

ARBOLES QUINTARIOS: ESTUDIO EXPERIMENTAL PARA INTERROGACIONES EXACTAS, PARCIALES, EN RANGO Y EN RANGO PARCIAL

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

Se presenta un estudio experimental del costo de almacenamiento y el tiempo de respuesta en una estructura de datos denominada Arbol Quintario para un sistema de recuperación de información basado en las interrogaciones exacta, parcial, en rango y en rango parcial.

Para valores moderados de la dimensionalidad, cardinalidad e intervalo de variación de los valores de clave se han obtenido árboles de grandes volúmenes, siendo el parámetro que más influye, la dimensionalidad.

Los números de accesos y comparaciones aumentan con la cardinalidad y la dimensionalidad en los cuatro tipos de interrogaciones. En lo relacionado con el intervalo de variación de los valores de clave se observa que los resultados se ven afectados por varios factores. En las interrogaciones exacta y parcial aunque se produce algún crecimiento, este es muy ligero. Y en las interrogaciones en rango y en rango parcial se produce un decrecimiento.

En definitiva, el tiempo de respuesta es bueno para interrogaciones exactas y parciales, y aceptable para interrogaciones en rango y en rango parcial, siendo la dimensionalidad el parámetro que influye de forma negativa para este factor.

ARBOLES QUINTARIOS: ESTUDIO EXPERIMENTAL PARA INTERROGACIONES EXACTAS, PARCIALES, EN RANGO Y EN RANGO PARCIAL

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

RESUMEN

En éste artículo se realiza un estudio experimental de una estructura de datos denominada Arbol Quintario, diseñada para un sistema de recuperación de información basado en los siguientes tipos de interrogaciones: exacta, parcial, en rango y en rango parcial.

0) INTRODUCCION

Han aparecido muchos métodos para la construcción de sistemas de recuperación de información, la mayoría de ellos se centran sobre alguno de los cuatro tipos de interrogaciones [2,3,4,5,6,10]. La idea del Arbol Quintario surge a partir de dos estructuras: el Arbol de Rango, desarrollado por Bentley y Friedman [1], especial para interrogaciones en rango; y el Arbol Complejo de Búsqueda Binaria (BST-Complejo), desarrollado por Lien, Taylor y Driscoll [8], adecuado, principalmente para interrogaciones exactas.

El Arbol Quintario fue propuesto por D.T. Lee y C.K. Wong [7], como una estructura de datos con un tiempo de respuesta bajo, para el peor caso, en los cuatro tipos de interrogaciones, a costa de un gasto de almacenamiento extra.

1) ARBOL QUINTARIO

Se considera un fichero F de n registros, donde cada registro es una nd -tupla de valores ordenados. Cada componente de la nd -tupla, $(k_0, k_1, \dots, k_{nd-1})$, se denomina atributo o clave. El número de valores distintos que puede tomar cada clave k_i es v_i .

En un Arbol Quintario hay nd niveles de nodos internos representando a las distintas claves. En cada uno de estos niveles se tiene un árbol binario cuyos nodos representan a los valores de clave, el número máximo de nodos en la estructura de árbol binario del nivel i es v_i . Cada uno de estos nodos, además de tener los dos hijos característicos de la estructura de árbol binario en el mismo nivel, tienen unidos tres árboles adicionales pertenecientes a un nivel inferior.

En general, considerando el dominio del fichero F como un espacio nd -dimensional y cada clave como un eje de coordenadas, se tiene que cada nodo en el nivel i representa un subfichero en un espacio $(nd-i)$ -dimensional, y tiene cinco ramificaciones: dos de ellas apuntando a los nodos en el mismo nivel, que representan a dos subficheros en el espacio $(nd-i)$ -dimensional que pertenecen a la izquierda y a la derecha del hiperplano asociado al nodo padre; y otras tres, que representan a tres subficheros en el espacio $(nd-i-1)$ -dimensional en el nivel $i+1$, dos de ellos son las proyecciones de los dos subficheros anteriores sobre el hiperplano, y

el tercero está constituido por los registros del hiperplano. El número total de registros de estos tres subficheros es igual al número de registros del subfichero representado por el nodo padre.

El árbol se construye como sigue: dado el fichero de n registros, se calcula la mediana de éste, rm , de tal forma que el registro rm divida al fichero en dos partes que no contengan más de $n/2$ registros cada una. Una vez hecho esto, se construyen tres subficheros *fizquierdo*, *fmid* y *fderecho*, tales que los valores de la primera clave en cada uno de ellos, sean menores, iguales y mayores que el valor de la primera clave del registro rm , respectivamente.

Se crea el nodo raíz con el valor de la primera clave del registro rm , y cinco punteros que son *izquierdo*, *medizquierdo*, *central*, *medderecho* y *derecho*. A partir de este punto, se generan los cinco subárboles, sabiendo que *izquierdo* y *derecho* son los punteros destinados a la estructura de árbol binario en cada nivel; y por lo tanto, apuntan a los subárboles generados a partir de los ficheros *fizquierdo* y *fderecho* en el mismo espacio nd -dimensional (representando a la misma clave). Sin embargo, los punteros *medizquierdo*, *central* y *medderecho* apuntan a los subárboles situados en un nivel inferior (espacio $(nd-1)$ -dimensional), que representan, en definitiva, a los mismos ficheros *fizquierdo*, *fmid* y *fderecho*, pero ignorando la clave 0, al descender de nivel se pasa a representar a la clave siguiente. El mismo proceso se repite para los subficheros generados mientras no se llegue al nivel $(nd-1)$ -ésimo.

En el nivel $(nd-1)$ -ésimo, además de los punteros *izquierdo* y *derecho* de la estructura de árbol binario, se mantienen los tres punteros centrales, pero apuntando a los nodos hoja del árbol: *central* apunta a los registros representados por el nodo, y *medizquierdo* y *medderecho* apuntan a los registros de los subárboles *izquierdo* y *derecho*, respectivamente.

La figura 1 muestra el Árbol Quintario para la base de datos bidimensional representada en la tabla. Está constituido por dos niveles de nodos internos: El nivel 0 que representa la clave A y el nivel 1 que representa a la B; además del nivel de nodos hoja que apuntan a los registros. Se selecciona el nodo c como nodo raíz de la estructura de árbol binario en el nivel 0, y los nodos b y f como raíces de los subárboles *izquierdo* y *derecho*, respectivamente. Los registros que constituyen el subfichero *fizquierdo* son los registros del 1 al 5, los de *fmid* son los del 6 al 8 y los de *fderecho*, del 9 al 15.

Según Lee y Wong el tiempo requerido para construir un árbol quintario es $O(n * (\log n)^{nd})$ y el máximo almacenamiento es $O(n * (\log n)^{nd})$.

2) BUSQUEDA Y RECUPERACION

Además de proponer la estructura de árbol quintario, D.T. Lee y C.K. Wong, esbozan los algoritmos dedicados a la búsqueda y recuperación de información de acuerdo con cada uno de los cuatro tipos de interrogaciones. En los próximos apartados se resume el proceso que siguen estos algoritmos.

2.1) Búsqueda exacta

La interrogación exacta es aquella en la que se especifica un valor para cada clave. El proceso de recuperación debe ir recorriendo la estructura de árbol binario de cada nivel i , de acuerdo con el valor de la clave k_i especificada, de la misma forma que una búsqueda binaria, con la salvedad que cuando el valor de un nodo coincida con el valor de la clave correspondiente, se desciende de nivel y se continúa la búsqueda a partir de la clave siguiente.

Ya que cada acceso a nodo descarta generalmente la mitad del fichero representado por el nodo, el número total de nodos accedidos es de casi $O(nd + \log n)$ [7].

2.2) Búsqueda parcial

La interrogación parcial especifica

$$s < nd$$

valores de clave, quedando las $nd-s$ claves restantes sin especificar. El proceso a seguir para recuperaciones de este tipo es similar al anterior, excepto en que cuando no se especifica una clave se tiene que buscar a través de los tres subárboles situados en el nivel inferior. Siendo t el número de registros recuperados, el número máximo de accesos a nodos está limitado por $O(3^{nd-s}(s + \log n) + t)$ [7].

2.3) Búsqueda en rango

La interrogación en rango especifica un intervalo de variación $[q(l_i, i), q(l_s, i)]$ para cada clave k_i . Para recuperar la información se accede al nodo raíz del árbol en el nivel 0. Si el valor del nodo está fuera del intervalo especificado por la interrogación, se siguen los punteros izquierdo o derecho dependiendo de si el valor del nodo es mayor que $q(l_s, 0)$ o menor que $q(l_i, 0)$. Si el valor del nodo está dentro del intervalo, se utilizan dos variables de control, I y D , que indican que el intervalo $[q(l_i, i), q(l_s, i)]$ ha sido partido en dos subintervalos $[q(l_i, i), x)$, $(x, q(l_s, i)]$ donde x es el valor almacenado en el nodo actual. De esta forma, si el nodo raíz está dentro del intervalo, se accede recursivamente al subárbol central en un nivel inferior, y a los subárboles izquierdo y derecho en el mismo nivel, poniendo las variables de control I y D a 1. La partición inicial del intervalo da lugar al recorrido de los subárboles central, izquierdo y derecho. Por otro lado, la partición siguiente de los subintervalos da lugar al recorrido de los subárboles central, izquierdo y medderecho (derecho y medizquierdo). Ya que cada intervalo $[q(l_i, i), q(l_s, i)]$ se puede dividir en hasta $\log n$ subintervalos en cada nivel i , el número de accesos a nodo está limitado por $O(t + (\log n) \cdot nd)$ [7], donde t es el número de registros encontrados.

2.4) Búsqueda en rango parcial

La interrogación en rango parcial es aquella que especifica

$$s < nd$$

intervalos de variación, dejando los nd-s restantes sin especificar. El algoritmo que proponen Lee y Wong para este tipo de búsqueda es similar al anterior, excepto en que si no se especifica una clave, se tiene que buscar a través de los tres subárboles centrales del nivel inferior.

El número máximo de nodos accedidos se puede obtener análogamente como en una búsqueda parcial, siendo t el número de registros recuperados: $O(3^{**}(nd-s)*\log n)^{**s+t}$ [7].

3) ESTUDIO EXPERIMENTAL PARA INTERROGACIONES EXACTAS, PARCIALES, EN RANGO Y EN RANGO PARCIAL.

A pesar de la importancia del estudio teórico de una estructura de datos, el estudio experimental proporciona una serie de valores añadidos al conocimiento que de la misma se tiene [9].

3.1) DESCRIPCION DE LOS FICHEROS

Según Lee y Wong [7] se concluye que, debido a que el almacenamiento requerido es $O(n(\log n)^{**nd})$, se tiene que para un tamaño de bases de datos moderado, en el cual la dimensionalidad es alta, es muy costoso implementar esta estructura. Sin embargo, si el almacenamiento no es el problema principal o el número de claves es pequeño, la implementación es abordable.

De acuerdo con todas las limitaciones que presenta la estructura, en relación con el almacenamiento requerido, para la realización de las pruebas se crearon 11 ficheros, cuyas características se pueden observar en la tabla 1. Tales ficheros están divididos en tres grupos, de acuerdo con la variación de cada uno de los siguientes parámetros: cardinalidad, n , dimensionalidad, nd , e intervalo de variación de los valores de clave, vi .

Nombre del Fichero	Dimensionalidad	Cardinalidad	Intervalo de Variación
QTRKR0	4	300	0..14
QTRKR1	4	1.000	0..14
QTRKR2	4	2.000	0..14
QTRKR3	4	4.000	0..14
QTRKN0	4	1.000	0..9
QTRKN1	4	1.000	0..11
QTRKN2	4	1.000	0..14
QTRKN3	4	1.000	0..16
QTRNR0	4	1.000	0..9
QTRNR1	8	1.000	0..9
QTRNR2	8	1.000	0..9

Tabla 1: Ficheros utilizados

Lee y Wong [7] puntualizan que, ya que el orden de las claves no sólo afecta al tiempo de respuesta, sino también al almacenamiento requerido, en esta estructura se debe tener en cuenta que la nd -tupla de cada registro se ordene de acuerdo con algún criterio: la muestra de interrogaciones planteadas por los usuarios del sistema, o las probabilidades de las claves especificadas en una interrogación. En la interrogación parcial o en rango parcial, por ejemplo, si una clave se especifica con mucha frecuencia, debe estar en uno de los

primeros niveles en la estructura del árbol. En este experimento todas las claves de un fichero tienen el mismo intervalo de variación, y las interrogaciones utilizadas en las pruebas de búsqueda se generan aleatoriamente con distribución uniforme, puesto que no se tiene en cuenta el problema de la ordenación de las claves.

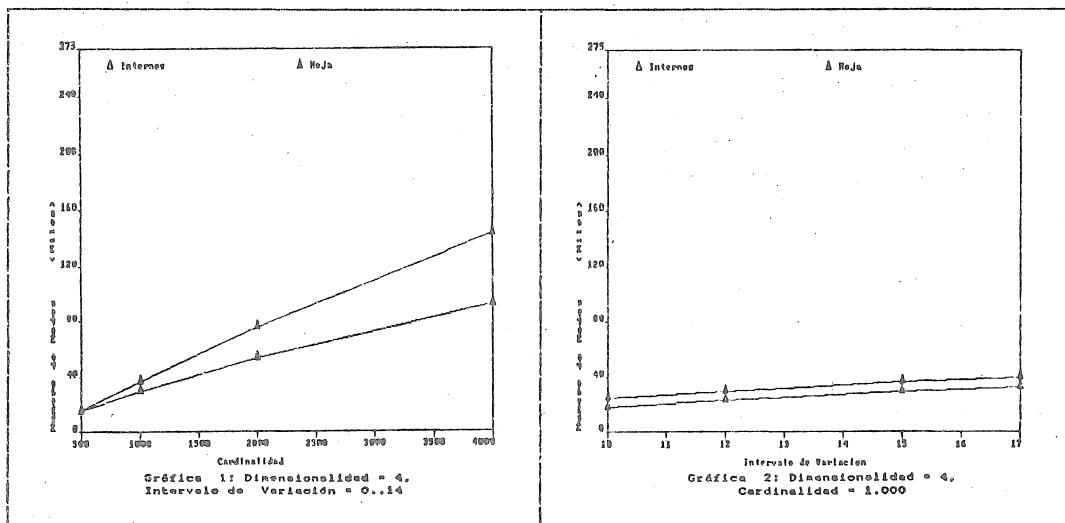
3.2) OBJETIVOS DEL EXPERIMENTO

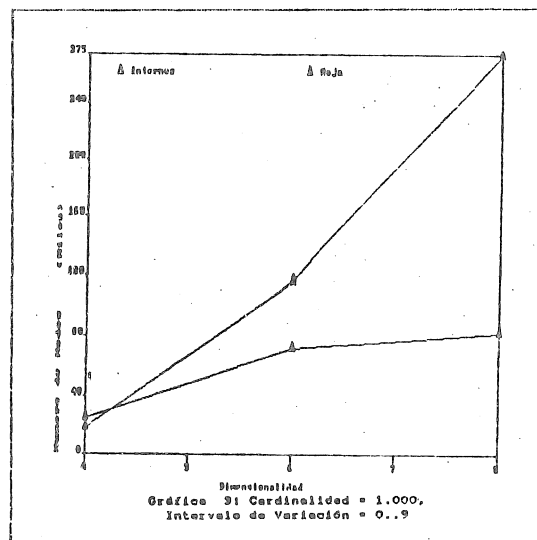
El objetivo de este trabajo es el estudio experimental del coato de almacenamiento y el tiempo de respuesta para los tipos de interrogaciones anteriormente citados, a través de los siguientes parámetros:

- 1) El número de nodos generados en la construcción de los árboles.
- 2) El número de accesos y comparaciones en los distintos tipos de interrogaciones, independientemente del número de claves sin especificar y el tamaño del hipercubo de interrogación.
- 3) El número de accesos y comparaciones en las interrogaciones parciales y en rango, en función de la variación del número de claves sin especificar y el tamaño del hipercubo de interrogación.

3.3) NUMERO DE NODOS GENERADOS EN LA CONSTRUCCION DE LOS ARBOLES:

La primera parte de las pruebas consiste en la creación de los árboles quintarios correspondientes a cada uno de los ficheros y la medición del número de nodos, tanto internos como hojas. Esto da una idea de los grandes volúmenes de memoria que pueden ocupar dichos árboles quintarios, y como puede influir en este aspecto la variación de alguno de los parámetros.





Al observar las gráficas 1, 2 y 3 se puede comprobar, tal y como se esperaba, que el número de nodos generados se incrementa al aumentar cualquiera de los parámetros.

En la gráfica 1 el crecimiento de los nodos hoja es, prácticamente, constante; el crecimiento más suave de los nodos internos indica el grado de compartición de los valores de clave de los registros.

En la gráfica 2 la diferencia entre el número de nodos internos y el número de hojas casi se mantiene constante.

Finalmente, el número de nodos internos se ve muy afectado ante una variación del número de claves, gráfica 3. Para 4 claves es menor que el número de nodos hoja, mientras que para 6 y 8 claves el número de nodos internos crece espectacularmente respecto al número de nodos hoja. Esto se debe a que al aumentar la dimensionalidad, el árbol tiene un mayor número de niveles, por tanto el número de nodos internos es mayor; sin embargo, el número de nodos hoja se encuentra limitado por el número de registros que permanece constante, y no puede crecer en la misma proporción.

3.4) NUMERO DE ACCESOS Y COMPARACIONES EN LOS DISTINTOS TIPOS DE INTERROGACIONES:

La segunda parte de las pruebas consiste en el estudio del número de accesos, comparaciones, celdas tratadas y registros recuperados en cada uno de los cuatro tipos de interrogaciones. Se entiende por número de accesos, el número de nodos internos a los que se accede durante la búsqueda. El número de comparaciones es el número de veces que se compara el campo de información de un nodo accedido con la interrogación. El número de celdas tratadas es el número de nodos hoja visitados durante el proceso. Y el número de registros recuperados es el número de registros que se obtienen debido a que cumplen las condiciones de la interrogación.

Todos los números aleatorios utilizados se generaron con

distribución uniforme.

El número de pruebas realizadas depende, en cierta forma, del tipo de interrogación:

Para el caso de la interrogación exacta se realizaron 500 pruebas.

En el caso de la interrogación parcial se realizaron pruebas con interrogaciones con una clave sin especificar, con dos, y hasta $nd-1$ claves sin especificar. De cada uno de estos casos se llevaron a cabo 500 pruebas. De esta forma, el número de pruebas realizadas viene dado por $500 \cdot (nd-1)$.

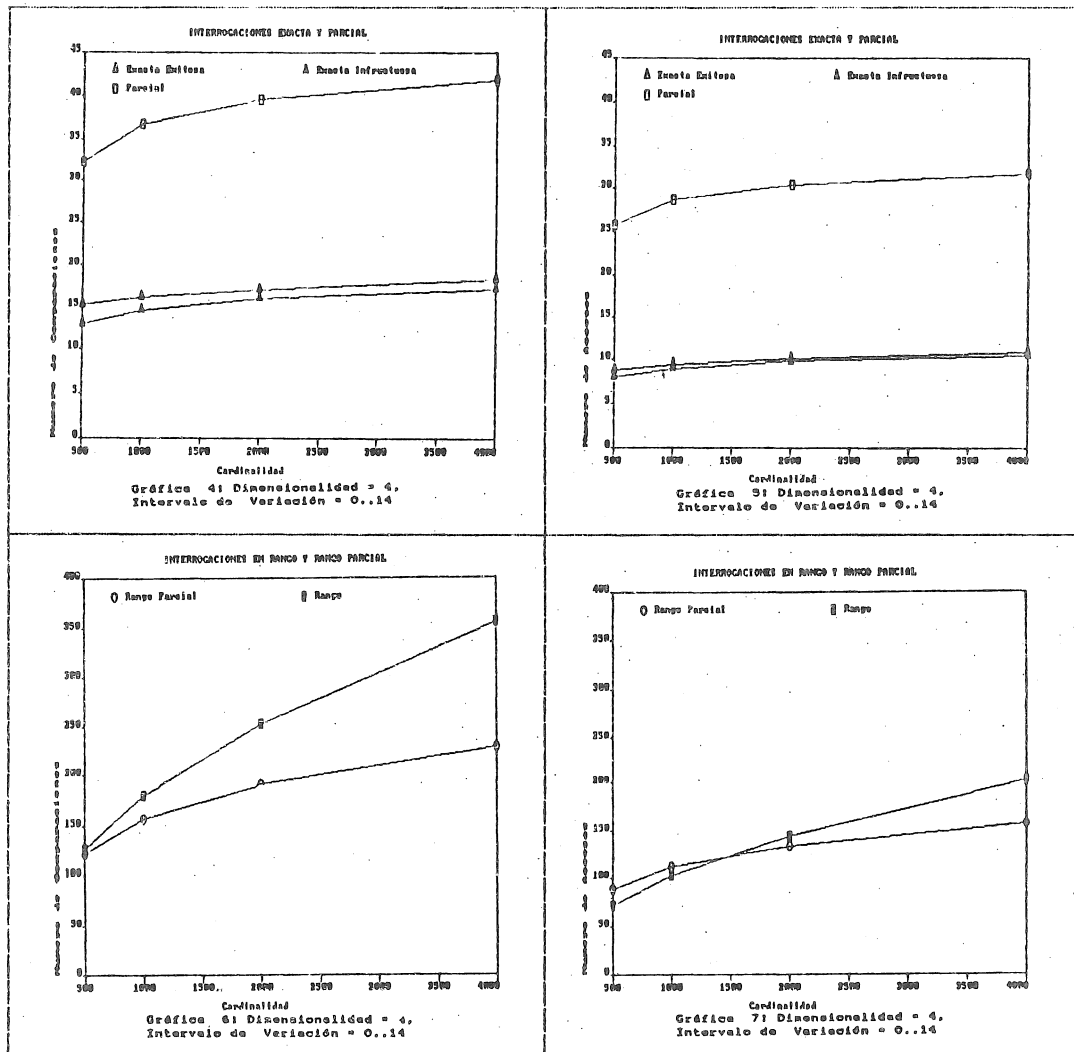
En el caso de la interrogación en rango se realizaron pruebas haciendo variar el tamaño del hipercubo desde 1 hasta 5, así el número total de pruebas realizadas fue de $500 \cdot 5 = 2.500$.

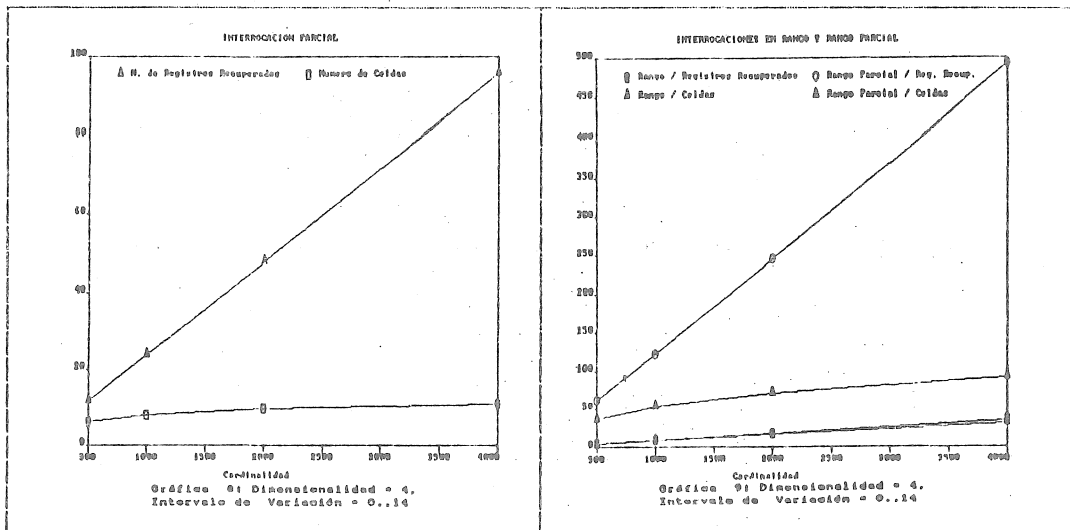
Finalmente, en el caso de la interrogación en rango parcial se varió por un lado, el número de claves sin especificar y, por otro, el tamaño del hipercubo. De acuerdo con lo anterior, el número de pruebas realizadas fue de $500 \cdot 5 \cdot (nd-1)$.

Para cada una de las 500 pruebas se necesitan nd números aleatorios limitados por el intervalo de variación de las claves, que representen los valores especificados por la interrogación. Además, en las interrogaciones parciales, se requieren $nd-1$ números aleatorios necesarios para indicar las posiciones de las claves sin especificar. Para las interrogaciones en rango se pueden utilizar como límites inferiores los nd valores generados, siempre que al calcular los límites superiores ninguno de ellos se salga del intervalo de variación de las claves, en caso de que esto ocurra se deben generar números que cumplan con esta restricción. Los límites superiores se calculan a partir de los límites inferiores, sumando a cada uno de ellos, inicialmente un 1, luego un 2, y así, hasta llegar a 5, de acuerdo con el tamaño del hipercubo de interrogación.

En la interrogación exacta las pruebas se realizaron para dos casos específicos: que todas las búsquedas resultaran exitosas y que todas resultaran infructuosas. Para el primer caso, se genera aleatoriamente un número entre 1 y el número de registros del fichero y éste, precisamente, será el registro a buscar. En el segundo caso se utiliza una interrogación aleatoria que cumpla con la condición de no existir en el fichero de datos. En la interrogación exacta no tiene sentido hablar de celdas tratadas puesto que en las búsquedas exitosas siempre se visita una celda, y en las infructuosas nunca se visita ninguna.

A) El parámetro variable es la cardinalidad:





Las gráficas anteriores muestran como influye la cardinalidad en el número de comparaciones, accesos, celdas tratadas y registros recuperados en los cuatro tipos de interrogaciones. Cualquiera de estos datos aumenta con el número de registros del fichero original. Esto es lógico si se parte del hecho de que un aumento en el número de registros produce un aumento en el número de nodos de árbol. Además, en general, el crecimiento del número de comparaciones, accesos y celdas es de menor orden a medida que aumenta la cardinalidad. Esto último se debe, fundamentalmente, a que los registros comparten más nodos internos.

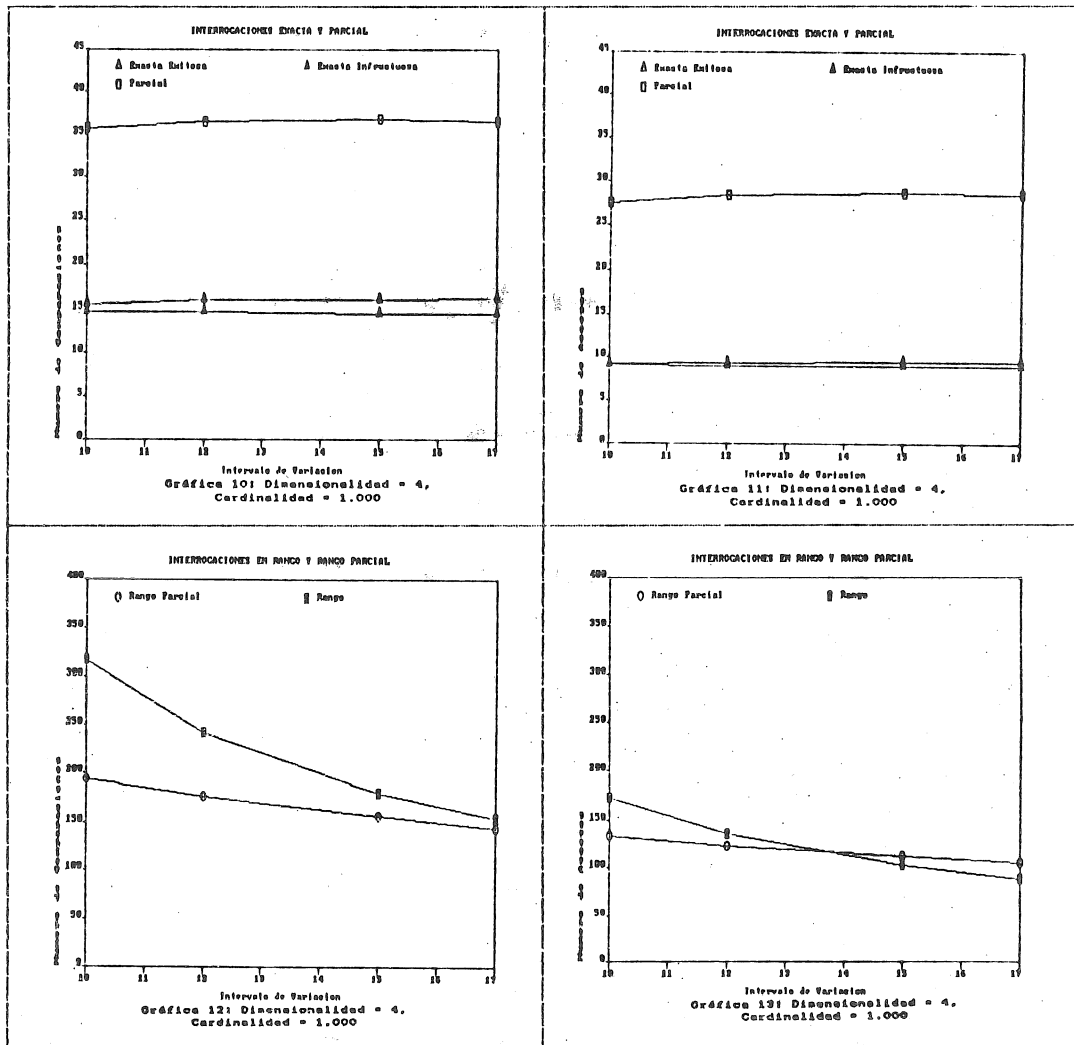
En las gráficas 4 y 5, para el caso de la interrogación exacta se obtiene que tanto el número de comparaciones como el de accesos es menor en las interrogaciones infructuosas que en las exitosas, aunque la diferencia no es muy notable. Esto se debe a que en las búsquedas infructuosas el proceso termina antes, una vez queda descartada la posibilidad de encontrar el registro; mientras que en las exitosas se llega hasta el nivel de nodos hoja para encontrar el registro solicitado. Para la interrogación parcial se observa un gran aumento tanto en el número de comparaciones como en el de accesos con respecto a la búsqueda anterior. Hecho muy natural teniendo en cuenta que en la búsqueda parcial, tras encontrar una clave sin especificar, se expande a través de los tres subárboles del nivel inferior.

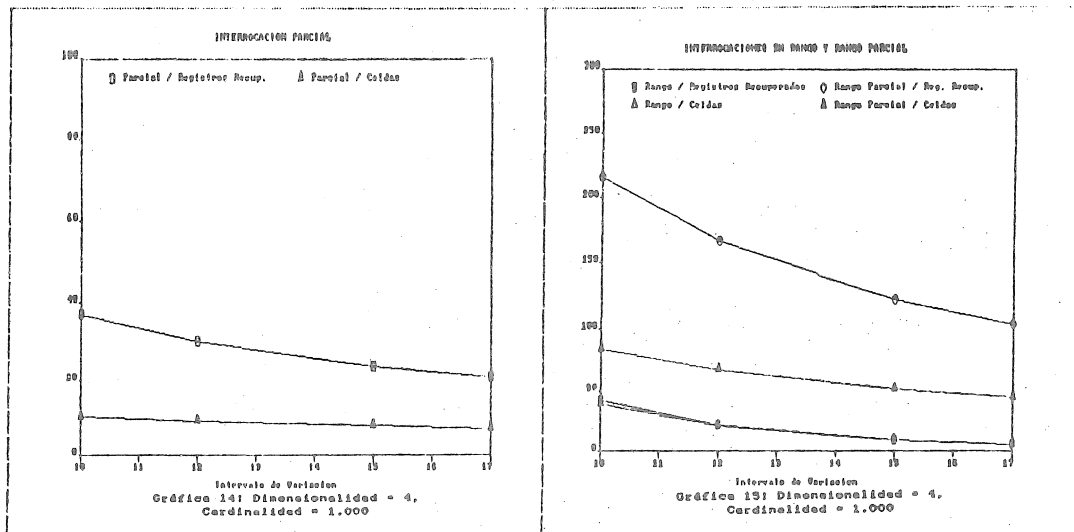
En las gráficas 6 y 7 se observa que el número de comparaciones y de accesos, en la interrogación en rango presenta un gran crecimiento ante un aumento de la cardinalidad, puesto que el tamaño de los árboles binarios de cada nivel crece con éstos. Sin embargo, la interrogación en rango parcial no se ve tan afectada debido a que al existir claves sin especificar, aunque se produce una expansión, se evita el recorrido de los árboles binarios de los niveles asociados a las claves no especificadas.

En las gráficas 8 y 9 se observa que en la interrogación en rango la diferencia entre el número de registros recuperados y el de celdas es mínimo. Sin embargo, en las interrogaciones parcial y en rango parcial el número de registros recuperados tiene un crecimiento constante y notable al aumentar la cardinalidad, mientras que el

número de celdas, aunque aumenta el crecimiento es ligero, esto es un indicador de lo que favorece el Arbol Quintario a las interrogaciones parciales. Además el número de registros recuperados es comparativamente inferior al de la interrogación en rango parcial, esto es debido a que las interrogaciones en rango representan un volumen de búsqueda pequeño con respecto al número de registros existente, mientras que el de las interrogaciones en rango parcial es mayor.

B) El parámetro variable es el intervalo de variación de los valores de clave:





En las gráficas 10 y 11 se muestran los resultados de las pruebas para la interrogación exacta. Las comparaciones en las búsquedas exitosas presentan un crecimiento significativo y las infructuosas presentan, incluso, oscilaciones en torno a la media. Un hecho parecido ocurre con el número de accesos.

Estos resultados responden, por un lado, a que al aumentar el intervalo de variación, las estructuras de árbol binario de cada nivel tienden a crecer, lo que explicaría el aumento del número de comparaciones y accesos; por otro lado, debido a que el número de registros permanece constante, y no es muy elevado, el crecimiento de las estructuras de árbol binario está limitado, y éstas tienen menor altura conforme se va descendiendo por el árbol. Al descender de nivel, no se utilizan todos los valores posibles del intervalo de variación de cada clave. Se realizaron pruebas con ficheros con un número de registros mayor en las que se comprobó que, efectivamente, disminuye el efecto del segundo factor mencionado.

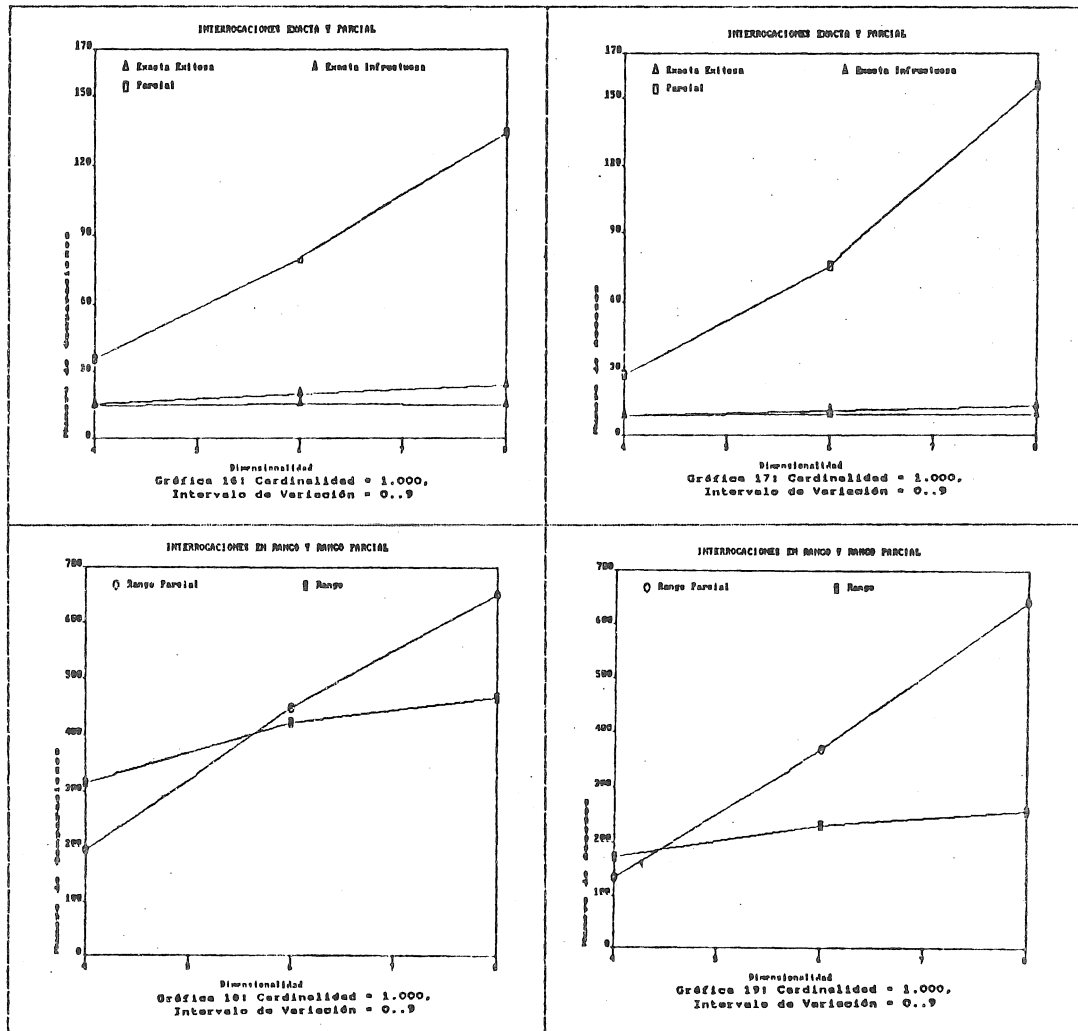
Para el caso de la interrogación parcial, influyen los mismos factores que en la interrogación exacta. Además, también hay un incremento general con respecto a ésta, por la expansión hacia los tres subárboles en el nivel inferior cuando se encuentra una clave sin especificar. Sin embargo, los registros recuperados, gráfica 14, presentan un decrecimiento a medida que el valor del intervalo de variación se hace mayor; esto se debe a que existe una mayor probabilidad de generar interrogaciones que supongan la recuperación de un menor número de registros. Así mismo, el número de celdas tratadas en la interrogación parcial, gráfica 14, presenta una ligera disminución al aumentar el tamaño del intervalo de variación de los valores de clave. Esto se debe a que aumenta la posibilidad de llegar a celdas vacías las cuales no se contabilizan.

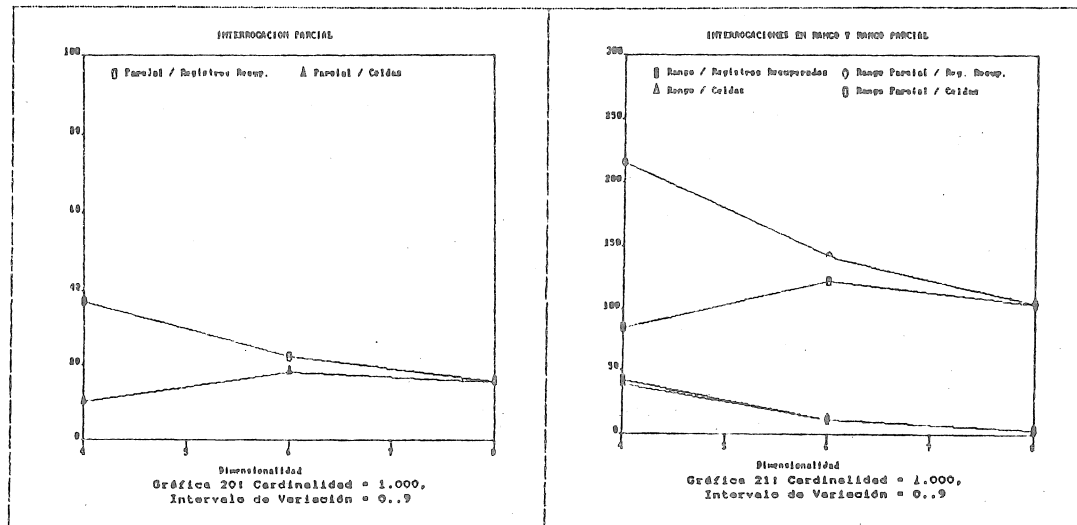
En las interrogaciones en rango y rango parcial, observando las gráficas 12 y 13 se tiene que, al aumentar el tamaño de intervalo de variación se produce un decrecimiento tanto en el número de comparaciones como en el de accesos en ambas búsquedas. Esto es una muestra de cómo está estructurado el árbol decreciendo la altura por niveles de los árboles binarios al aumentar el intervalo de

variación. Además, en este caso, al igual que se observó en el apartado A, la interrogación en rango se ve más afectada; el orden de decrecimiento de ésta es mayor. Se realizaron pruebas con ficheros con más registros, y los resultados también se suavizaron.

En los números de celdas y registros recuperados influye, tal como se observa en la gráfica 15, que hay una mayor probabilidad de generar una interrogación que permita la recuperación de un menor número de registros al aumentar el tamaño del intervalo de variación. Se vuelve a verificar la coincidencia de los valores de los números de celdas y registros recuperados para la interrogación en rango.

C) El parámetro variable es la dimensionalidad:





En las gráficas 16 y 17 se observa que en la interrogación exacta el número de accesos y de comparaciones aumenta con la dimensionalidad. En las búsquedas infructuosas, con un número de claves mayor es más probable detectar antes la no existencia del registro buscado, debido a esto, tanto el número de comparaciones como el de accesos es menor que para las exitosas.

En la interrogación parcial, los números de accesos y comparaciones aumentan más rápidamente a medida que aumenta la dimensionalidad, ya que el árbol tiene más niveles.

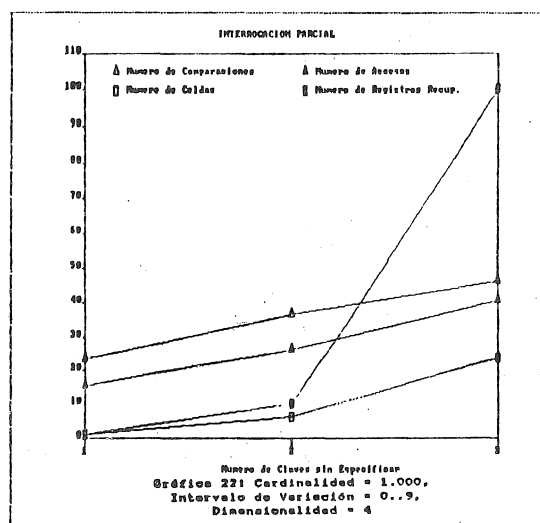
En las gráficas 18 y 19 se observa el comportamiento de la estructura en las interrogaciones en rango y rango parcial. Se muestra que en las interrogaciones en rango, aunque los números de accesos y comparaciones aumentan con la dimensionalidad, este crecimiento es menor que el registrado para la interrogación en rango parcial. Esto se debe a que en la interrogación en rango a medida que aumenta la dimensionalidad hay más búsquedas infructuosas por lo que se produce el mismo efecto que en la interrogación exacta con búsquedas infructuosas. De hecho, se realizaron pruebas con búsquedas exitosas y se produjo un incremento mayor. En la interrogación en rango parcial ocurre lo contrario, el crecimiento resultante es mayor debido a los mismos factores que influyen en la interrogación parcial.

En cuanto al número de celdas y registros recuperados, se pueden realizar las siguientes observaciones: en la interrogación en rango el número de celdas tratadas es pequeño, y disminuye conforme aumenta la dimensionalidad, debido a que las interrogaciones en rango seleccionadas de forma aleatoria, la mayoría de las veces no recuperan información por ser el número de registros muy pequeño, en comparación con el número de todas las posibles combinaciones que pueden tomar tales registros. En las interrogaciones parcial y en rango parcial se produce un crecimiento en el número de celdas al aumentar el número de claves de 4 a 6 y un ligero decrecimiento al aumentarlo de 6 a 8; esto último se debe a la limitación del número de registros. El número de registros recuperados siempre disminuye demostrando que decrece la probabilidad de encontrar un registro al

aumentar la dimensionalidad.

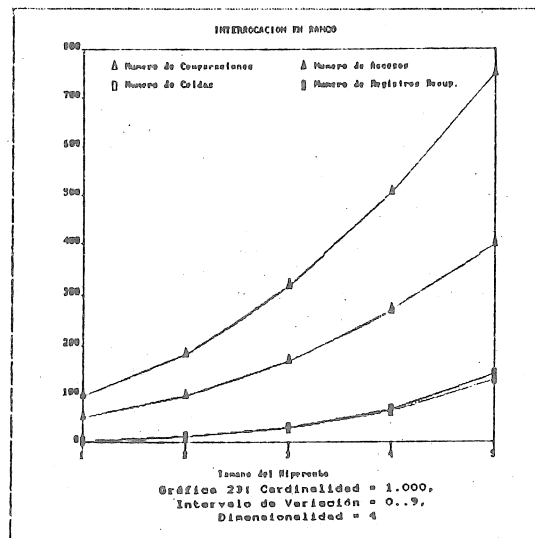
3.5) NUMERO DE ACCESOS Y COMPARACIONES EN FUNCION DEL NUMERO DE CLAVES SIN ESPECIFICAR Y EL TAMANO DEL HIPERCUBO:

A) Interrogación parcial en función del número de claves sin especificar:



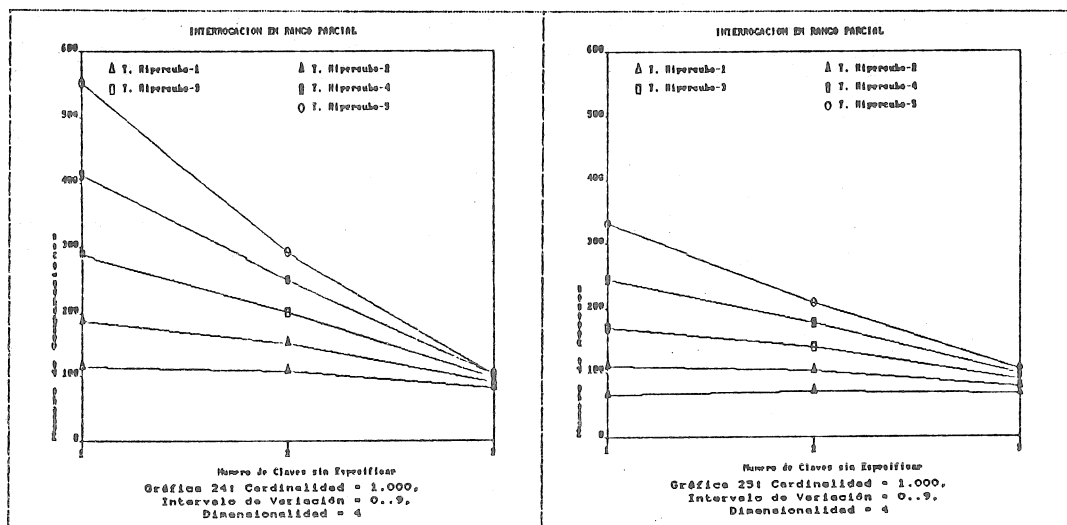
En la gráfica 22 se observa que a medida que crece el número de claves sin especificar, crece, asimismo, el número promedio de accesos, comparaciones, celdas y registros recuperados en una proporción mucho más representativa. Estos resultados eran los esperados si se tiene en cuenta que al aumentar el número de claves sin especificar, aumenta la probabilidad de que sean más los registros que cumplan con las características especificadas por la interrogación.

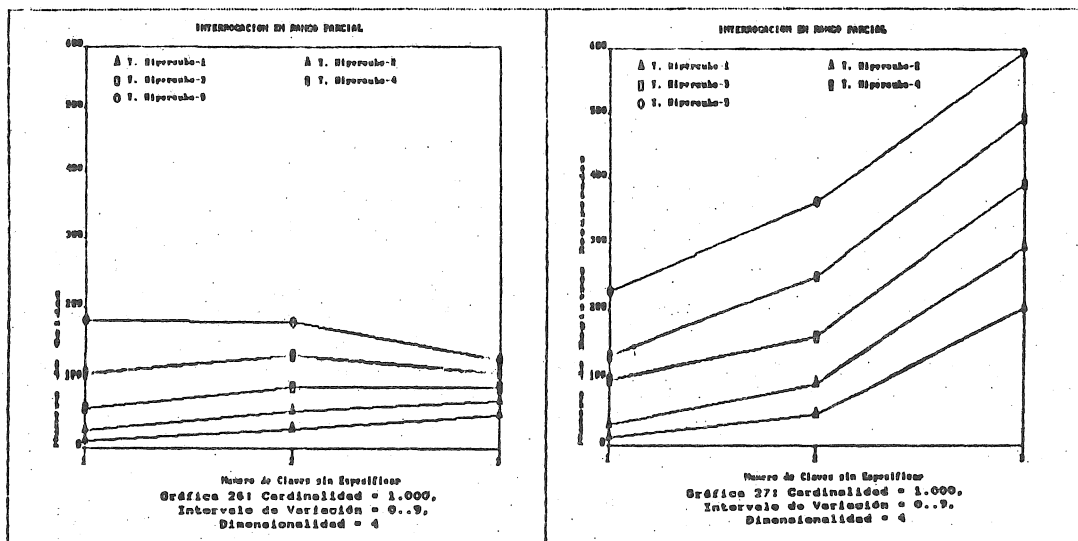
B) Interrogación en rango en función del tamaño del hipercubo:



En la gráfica 23 se observa el comportamiento de la interrogación en rango ante variaciones del tamaño del hipercubo. En cuanto a la variación del número de accesos, comparaciones, celdas tratadas y registros recuperados, se tiene que, a medida que el tamaño del hipercubo aumenta, evidentemente crecen los cuatro parámetros. Nuevamente, las curvas del número de celdas y registros recuperados casi coinciden.

C) Interrogación en rango parcial en función del tamaño del hipercubo y del número de claves sin especificar:





En las gráficas 24 y 25 se observa un incremento en el número de comparaciones y accesos al aumentar el tamaño del hipercubo, como cabe esperar ya que aumenta la información a recuperar, tal y como se observa en la gráfica 27, en la que, evidentemente, se refleja que el número de registros recuperados crece con el tamaño del hipercubo y con el número de claves sin especificar.

Para cada tamaño de hipercubo, al aumentar el número de claves sin especificar, el comportamiento general tiende a descender en los promedios de los números de comparaciones y accesos, debido a que una clave sin especificar significa tres accesos al siguiente nivel, mientras que una especificada supone un cierto número de comparaciones y accesos para situarse en todos los nodos cuyos campos de información estén dentro del rango especificado en la interrogación, y los accesos al siguiente nivel que son función del tamaño del hipercubo.

En la gráfica 26 el número de celdas visitadas crece con el tamaño del hipercubo, ahora bien, para cada tamaño la variación del número de claves sin especificar provoca un comportamiento variable de evolución continua, indicando que se accede a celdas con más de un registro.

4) ESTUDIO PROBABILISTICO DE LA INFORMACION RECUPERADA

Tras realizar un estudio probabilístico del número esperado de registros a recuperar en los tres tipos de interrogaciones, a través de:

Interrogación parcial:

$$\frac{\sum_{j=1}^{nd-1} vi \cdot j}{nd-1} = \frac{n}{vi \cdot nd}$$

Interrogación en rango:

$$\sum_{j=1}^{nd-1} \left(\sum_{p=2}^6 p^{** (nd-j)} \right) * vi^{**j} \quad n$$

Interrogación en rango parcial:

$$\frac{\sum_{p=2}^6 p^{***}nd}{5} = \frac{n}{vi^{***}nd}$$

Nótese que en este caso, v se trata como una constante ya que tiene el mismo valor para todas las claves del fichero.

Se comparan los resultados con los obtenidos en el estudio experimental, tal y como se observa en las tablas 2, 3 y 4, en la interrogación parcial los valores, prácticamente, coinciden; y en las interrogaciones en rango y en rango parcial, la diferencia es muy pequeña.

Cardinalidad	Registros Recuperados					
	I. PARCIAL		I. EN RANGO		I. EN RANGO PARCIAL	
	Experim.	Probab.	Experim.	Probab.	Experim.	Probab.
300	11,63	11,90	4,48	4,49	61,60	62,12
1.000	23,93	23,80	6,70	6,98	122,37	124,25
2.000	47,60	47,60	17,23	17,98	243,13	246,49
4.000	93,54	93,21	33,07	33,93	492,99	496,99

Tabla 2: Dimensionalidad = 4, Intervalo de Variación = 0..14

Intervalo de Variación	Registros Recuperados					
	I. PARCIAL		I. EN RANGO		I. EN RANGO PARCIAL	
	Experim.	Probab.	Experim.	Probab.	Experim.	Probab.
0.. 9	37,02	37,00	42,17	43,48	216,14	222,67
0..11	30,29	30,29	21,24	21,93	166,86	169,75
0..14	23,93	23,80	6,70	6,98	122,37	124,25
0..16	20,86	20,83	5,17	5,43	103,73	103,16

Tabla 3: Dimensionalidad = 4, Cardinalidad = 1.000

Dimensionalidad	Registros Recuperados					
	I. PARCIAL		I. EN RANGO		I. EN RANGO PARCIAL	
	Experim.	Probab.	Experim.	Probab.	Experim.	Probab.
4	37,02	37,00	42,17	43,48	216,14	222,67
6	22,21	22,22	11,82	13,43	143,09	147,56
8	15,61	15,67	3,67	4,29	103,39	106,41

Tabla 4: Cardinalidad = 1.000, Intervalo de Variación = 0..9

Además, las tablas 5, 6 y 7 muestran los resultados de dicho estudio para el caso concreto de variación del tamaño del hipercubo y número de claves sin especificar. Se comprueba que se repite la situación anterior.

Claves sin especificar	Registros Recuperados		Tamaño de hipercubo	Registros Recuperados	
	Experim.	Probab.		Experim.	Probab.
1	1,01	1,00	1	3,44	1,60
2	10,12	10,00	2	11,46	6,10
3	99,92	100,00	3	30,84	25,60
			4	69,22	62,60
			5	141,61	129,60

Tabla 5: Interrogación Parcial

Tabla 6: Interrogación en Rango

Tamaño del hipercubo	Claves sin especificar	Registros Recuperados	
		Experim.	Probab.
1	1	12,46	8,00
1	2	47,84	40,00
1	3	202,26	200,00
2	1	32,06	27,00
2	2	93,99	90,00
2	3	294,28	300,00
3	1	68,89	64,00
3	2	160,63	160,00
3	3	392,12	400,00
4	1	131,67	125,00
4	2	249,79	250,00
4	3	492,08	500,00
5	1	227,83	216,00
5	2	363,13	360,00
5	3	594,82	600,00

Tabla 7: Interrogación en Rango Parcial

CONCLUSIONES

Se presenta en este trabajo un estudio experimental de la estructura de Arbol Quintario respecto a la ocupación y al tiempo de respuesta para las interrogaciones exacta, parcial, en rango y en rango parcial.

Para valores moderados de los parámetros: cardinalidad, dimensionalidad e intervalo de variación de los valores de clave, se han obtenido árboles de grandes volúmenes; el parámetro que más influye es la dimensionalidad.

El tiempo de respuesta medido en función del número de accesos y comparaciones es bueno para interrogaciones exactas y parciales y aceptable para interrogaciones en rango y en rango parcial, siendo la dimensionalidad el parámetro que influye de forma negativa para este factor.

REFERENCIAS

1. BENTLEY, J.L., and FRIEDMAN, J.H. Data structures for range searching. Computing Surveys 11,4 (1979), 397-409.
2. CARDENAS, A.F. and SAGAMANG, J.P. Doubly-Chained Tree Data Base Organization-Analysis and Design Strategies. Comput. J. 20(1977), 15-26.
3. GUTING, H. and KRIEGEL, H.P. Multidimensional B-Tree: An Efficient Dynamic File Structure for Exact Match Queries, 1980.
4. GUTING, H. and KRIEGEL, H.P. Multidimensional B-Tree used for Partial Match Retrieval. Computer Science Technical Report, Universitat Dortmund, 1980.
5. KASHYAP, R.L.; SUBAS, S.K. and YAO, S.B. Analysis of the Multiple-Attribute-Tree Data Base Organization. IEEE Trans. Softw. Eng. SE-6 (1977) 451-467.
6. KNUTH, D. The Art of Computer Programing. Volumen 3/Sorting and Searching, Addison-Wesley Publishing Company, 1973.
7. LEE, D.T., and WONG, C.K. Quintary Trees: A file structure for multidimensional database systems, vol 5, N° 3, 1980, 339-353.
8. LIEN, Y.E., TAYLOR, C.E., and DRISCOLL, J.R. A design and implementation of a relational data base. Proc. Very Large Data Base Conf., 1975.
9. SANTANA, O.; CABRERA, J.; DIAZ, M. y MAYOR, O. Análisis Experimental de los Esquemas de Inserción del Arbol B. Informática Automática, 1986.
10. TREMBLAY, J.P. and SORENSON, P.G. An Introduction to Data Structures with Applications, Mc Graw Hill International Book Company, 1984.