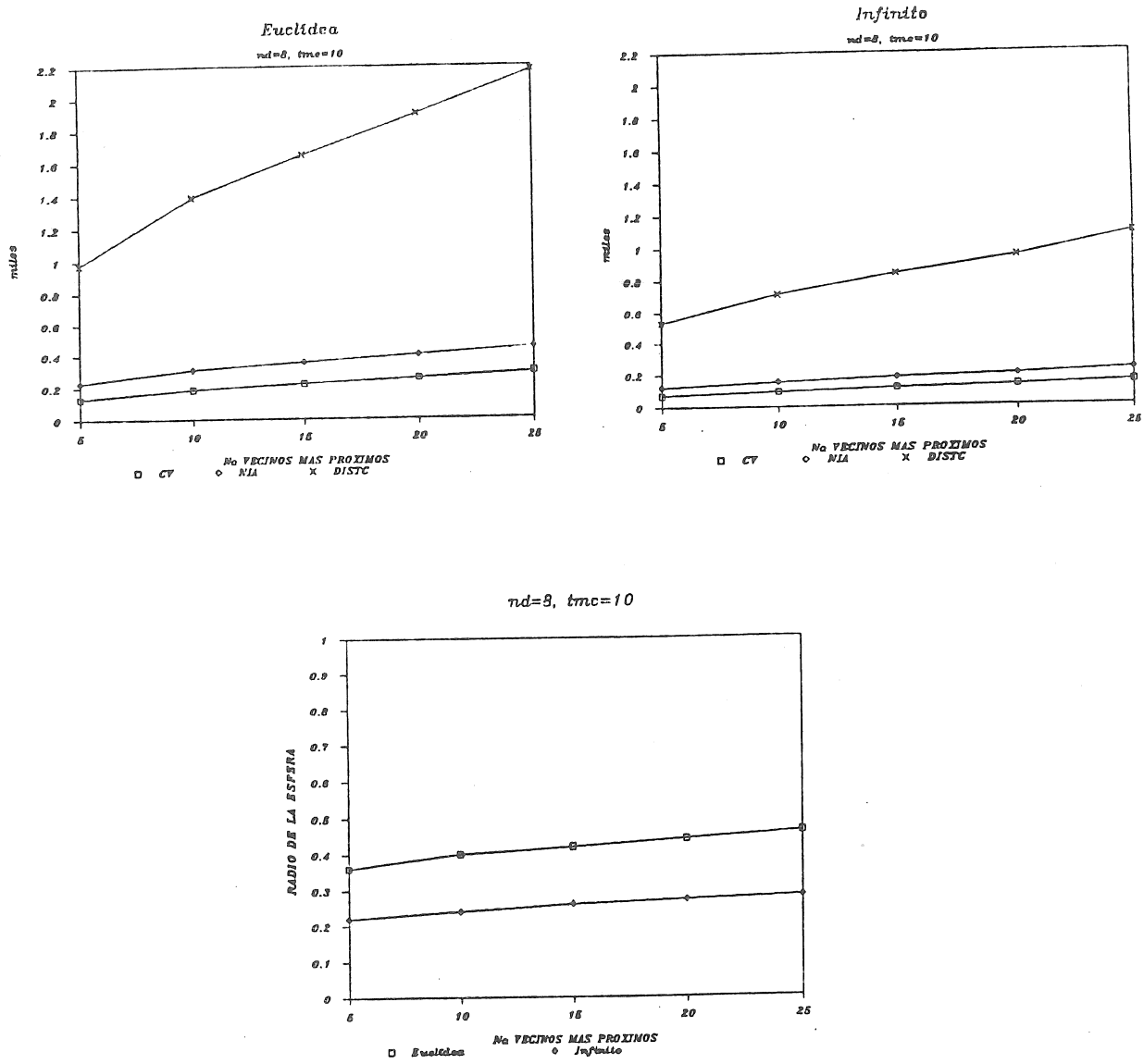

Figura 14

3) CONCLUSIONES

Se ha de citar en primer lugar que se demostró la utilidad de usar el test de solapamiento exterior de interior, al menos para tamaños máximos de celda mayores que 5, dado que el ahorro en el tiempo utilizado en calcular distancias es mayor que el tiempo perdido en realizar los test indicados. Asimismo, la distancia infinito mostró siempre mejor comportamiento que la euclídea. En cuanto a los parámetros medidos, debe indicarse que tanto el número de nodos internos accedidos, como el número de celdas visitadas, disminuyen al aumentar el tamaño máximo de celda y crecen al aumentar la dimensionalidad. El número de distancias calculadas esta en correlación con el tamaño máximo de celda y la

dimensionalidad. Tanto el número de tests de solapamiento con el interior como el de exterior de interior tienen comportamientos muy similares al del número de nodos internos accedidos. El número de tests de solapamiento con el exterior y el de contenciones de la esfera están en correlación con las alturas promedio y con la dimensionalidad. Por último, el aumento del número de vecinos a localizar provoca un aumento de todos los parámetros medidos.

BIBLIOGRAFIA

[1] DANDAMUDI, S.P. AND SORENSON, P.G.
Algorithms for the BD Tree Structure.
Dept. of Computational Science. University of Saskatchewan Saskatoon. Saskatchewan, Canada. 1984.

[2] DANDAMUDI, S.P. AND SORENSON, P.G.
An Empirical Performance Comparison of Some Variations of the K-d Tree and BD Tree.
International Journal of Computer and Information Sciences, Vol. 14, Nº 3, 1985.

[3] FRIEDMAN, J.H.; BENTLEY, J.L. AND FINKEL, R. A.
An Algorithm for Finding Best Matches in Logarithmic Expected Time.
ACM Transactions on Mathematical Software. Vol. 3, Nº 3:209-226, 1977.

[4] OHSAWA, Y. AND SAKAUCHI, M.
The BD-Tree-A New N-Dimensional Data Structure with Highly Efficient Dynamic Characteristics.
Institute of Industrial Science, University of Tokyo 22-1, Roppongi 7, Minato-ku, Tokyo 106, Japan. 1983.

[5] O.SANTANA, O. MAYOR, M. DIAZ, G. LOPEZ
Comportamiento del árbol_BD en las fases creciente, decreciente y estacionaria.
E.U. Informática, Universidad Politécnica de Canarias. España. 1987.