COMPORTAMIENTO DEL ARBOL-BD EN LAS FASES CRECIENTE, DECRECIENTE Y ESTACIONARIA

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

El presente artículo lleva a cabo un estudio experimental de la estructura denominada árbol-BD. Esta estructura es mejorada en las extracciones mediante una operación de recombinación que no permite punteros nulos y optimiza la ocupación. Dicho estudio aparece dividido en tres fases, cada una de ellas caracterizada por el tipo de operaciones a realizar. Previamente, se han realizado una serie de pruebas destinadas a comparar las características de dos posibles redistribuciones locales alternativas, a fin de escoger la más adecuada.

La primera fase, de estado creciente, corresponde a la etapa de crecimiento del árbol. En ella, se procede a realizar diez mil inserciones en cada árbol, midiendo durante este proceso una serie de parámetros que han de mostrar el comportamiento de la estructura.

La segunda fase, es la de estado estacionario. En ella se mantiene prácticamente fijo el número de elementos mediante una mezcla aleatoria de inserciones y extracciones, a partes iguales. En esta fase, a parte de la medición de parámetros de la estructura frente a operaciones alternadas, se llevan a cabo pruebas de las diferentes peticiones de información.

Por último, en la tercera fase se realiza la prueba del estado decreciente: extrayendo diez mil elementos hasta vaciar el árbol. Así, se pueden medir parámetros de la estructura ante situaciones críticas, simétricas, en cierto modo, de las de la primera fase.

COMPORTAMIENTO DEL ARBOL-BD EN LAS FASES CRECIENTE, DECRECIENTE Y ESTACIONARIA

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

0) INTRODUCCION

A menudo, es necesario recuperar información de una base de datos que está localizada en un espacio n-dimensional. Por ejemplo, en una base de datos puede ser necesario ver todos los datos con los atributos localizados en una zona rectangular dada. Una recuperación de este tipo se denomina búsqueda multidimensional.

El campo de aplicación de estas estructuras de ficheros no se queda sólo 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.

La estructura de árbol-BD, propuesta por Ohsawa y Sakauchi [4], realiza una partición dinámica del espacio n-dimensional, adaptándose al flujo de datos de entrada, y se lleva a cabo eficientemente usando la idea de zonas discriminadoras y sus expresiones derivadas. El origen del nombre es: división Binaria n-dimensional con expresión de zona Discriminadora.

En el crecimiento de la estructura de árbol-BD se utiliza el método de partición descrito en [1,2,4], para resolver las situaciones de sobrecarga de una celda. La información de los nodos internos es una expresión de zona discriminadora, EZD, [1,2,4], que representa a la zona correspondiente.

La estructura de árbol-BD puede no conducir a un árbol bien equilibrado. Se utilizan tres mejoras para resolver este problema: intercambio de EZD [4], redistribuciones locales [1,2,4] y arreglo local [4].

Intercambio de EZD

El árbol-BD tiene la propiedad de que para cada nodo el subárbol izquierdo representa un volumen menor o igual que el representado por el subárbol derecho. Por tanto pueden ocurrir situaciones no deseadas, tal como tener un nodo con un hijo derecho cuya EZD tiene menos longitud que la de su padre, lo cual significa que el hijo representa un volumen mayor que el de su padre. En este caso el árbol se puede desequilibrar y por tanto ser ineficaz para alguna secuencia de datos de entrada.

Se distinguen dos casos. En el primero, la EZD del hijo, prefija a la de su padre. En el segundo, las dos EZD son diferentes.

Redistribución local

Esta técnica propone posponer la división de una celda si el nuevo punto se puede acomodar entre las dos celdas hermanas. Esto es, cuando una celda está sobrecargada, antes de dividirla, se redistribuyen los puntos de la misma y de la celda hermana. Usando este esquema, se puede mejorar la utilización del fichero. Existen dos tipos alternativos de redistribución local uno propuesto por Dandamudi y Soreson [1,2], que se denota por RL-DS, y otro propuesto por Ohsawa y Sakauchi [4], que se denota por RL-OS.

Arreglo local

El Árbol-BD tiende a partir más frecuentemente el subárbol derecho que el izquierdo. Si se generan zonas discriminadoras con el mismo volumen configuran el árbol de forma similar a una escalera. Esta situación no es deseable, y para mantener un árbol equilibrado se introduce un esquema de arreglo local.

En el proceso de decrecimiento del árbol se presenta en este trabajo una reestructuración denominada recombinación. La recombinación geométrica de la celda correspondiente al interior de una zona con la celda de su exterior, se produce cuando la suma del número de puntos de las dos celdas no exceda de un cierto umbral del tamaño máximo de celda, tmc. Para llevar a cabo dicha recombinación se elimina la celda correspondiente al interior de la zona y esto provoca una extracción de un nodo interno.

BUSQUEDA Y RECUPERACION

Una petición de búsqueda asociativa consiste en especificar ciertas condiciones que deben de satisfacer las componentes de los puntos del espacio que se desean recuperar.

En el Árbol-BD se llevan a cabo los siguientes tipos de recuperaciones de información [2]:

I) Petición Exacta
II) Petición Parcial
III) Petición en Rango
IV) Petición en Rango Parcial

La petición exacta especifica un valor para cada clave.

Una búsqueda en rango en un espacio n-dimensional consiste en recuperar todos los puntos, r(j) j=1..n, tales que

$$rmin(j) \leq r(j) \leq rmax(j) \quad j=1..n$$

donde las componentes de los vectores rmin y rmax son los límites inferiores y superiores, respectivamente, del rango. Una petición en rango consistirá, por tanto, en especificar rmin(j) y rmax(j) para j=1..n.

La petición parcial especifica un valor para nce claves, con nce < n, y deja las restantes sin especificar.

La petición en rango parcial es similar a una petición en rango en la que algunas claves tienen el rango sin especificar.

La sección 1 de este artículo sirve como introducción al estudio experimental que se lleva a cabo sobre la estructura de árbol-BD. Dicho estudio se va desgranando en las sucesivas secciones de una forma acorde con el carácter dinámico de la estructura [3]. Así, en la sección 2 se desarrolla la fase del experimento que estudia la estructura en crecimiento continuo. Previamente, en esta misma sección se presentan las pruebas necesarias para dilucidar entre los dos tipos de redistribuciones locales. La sección 3 contiene el estudio de la fase estacionaria y las pruebas realizadas para medir diferentes parámetros de la estructura ante las peticiones de información solicitadas. Por último, la sección 4 estudia la fase decreciente del experimento. En la sección 5 se presentan las conclusiones de este trabajo.

1) DESCRIPCION DEL ESTUDIO EXPERIMENTAL

El rendimiento de una estructura de datos está determinado por dos criterios: el tiempo de procesamiento y la utilización de memoria.

Se han realizado varias simulaciones para estudiar el rendimiento del Árbol-BD, midiendo estos criterios en base a los siguientes parámetros:

a) número de nodos internos totales, NIT
b) número de celdas totales, CT
c) número de nodos internos accedidos, NIA
d) ocupación, PO
e) altura promedio, HP
f) reestructuraciones, R

De acuerdo con el caracter dinámico de la estructura el estudio se ha realizado en las siguientes fases:

1) Estado creciente. Inserciones repetidas.
2) Estado estacionario. Inserciones y extracciones manteniendo aproximadamente constante el número de registros.
3) Estado decreciente. Extracciones repetidas.

Todos los números aleatorios utilizados en los experimentos se han obtenido a partir de un generador de números pseudoaleatorios de distribución continua uniforme en [0,1]. Para cada registro se generaron nd, siendo nd la dimensionalidad del espacio, valores aleatorios pertenecientes a una distribución uniforme discreta entre 0,01 y 1,00.

2) ESTADO CRECIENTE

El estado creciente corresponde a la fase de ampliación del árbol a través de sucesivas inserciones.

Se construyeron árboles de 10000 registros. Los párametros anteriormente descritos, desde el a hasta el e, se midieron en intervalos regulares de 100 inserciones, excepto la altura promedio que se midió en intervalos de 1000. El número de nodos internos difiere del número de celdas en uno, ya que no existen punteros nulos. Las reestructuraciones se midieron a nivel de promedios y los tipos estudiados fueron: particiones, intercambio de EZD, redistribución local y arreglo local.

Una cuestión previa es el estudio comparativo de la estructura para los dos tipos de redistribuciones locales: la RL-OS y la RL-DS. Para ello se construyeron dos árboles de dimensionalidad 6 y tmc 5, obteniéndose que:

- La diferencia del número de nodos internos de ambos se fue incrementando a lo largo del proceso de inserción, alcanzándose al final un árbol de 2762 con la primera y de 2564 con la segunda.

- El número de nodos internos accedidos por inserción coincidió para ambos árboles a lo largo de todo el proceso.

- La ocupación fue próxima a un 78% con la RL-DS y a un 73% con la RL-OS, manteniendo una diferencia constante a lo largo de todo el proceso, como se observa en la figura 1.

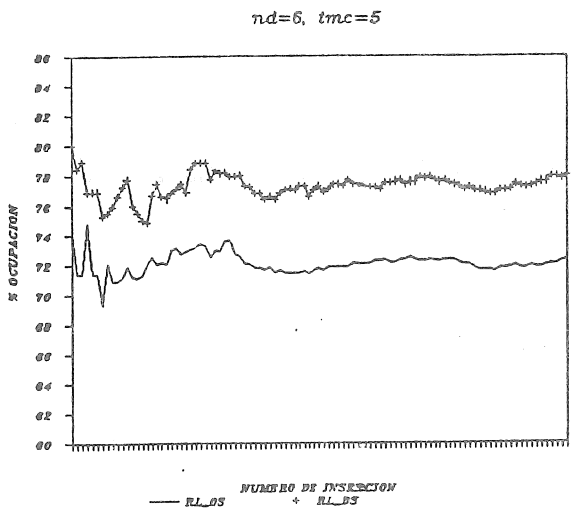

Figura 1

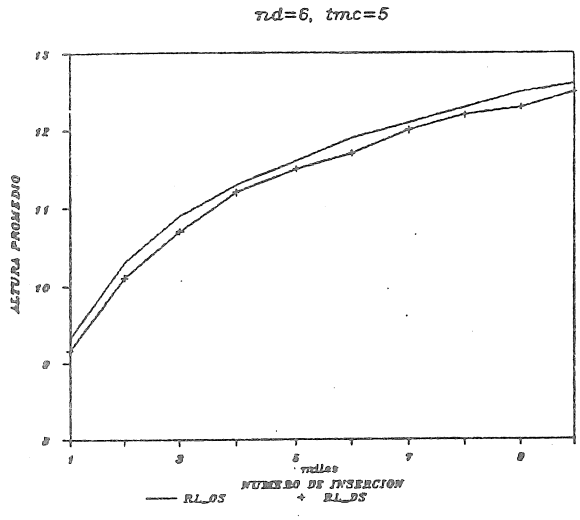

Figura 2

- La altura promedio es creciente y mantiene una diferencia mínima constante entre ambos árboles, a favor de la RL-DS, como se observa en figura 2.

- La estructura requiere un gasto adicional para mantener el árbol equilibrado que viene dado por los promedios de intercambios de EZD, arreglos locales y redistribuciones. Cuando se utiliza la RL-DS no se realiza ningún arreglo local, mientras que si se realizan para la RL-OS. En la tabla 1 se muestran estos resultados, además del número de particiones y del tiempo por inserción obtenidos en la comparación:

PROMEDIO DE	BD con RL-OS	BD con RL-DS
RL-DS		0.1244
RL-OS	0.0332	
INTERCAMBIO DE EZD I	0.0445	0.0324
INTERCAMBIO DE EZD II	0.0447	0.0353
ARREGLO LOCAL	0.0037	
PARTICION	0.2768	0.2562
TIEMPO/INSERCION (sg)	0.0423	0.0398

Tabla 1

Estos resultados indican un mejor rendimiento de la estructura con la RL-DS. Por tanto, todos los experimentos que se presentan a continuación se realizaron utilizando dicha redistribución local.

Se han construído 20 árboles, teniendo en cuenta la variación de la dimensionalidad y tmc, para los siguientes valores:

nd : 4, 6, 8, 10
tmc: 5, 10, 15, 20, 25

obteniéndose que:

- El número de nodos internos se incrementó en todos los casos, a lo largo del proceso, obteniéndose finalmente árboles cuyo NIT disminuye al crecer tmc. Por otro lado NIT no aparece afectado por la dimensionalidad.

- El número de nodos internos accedidos para cada valor de tmc, se comprobó que prácticamente no variaba con la dimensionalidad. Para una dimensionalidad dada al aumentar tmc tal y como se observa en la figura 3 el número de accesos disminuye.

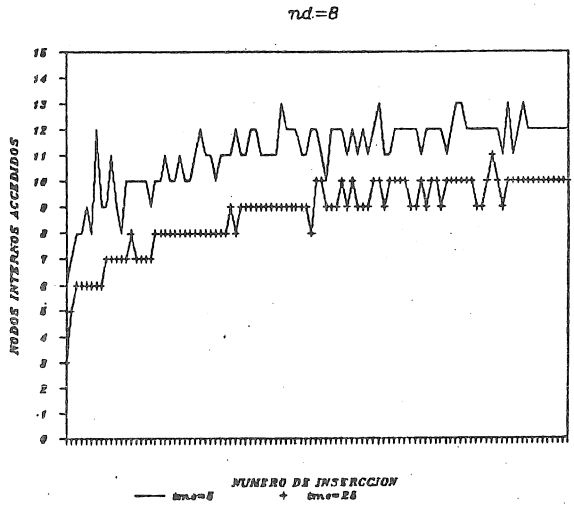

Figura 3

- En la ocupación tampoco influye la dimensionalidad al fijar tmc, como se observa en las figuras 4 y 5, donde cada línea representa una dimensionalidad distinta. Así mismo la ocupación crece al disminuir tmc, observándose que en cada caso están presentes oscilaciones de tipo sinusoidal, asintóticas a un valor constante, siendo estas oscilaciones mayores cuanto mayor es tmc.

- La altura promedio para cada valor de tmc coincidió exactamente al variar la dimensionalidad. Es mayor para valores de tmc menores, manteniéndose una diferencia constante durante el proceso de inserción, como se observa en la figura 6.

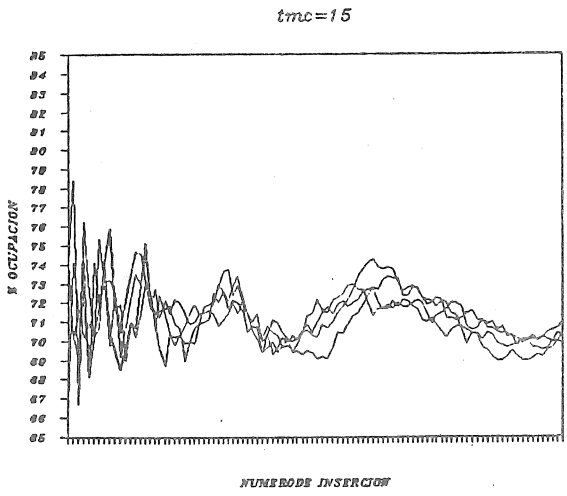

Figura 4

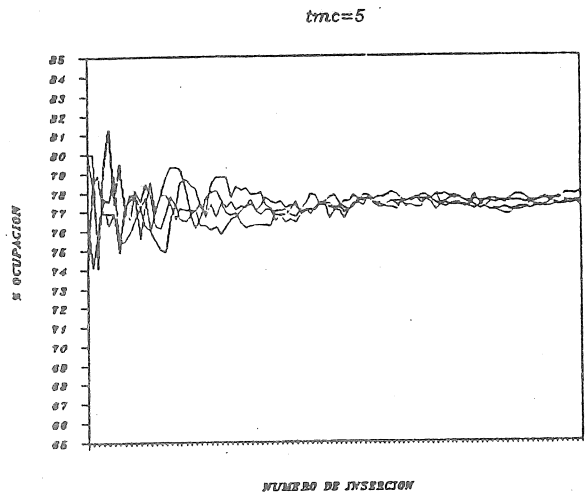

Figura 5

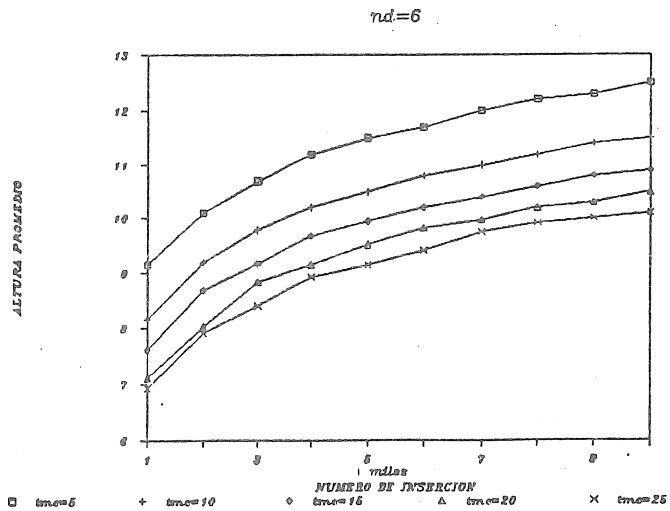

Figura 6

3) ESTADO ESTACIONARIO

El estado estacionario corresponde a la fase de actualización del árbol, a través de la selección de inserciones y extracciones de forma aleatoria.

A partir de un árbol de 10000 registros construidos en el estado creciente, se insertaron 1000 nuevos registros y se extrajeron 1000 de los ya existentes.

Debido a que la variación de la dimensionalidad prácticamente no influye en los parámetros medidos, para los valores estudiados, y que la variación de tmc

presenta resultados análogos, el estudio experimental en este estado se realizó con dimensionalidad 6 y tmc 10.

Los parámetros anteriormente descritos, desde el a hasta el e, se midieron en intervalos regulares de 100 inserciones+extracciones, excepto la altura promedio que se midio en intervalos de 400. Las reestructuraciones utilizadas fueron, por un lado, en las inserciones las mismas que en el estado creciente, y por otro, en las extracciones: la extracción de un nodo interno, la recombinación en función de los distintos umbrales. Se midió el promedio de las reestructuraciones de forma conjunta.

- Se observó durante el proceso un incremento del número de nodos internos, indicando un aumento de entropia de la estructura, debido a la diferencia de eficacia en la aplicación de los procesos de mantenimiento en las fases de inserción y extracción.

- El número de nodos internos accedidos fue prácticamente constante durante el proceso.

- En correspondencia con el hecho de que el número de nodos internos aumenta, se obtuvo que la ocupación disminuye a medida que aumenta el número de inserciones+extracciones.

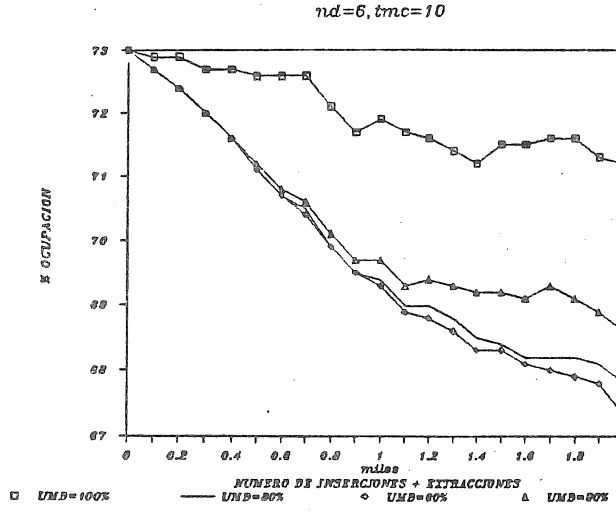

Figura 7

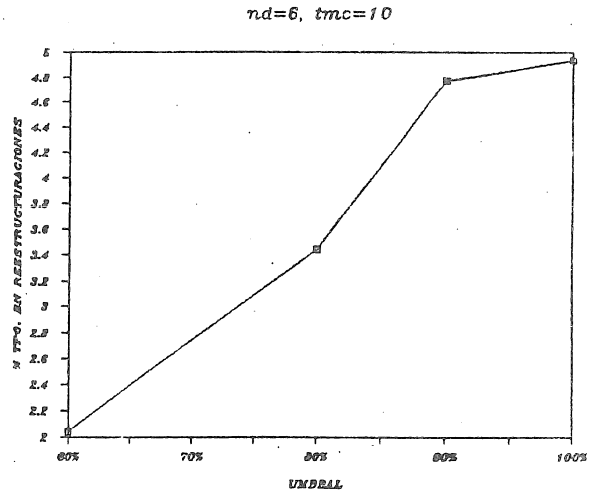

Figura 8

Como se observa en la figura 7, la disminución del umbral de recombinación produce un efecto negativo en la ocupación de la estructura. Este hecho, aparece agravado por la tendencia del proceso a aumentar el número de nodos, tendencia que aparece más fuerte cuanto menor sea el umbral de recombinación.

En la figura 8 se muestra la porción del tiempo total que se emplea en las operaciones de reestructuración. Como era de esperar se comprueba que el aumento del umbral de recombinación hace crecer el número de operaciones de mantenimiento precisas.

3.1 Estudio experimental de las peticiones exacta, parcial, en rango y en rango parcial

Otro criterio para determinar el rendimiento de una estructura es el tiempo de respuesta a los diversos tipos de peticiones.

El tiempo de respuesta se mide en función de los siguientes parámetros:

c) número de nodos internos accedidos, NIA.
g) número de celdas visitadas, CV.
h) número de registros recuperados, RR.
i) eficiencia de la petición = (RR / RT) / (CV / CT), donde RT es el número de registros totales.

Durante el estado estacionario se midieron estos parámetros como promedio, con umbral de recombinación 100%, para los distintos tipos de peticiones, en cinco instantes (inserción+extracción) del mismo: 0, 500, 1000, 1500 y 2000.

Las peticiones exactas se eligieron de los ficheros utilizados para la construcción de cada árbol. Las parciales se generaron a partir de las exactas seleccionando aleatoriamente las posiciones de las claves sin especificar, variando el número de claves sin especificar de 1 a 4. Las peticiones en rango se generaron a partir de las exactas, considerándolas como límite inferior del rango, sumándole a cada componente una longitud fija para obtener el límite superior; se tomaron longitudes de 0.1, 0.2, 0.3 y 0.4. Sólo se tuvieron en cuenta aquellos hipercubos que estuviesen contenidos en el espacio considerado. Las peticiones en rango parcial se generaron a partir de las de rango seleccionando las posiciones y el número de claves sin especificar como en las peticiones parciales.

El número de pruebas realizadas en cada instante y para cada petición fue de: 500 en la exacta, 2000 en la parcial (correspondiendo 500 a cada valor del número de claves sin especificar), 2000 en rango (correspondiendo 500 a cada tamaño del hipercubo), y 8000 en rango parcial (teniendo en cuenta el número de claves sin especificar y el tamaño del hipercubo).

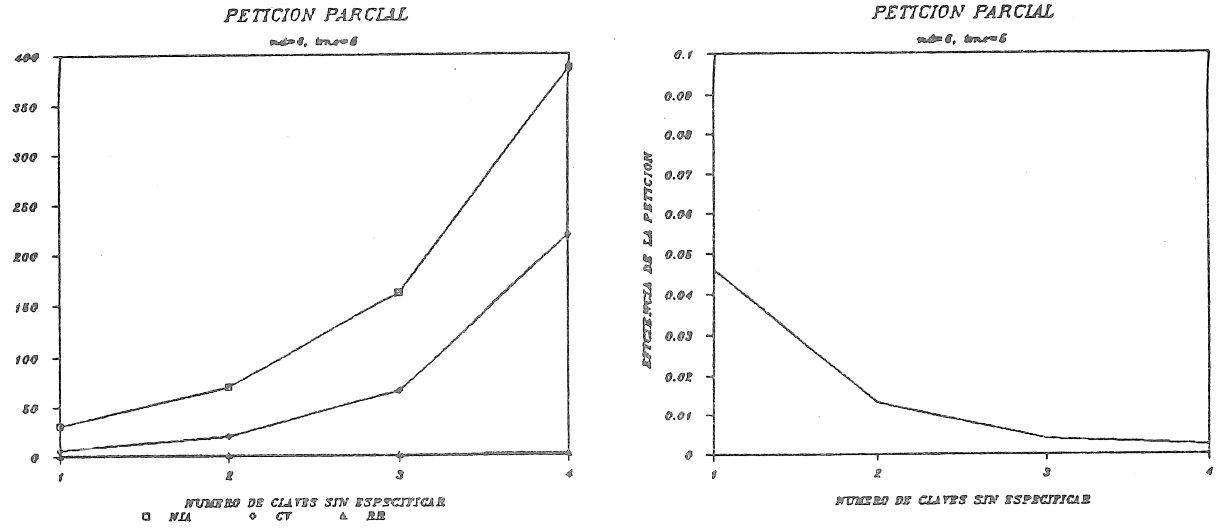

Figura 9

Los resultados obtenidos mostraron la independencia de los parámetros c, g, h e i, del instante en que se realice cualquier petición.

En una petición exacta se visita una sola celda y el número de puntos recuperados puede ser uno o ninguno, dependiendo de si la búsqueda es exitosa o infructuosa. El número promedio de accesos obtenido fue 12.4128.

El comportamiento de la petición parcial en función del número de claves sin especificar se presenta en la figura 9. El crecimiento del número de registros recuperados no es tan significativo como el de celdas debido al gran volumen del espacio en el que se está trabajando. La eficiencia de la petición es baja por la misma razón.

Los resultados de la petición en rango en función del tamaño del hipercubo, th, se muestran en la figura 10. Nuevamente se observa que el crecimiento del número de registros recuperados es inferior al número de celdas debido a la baja densidad de puntos. La eficiencia de la petición es baja por la misma razón.

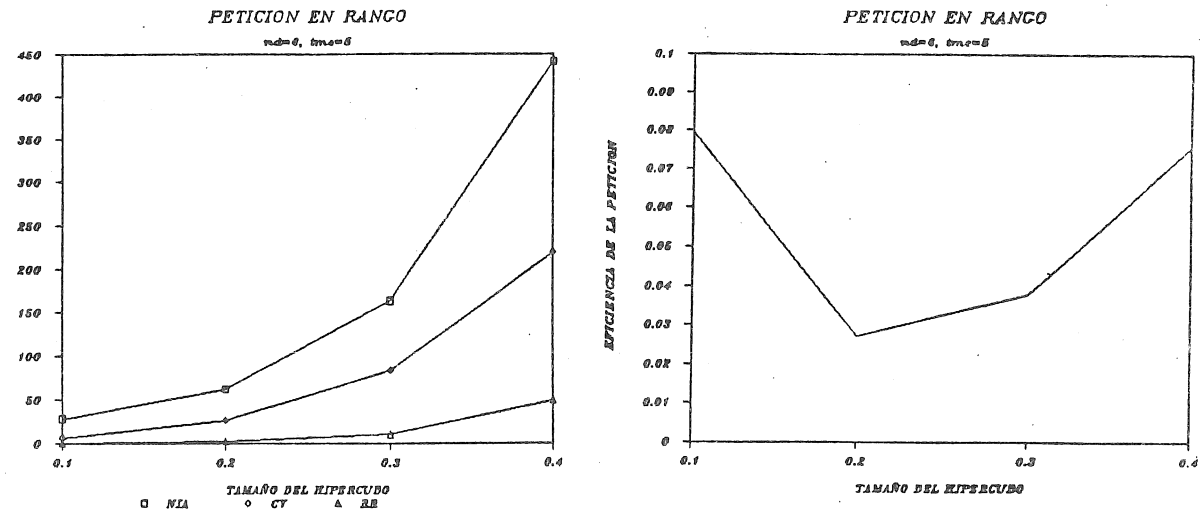

Figura 10

En la petición en rango parcial se presentan, en la figura 11, los resultados en función del número de claves sin especificar y del tamaño del hipercubo. Como era de esperar, los cuatro valores medidos aumentan tanto con el número de claves sin especificar como con el tamaño del hipercubo. Con estos niveles de vaguedad en la petición y a pesar de la baja densidad de puntos del árbol, se presenta por primera vez en el número de registros recuperados un crecimiento similar al del número de celdas visitadas.

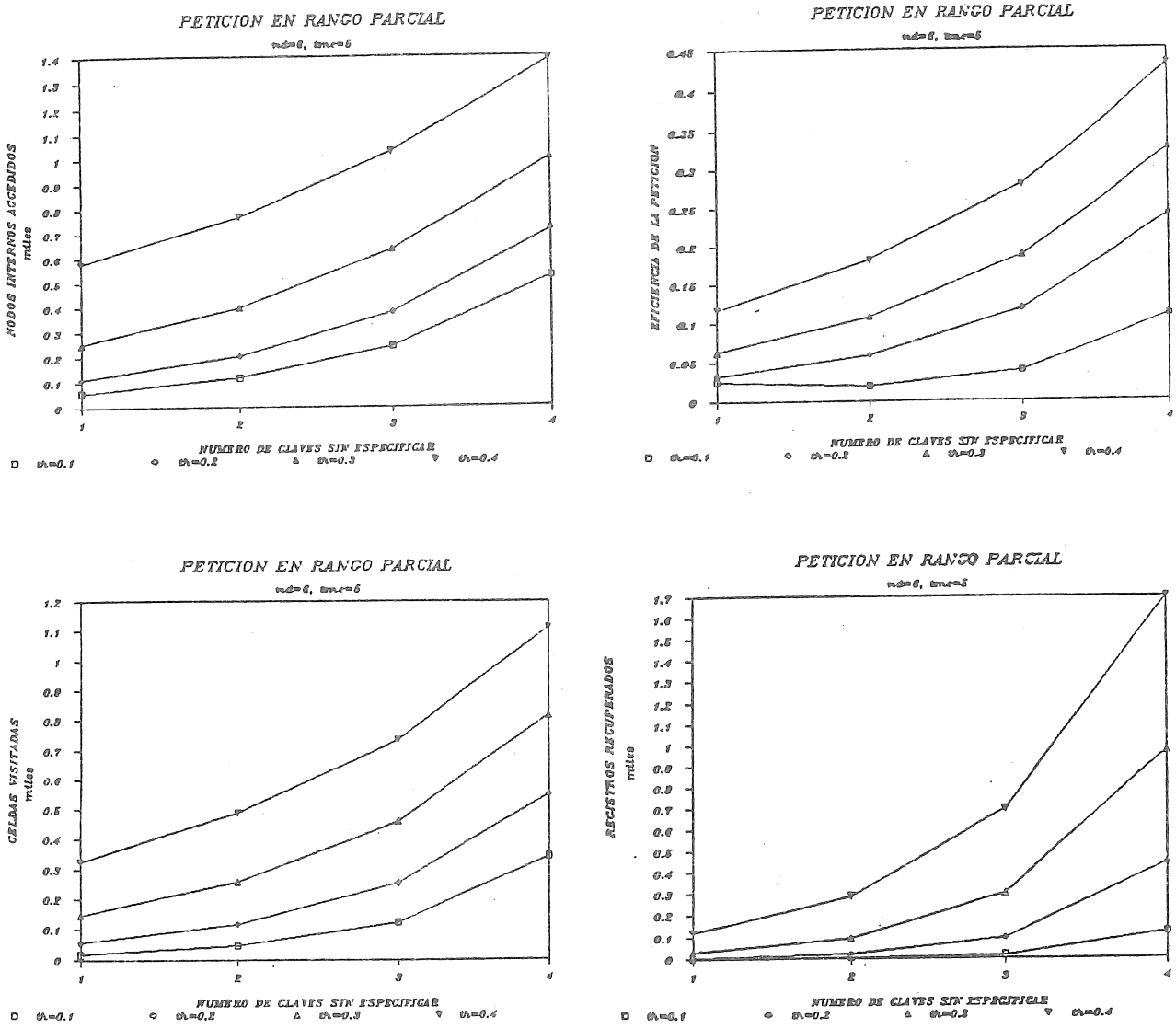

Figura 11

4) ESTADO DECRECIENTE

El estado decreciente corresponde a la fase de disminución del árbol a través de sucesivas extracciones.

A partir del árbol final del estado estacionario se realizaron las extracciones de los 10000 registros, midiendo los mismos parámetros y en iguales intervalos que en el estado creciente. Cada uno de los parámetros fue estudiado para diferentes umbrales de recombinación. Obteniéndose que:

- El número de nodos internos accedidos, presenta una ligera disminución con el número de extracciones hasta un punto muy cerca del final del proceso en el cual desciende drásticamente. La variación del umbral no presenta diferencias significativas.

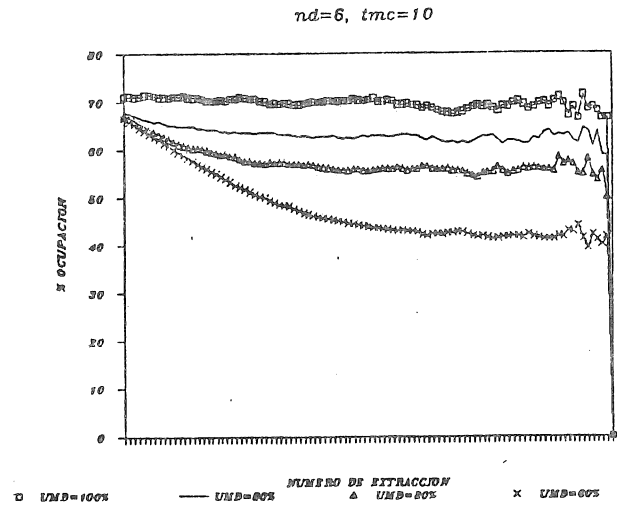

Figura 12

- En cuanto a la ocupación se observa, figura 12, de forma general, una ligera disminución a lo largo del proceso. Tan solo en la extracción de los últimos registros aparece un descenso brusco. Esto hace pensar en lo adecuado del sistema que permite mantener un alto nivel de ocupación hasta alcanzar las últimas extracciones. En cuanto al estudio de los distintos umbrales aparece una caída del nivel de ocupación más precipitada cuanto menor sea el umbral de recombinación.

5) CONCLUSIONES

El presente estudio experimental se ha llevado a cabo utilizando la redistribución local RL-DS, puesto que en las comparaciones previamente efectuadas se manifestó como la más adecuada, tanto en compactación del almacenamiento utilizado, como en tiempo de proceso.

En la fase de crecimiento se observó la independencia de los parámetros medidos con respecto a la dimensionalidad. Sin embargo, y tal como era de esperar, el número de nodos internos totales, el número de nodos accedidos, la ocupación y la altura promedio presentaron crecimientos al disminuir el tamaño máximo de la celda.

La fase estacionaria presentó un ligero, pero continuo, aumento del número de nodos internos totales. Sin embargo, los valores del número de nodos internos accedidos sí que se mantuvieron prácticamente constantes. La ocupación presentó también una tendencia a la baja, tendencia que se ve agravada al producirse la más mínima variación del umbral de recombinación óptimo (100%). Como es lógico el aumento del umbral de recombinación trajo consigo un crecimiento del tiempo empleado en operaciones de reestructuración. El estudio de los distintos tipos de peticiones se mostró independiente del punto de la fase estacionaria en que se lleva a cabo.

Por último, la fase decreciente mostró como, a pesar de que la ocupación es menor cuanto menor es el umbral de recombinación, el descenso brusco se produce para todos los umbrales tan solo en la extracción de los últimos registros.

REFERENCIAS:

[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] NIEVERGELT, J.; HINTERBERGER H. AND SEVCIK K.C.
The Grid File: An Adaptable, Symmetric Multi-key File Structure.
Institut für Informatik.
ETH Zürich. 1981.

[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.