Análisis del Coste de Búsquedas por Rango en Estructuras de Datos Multidimensionales[1]

A. Duch y C. Martínez[2]

Resumen

En este trabajo presentamos el análisis del coste promedio de búsquedas por rango en diversas estructuras de datos multidimensionales. En primer lugar, consideramos el caso de los árboles K-dimensionales relajados y más tarde extendemos nuestros resultados a otras estructuras de datos multidimensionales. Demostramos que el coste de las búsquedas por rango está relacionado con el coste de una variante de las búsquedas parciales, usando para ello una mezcla de argumentos combinatorios y geométricos. Esta reducción simplifica el análisis y nos permite obtener cotas inferiores y superiores del rendimiento y una útil caracterización del coste de las búsquedas por rango como suma de costes de búsquedas parciales. Estos resultados se emplean después para encontrar estimaciones asintóticas precisas del coste promedio de las búsquedas por rango.

1 Introducción

Las *búsquedas por rango* aparecen frecuentemente en aplicaciones de grandes bases de datos, sistemas de información geográfica, informática gráfica, etc. [Sam90]. Dada una colección de puntos K-dimensionales y un rectángulo, el objetivo de una búsqueda por rango (en inglés, *orthogonal range search*) es hallar todos los puntos de la colección que están en el interior del rectángulo dado. Aparte de las aplicaciones obvias e inmediatas de las búsquedas por rango, este tipo de búsquedas está implícita en búsquedas por región más complejas y en otras búsquedas *asociativas*.

Son muchas las estructuras de datos que se han propuesto para el mantenimiento y gestión de información multidimensional, y en particular, para soportar eficientemente las búsquedas por rango (véase por ejemplo [BF79]). Entre otras, están los árboles K-d [Ben75], los *quadtrees* [BF74], los *tries* K-dimensionales [Riv76] y múltiples variantes de los anteriores.

No obstante, el análisis matemático del rendimiento de las búsquedas por rango ha resultado ser una tarea difícil. Muchos trabajos anteriores se fundamentan en la hipótesis de que la estructura arborescente está perfectamente balanceada y en consecuencia los resultados obtenidos son excesivamente optimistas. Sólo recientemente ha habido un notable progreso en esta dirección con los trabajos de Devroye y sus coautores [CDZC01, DJZC00], donde se calculan cotas inferiores (Ω) y superiores ($\mathcal{O}$-grande) para el rendimiento promedio de las búsquedas por rango en *squarish K-d trees* (una variante de los árboles K-d) y otras estructuras de datos multidimensionales.

En nuestro trabajo hemos utilizado el mismo modelo aleatorio que en [CDZC01, DJZC00], pero obtenemos resultados más precisos, al obtener cotas inferiores y superiores exactas, y una caracterización precisa del coste de las búsquedas por rango como suma de los costes de búsquedas parciales *en rebanadas* (ver más abajo en esta misma sección). Usando estos resultados podemos dar una estimación asintótica ajustada del coste promedio de las búsquedas por rango. Nuestras técnicas de demostración son bastante diferentes de las utilizadas en [CDZC01, DJZC00] e involucran argumentos geométricos y combinatorios. Pero al igual que en los trabajos antes mencionados, son resultados de tipo general y fácilmente aplicables a una gran variedad de estructuras de datos multidimensionales.

El presente trabajo resume los resultados pre-

[1]Trabajo parcialmente subvencionado por el proyecto AEDRI (DGES PB98-0926) del Ministerio de Educación y Cultura, y el proyecto ALCOM-FT (IST-1999-14186) del Programa *Future and Emergent Technologies* de la Unión Europea.

[2]Dpto. de Lenguajes y Sistemas Informáticos. Universitat Politècnica de Catalunya. E-mail: {duch,conrado}@lsi.upc.es

sentados en [DM02a] y [DM02b].

1.1 Definiciones básicas

Sea $F = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, $n \geq 0$, la colección de n puntos K-dimensionales dada. Asumiremos que cada $x \in F$ es una K-tupla $x = (x_0, \ldots, x_{K-1})$ perteneciente a $[0,1]^K$.

Una *pregunta de rango* Q es un hiperrectángulo K-dimensional con lados paralelos a los ejes de coordenadas:

$$Q = [\ell_0, u_0] \times [\ell_1, u_1] \times \cdots \times [\ell_{K-1}, u_{K-1}],$$

donde $\ell_i \leq u_i$, para todo $0 \leq i < K$.

El modelo probabilístico que hemos empleado en nuestro análisis introduce una leve diferencia con el modelo presentado en [CDZC01, DJZC00]. En nuestro modelo, los lados de una pregunta de rango aleatorio tienen longitudes dadas $\Delta_0, \Delta_1, \ldots \Delta_{K-1}$, siendo $0 \leq \Delta_i \leq 1/2$, para toda $0 \leq i < K$ y el centro de la pregunta es un punto z obtenido independientemente al azar, de alguna distribución continua en

$$Z_\Delta = \prod_{0 \leq r < K} [-\Delta_r/2, 1 + \Delta_r/2]$$

Por lo tanto, $\ell_i = z_i - \Delta_i/2$ y $u_i = z_i + \Delta_i/2$, for $0 \leq i < K$. En nuestro modelo una pregunta de rango Q puede "caer" parcialmente fuera de $[0,1]^K$, por lo que en general lo que tendremos es

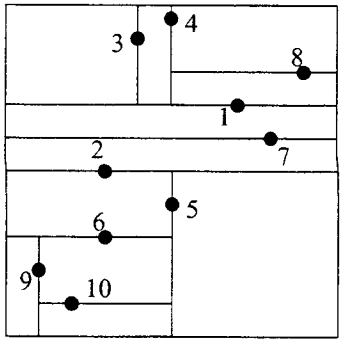
$$Q \subset C_\Delta = \prod_{0 \leq r < K} [-\Delta_r, 1 + \Delta_r].$$

1.2 Árboles K-d relajados

Un *árbol K-d relajado* [DECM98] que representa al conjunto F es un árbol binario en el cual se cumplen las dos siguientes condiciones: (a) cada nodo contiene uno de los n puntos K-dimensionales y tiene asociado un *discriminante* $j \in \{0, 1, \ldots, K-1\}$; (b) para cualquier nodo x con discriminante j se cumple que cualquier punto y almacenado en su subárbol izquierdo satisface $y_j < x_j$ y cualquier punto z en el subárbol derecho satisface $z_j \geq x_j$ (véase la figura 1).

Obsérvese que la secuencia de discriminantes en un camino desde la raíz hasta una hoja es arbitraria, al contrario de lo que sucede con los

$x_1 = (0.692, 0.703)$
$x_2 = (0.286, 0.495)$
$x_3 = (0.410, 0.895)$
$x_4 = (0.522, 0.953)$
$x_5 = (0.507, 0.394)$
$x_6 = (0.295, 0.300)$
$x_7 = (0.811, 0.605)$
$x_8 = (0.912, 0.807)$
$x_9 = (0.093, 0.210)$
$x_{10} = (0.188, 0.109)$



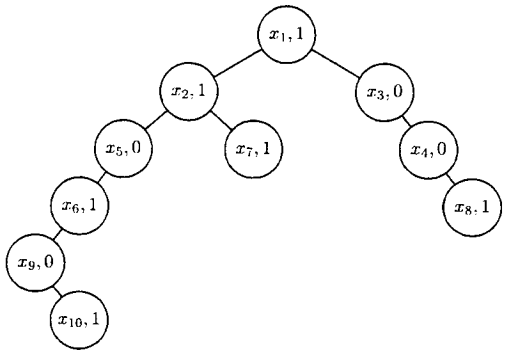


Figura 1: Un árbol 2-d relajado y la partición de $[0,1]^2$ inducida por dicho árbol

árboles K-d estándar [Ben75], donde la secuencia de discriminantes a lo largo de cualquier camino es cíclica; en particular, la raíz tiene discriminante 0, sus hijos tienen discriminante 1, y en general, los nodos del nivel m tienen discriminante $j = m \bmod K$.

En nuestro análisis asumiremos que los árboles K-d relajados son aleatorios, es decir, que se construyen insertando n puntos obtenidos de forma independiente de alguna distribución continua en $[0,1]^K$ y que los discriminantes asociados son independientes y obtenidos uniformemente al azar entre 0 y $K-1$. Esto significa que las $n!^K K^n$ posibles configuraciones de puntos/discriminantes son equiprobables [DECM98]. En particular, la probabilidad de que un árbol K-d relajado alaetorio de tamaño n tenga un subárbol izquierdo de tamaño ℓ y el discriminante de su raíz sea j es $1/(n \cdot K)$, para cualquier $j \in \{0, \ldots, K-1\}$ y cualquier ℓ,

$0 \leq \ell < n$.

Las búsquedas por rango en árboles K-d (en cualquiera de sus variantes, y en concreto en los relajados) son muy simples. Al visitar un nodo x con discriminante j se compara x_j con el j-ésimo rango $[\ell_j, u_j]$ de la pregunta. Si la pregunta está totalmente por encima (o por debajo) de x_j, la búsqueda continuará en el subárbol apropiado de x. Si, por el contrario, $\ell_j \leq x_j \leq u_j$ entonces la búsqueda continúa, recursivamente, en ambos subárboles de x. Adicionalmente habría de comprobarse si x está contenido en el rectángulo o no. Las invocaciones recursivas finalizan cuando se alcanzan subárboles vacíos.

Si en el esquema recursivo siempre continuásemos por ambos subárboles el coste sería $\Theta(n)$ ya que la totalidad del árbol sería explorada. Si, en cambio, en cada nodo visitado, la recursividad continuáse sólo por uno de los subárboles el coste (promedio) sería $\Theta(\log n)$. Pero en una búsqueda por rango la "alternancia" entre una y dos invocaciones recursivas siguen un patrón complejo originando un coste promedio que estará a medio camino entre logarítmico y lineal.

En nuestro análisis mediremos el coste de las búsquedas por rango como el número de nodos del árbol K-d visitados durante la búsqueda. Si el número de puntos contenidos en la respuesta es P, es evidente que el coste R_n de la búsqueda será de la forma $P + W_n$, siendo W_n lo que se conoce como *overhead*.

2 Análisis del coste de búsquedas por rango

2.1 Envolventes, rebanadas y búsquedas parciales

La *envolvente* (*bounding box*) $B(x)$ de un punto x en un árbol K-d t es la región de $[0,1]^K$ asociada a la hoja que fue reemplazada por x al ser insertado en t. Por ejemplo, si x está en la raíz de t entonces $B(x) = [0,1]^K$. Si z es el hijo izquierdo de la raíz x y la raíz discrimina respecto a la coordenada j entonces $B(z) = [0,1] \cdots \times [0, x_j] \times \cdots \times [0,1]$.

Lema 1 *Un punto x con envolvente $B(x)$ es visitado por una búsqueda por rango con pregunta Q si y sólo si $B(x)$ intersecta a Q.*

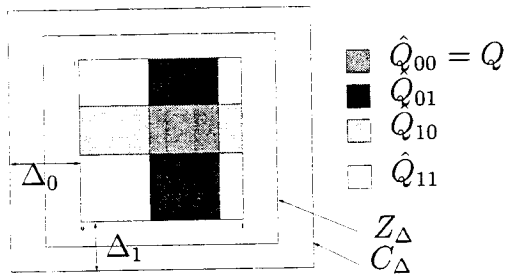


Figura 2: Ejemplo de rebanadas propias inducidas por una pregunta Q

Demostración: Véase por ejemplo [DJZC00]. ■

Dada una cadena de bits $w = w_0 \cdots w_{K-1}$, la *rebanada* (*slice*) Q_w es el hiperrectángulo K-dimensional definido por

$$Q_w = \prod_{0 \leq r < K} [\ell'_r, u'_r],$$

donde $[\ell'_i, u'_i] = [\max\{0, \ell_i\}, \min\{u_i, 1\}]$ si $w_i = 0$, y $[\ell'_i, u'_i] = [0,1]$ si $w_i = 1$. Por definición, $Q_{00\ldots0} = Q \cap [0,1]^K$ y $Q_{11\ldots1} = [0,1]^K$.

Una *rebanada propia* $\hat{Q}_w$ es la región definida por

$$\hat{Q}_w = Q_w - \bigcup_{v<w} Q_v,$$

donde $v < w$ si y sólo si $v_i < w_i$ para toda $0 \leq i < K$. En otras palabras, la rebanada propia $\hat{Q}_w$ es la región resultante de substraer todas las rebanadas estrictamente contenidas en Q_w (véase Fig. 2).

El concepto más importante de esta subsección es el de *búsqueda parcial en rebanadas*; pero antes revisamos brevemente las búsquedas parciales estándar. En una búsqueda parcial se nos da una pregunta q de la forma $q = (q_0, q_1, \ldots, q_{K-1})$, donde $q_i \in [-\Delta_i, 1 + \Delta_i] \cup \{*\}$ y el objetivo es hallar todos los puntos que satisfacen la pregunta, es decir, los puntos x tales que $x_i = q_i$ si $q_i \neq *$, para toda $0 \leq i < K$. Alternativamente, una pregunta q puede contemplarse como un par (y, w) donde y es un punto de C_Δ y w es una cadena de K bits llamada *patrón de especificación*, de manera que $q_i = y_i$ si $w_i = 1$ y $q_i = *$ si $w_i = 0$. Las búsquedas parciales tienen sentido cuando al menos una de las

coordenadas está especificada y al menos una de las coordenadas no lo está. El coste de las búsquedas parciales ha sido ampliamente estudiado para diversas estructuras multidimensionales (véase por ejemplo [DJZC00, FP86, KP94, MPP01]).

Dado un hiperrectángulo Q, una cadena de bits w y un punto $y \in C_\Delta$, una búsqueda parcial en rebanada actúa como una búsqueda parcial convencional con pregunta (y, w), pero a diferencia de ésta, la búsqueda parcial en rebanada sólo reportará un punto x si, además de ser visitado por la búsqueda parcial con pregunta (y, w), éste pertenece a $\hat{Q}_w$.

Para cualquier búsqueda parcial (en rebanada o estándar) con pregunta (y, w) definimos el hiperplano $H(y, w) = \{x \in C_\Delta \mid \forall i : w_i = 1 \implies x_i = y_i\}$. Por ejemplo, si $K = 2$ entonces $H(y, 00) = C_\Delta$, $H(y, 11) = \{y\}$, $H(y, 01)$ es la línea paralela al eje de abcisas que pasa por y y $H(x, 10)$ es la línea paralela al eje de ordenadas que pasa por x.

Lema 2 *Un punto x con envolvente $B(x)$ es visitado por una búsqueda parcial con pregunta (y, w), si y sólo si $B(x)$ intersecta el hiperplano $H(y, w)$.*

Un punto x con envolvente $B(x)$ es visitado y reportado por una búsqueda parcial en rebanada con hiperrectángulo Q, patrón de especificación w y punto y, si y sólo si $x \in \hat{Q}_w$ y la envolvente $B(x)$ intersecta el hiperplano $H(y, w)$.

2.2 Caracterización combinatoria

En esta subsección encontraremos varias relaciones entre el coste $R(t)$ de una búsqueda por rango en un árbol K-d t y el coste de búsquedas parciales en rebanada $P_w(t, y)$ con patrón de especificación w y punto y en el árbol t. En ambos casos, el hiperrectángulo Q será el mismo (y se deja implícito).

Teorema 3 *Sea Q un hiperrectángulo con esquinas $v_0, v_1, \ldots, v_{2^K-1}$. Para cualquier árbol K-d t*

$$R(t) \leq \sum_{0 \leq j < 2^K} \sum_{w \in (0+1)^K} P_w(t, v_j).$$

Demostración: Sea x un punto visitado por la búsqueda por rango y sea w la cadena de bits correspondiente a la rebanada propia $\hat{Q}_w$ que contiene a x. Puesto que x es visitado por la búsqueda por rango, sabemos que $B(x) \cap Q \neq \emptyset$ (Lema 1). Bastará, por el lema 2, demostrar que si $B(x)$ intersecta Q entonces existe al menos una esquina v_j de Q tal que el hiperplano $H(v_j, w)$ intersecta $B(x)$; al menos un término de la parte derecha de la desigualdad dada en el enunciado del teorema contabiliza entonces a x.

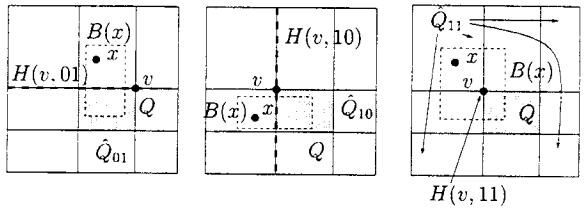


Figura 3: Ilustración gráfica de la demostración del Teorema 3

Si $B(x)$ contiene a alguna de las esquinas v_j de Q entonces claramente $B(x)$ intersecta a $H(v_j, w)$ ya que el hiperplano en cuestión también contiene a v_j. Si $B(x)$ no contiene a ninguna de las esquinas de Q —pero intersecta a Q— sólo existen dos posibilidades: $B(x)$ está completamente contenido en Q ó $B(x)$ interesecta una o más "caras" de Q. En el primer caso, se sigue que $w = 00\cdots0$, y entonces $H(v_j, 00\cdots0) = C_\Delta$ intersecta a $B(x)$, para cualquier v_j. En el segundo caso, puesto que $B(x)$ no contiene ninguna esquina y no está contenida en Q se sigue que $w \neq 11\cdots1$ y es fácil ver que una de las caras de Q intersectada por $B(x)$ tiene que ser a su vez parte de la frontera de $\hat{Q}_w$. Si v_j es una de las esquinas de la cara intersectada entonces $H(v_j, w)$ contiene a dicha cara y por tanto intersecta a $B(x)$ (véase Fig. 3 para una ilustración gráfica de esta demostración para el caso $K = 2$). ■

Teorema 4 *Sea Q un hiperrectángulo con centro en z. Para cualquier árbol K-d t,*

$$R(t) \geq \sum_{w \in (0+1)^K} P_w(t, z).$$

Demostración: El enunciado del teorema es inmediato, una vez que demostremos que todo punto reportado por una búsqueda parcial en rebanada con parámetros (z, w, Q) es también visitado por la búsqueda por rango con pregunta Q. Formalmente,

será suficiente demostrar que si $x \in \hat{Q}_w$ y $H(z,w)$ intersecta la envolvente $B(x)$ entonces $B(x)$ intersecta a Q.

Si $w = 00\ldots0$ lo anterior es trivialmente cierto ya que $x \in Q$. Por otro lado, si $w \neq 00\ldots0$ entonces $H(z,w)$ y $\hat{Q}_w$ son disjuntos. Puesto que $B(x)$ intersecta $\hat{Q}_w$ (x forma parte de ambas regiones por hipótesis), concluímos que $B(x)$ tiene que intersectar Q también. Entonces, por el lema 1, se sigue que x es visitado por la búsqueda por rango. ∎

La cota superior del Teorema 3 puede ser refinada, clasificando los costes por cuadrantes.

Teorema 5 *Dado un hiperréctangulo Q con centro en z, divídase C_Δ en 2^K cuadrantes $C_0, C_1, \ldots, C_{2^K-1}$, siendo z el centro de la partición. Sea $R^{(i)}(t)$ el número de puntos del cuadrante i-ésimo visitados por la búsqueda por rango con pregunta Q en el árbol t. Análogamente, sea $P_w^{(i)}(t,y)$ el número de puntos del i-ésimo cuadrante reportados por la búsqueda parcial en rebanada con parámetros (w,y,Q) en el árbol t. Sea v_i la (única) esquina de Q que pertenece al i-ésimo cuadrante. Entonces*

$$R^{(i)}(t) = \sum_{w \in (0+1)^K} P_w^{(i)}(t, v_i).$$

Demostración: Omitimos la prueba detallada. Baste decir, que en el análisis por casos utilizado en la demostración del Teorema 3, una inspección cuidadosa revela que si el punto considerado x está en el i-ésimo cuadrante entonces la esquina que hay que considerar en la búsqueda parcial en rebanada es necesariamente v_i. ∎

2.3 Coste esperado de las búsquedas por rango

Los teoremas de la subsección previa nos muestran como se puede reducir el análisis de la búsqueda por rango al análisis de búsquedas parciales en rebanada.

En esencia, tomando esperanzas matemáticas (el argumento es un poco más delicado) en ambos lados del Teorema 5 tendremos una caracterización del coste promedio de las búsquedas por rango como suma de costes promedio de búsquedas parciales en rebanada.

Teorema 6 *Sea $\mathbb{E}[R_n]$ el coste esperado de una búsqueda por rango con pregunta aleatoria en un árbol K-d aleatorio de tamaño n. Sea $\mathbb{E}[P_{n,w}]$ el coste esperado de una búsqueda parcial en rebanada con patrón w en un árbol K-d aleatorio de tamaño n, con respecto a un hiperrectángulo aleatorio Q y un punto aleatorio indpendiente en $[0,1]^K$. Entonces*

$$\frac{1}{V(Z_\Delta)} \cdot \sum_{w \in (0+1)^K} \mathbb{E}[P_{n,w}] \leq \mathbb{E}[R_n],$$

$$\mathbb{E}[R_n] \leq V(Z_\Delta) \cdot \sum_{w \in (0+1)^K} \mathbb{E}[P_{n,w}],$$

donde $V(Z_\Delta) = \prod_{0 \leq r < K}(1 + \Delta_r)$.

Ahora necesitamos hallar los valores de $\mathbb{E}[P_{n,w}]$ en árboles K-d aleatorios. El siguiente resultado expresa la intuición de que el coste buscado es el de una búsqueda parcial estándar multiplicado por la probabilidad de que un punto esté en la rebanada propia $\hat{Q}_w$ (a la que llamaremos *volumen* de la rebanada). Puesto que el coste esperado de las búsquedas parciales estándar en árboles K-d relajados es bien conocido [MPP01], obtenemos el siguiente resultado.

Teorema 7 *Sea orden(w) el número de coordenadas especificadas en w (el número de unos en w). Si $w \neq 00\ldots0$ y $w \neq 11\ldots1$, entonces el coste esperado $\mathbb{E}[P_{n,w}]$ de una búsqueda parcial en rebanada sobre un árbol K-d relajado aleatorio de tamaño n con respecto al patrón w, un punto aleatorio en $[0,1]^K$ y un hiperrectángulo aleatorio Q es*

$$\mathbb{E}[P_{n,w}] = \mathit{Vol}(\hat{Q}_w) \cdot \beta(\rho) \cdot n^{\alpha(\rho)} + \mathcal{O}(1),$$

donde $\rho = \mathit{orden}(w)/K$, $\alpha \equiv \alpha(x) = (\sqrt{9-8x} - 1)/2$, $\beta(x) = \Gamma(2\alpha+1)/((1-x)(\alpha+1)\alpha^3\Gamma^3(\alpha))$, y $\mathit{Vol}(\hat{Q}_w)$ es la probabilidad de que un punto x pertenezca a la rebanada propia $\hat{Q}_w$ de un hiperrectángulo aleatorio Q. Adicionalmente, $\mathbb{E}[P_{n,00\ldots0}] = \mathit{Vol}(\hat{Q}_{00\ldots0}) \cdot n$ y $\mathbb{E}[P_{n,11\ldots1}] = 2 \cdot \mathit{Vol}(\hat{Q}_{11\ldots1}) \cdot (H_{n+1} - 1)$, donde $H_n = \sum_{1 \leq j \leq n} 1/j = \log n + \gamma + \mathcal{O}(1/n)$ denota el n-ésimo número armónico.

El cálculo de los volúmenes $\mathrm{Vol}(\hat{Q}_w)$ puede ser una tarea ardua, en tanto que nuestro modelo admite que los hiperrectángulos Q caigan parcialmente fuera de $[0,1]^K$ y especialmente si los puntos

considerados no están uniformemente distribuídos. Adicionalmente, si los Δ_i's son grandes la diferencia entre las cotas inferior y superior del Teorema 6 es significativa y el modelo empleado pierde interés ya que el "marco" en torno a la región donde están los datos es demasiado grande. Pero si los Δ_i's son pequeños (tienden a 0 para $n \to \infty$, es decir, el número esperado de puntos que caen dentro del hiperrectángulo no crece linealmente con n) y los puntos están uniformemente distribuídos en $[0,1]^K$ entonces es sencillo establecer el siguiente corolario.

Corolario 8 *Dado un árbol K-d relajado aleatorio que almacena n puntos obtenidos independiente y uniformemente al azar en $[0,1]^K$, el coste esperado de una búsqueda por rango con una pregunta de dimensiones $\Delta_0, \ldots, \Delta_{K-1}$ tales que $\Delta_i \to 0$ para $n \to \infty$ y centro independiente y uniformemente obtenido al azar en Z_Δ satisface*

$$\mathbb{E}[R_n] \sim \Delta_0 \cdots \Delta_{K-1} \cdot n + \sum_{1 \le j < K} c_j \cdot n^{\alpha(j/K)} + 2 \cdot (1-\Delta_0) \cdots (1-\Delta_{K-1}) \cdot \log n + \mathcal{O}(1),$$

donde

$$c_j = \beta(j/K) \cdot \sum_{w:orden(w)=j} \left(\prod_{i:w_i=0} \Delta_i \right) \cdot \left(\prod_{i:w_i=1} (1-\Delta_i) \right).$$

Demostración: El corolario resulta de combinar el Teorema 6 y la estimación asintótica del coste esperado de búsquedas parciales en rebanada dada en el Teorema 7. Puesto que $\Delta_i \to 0$ para $n \to \infty$, se sigue que $V(Z_\Delta) \to 1$ y entonces las cotas inferior y superior del Teorema 6 se igualan, obteniéndose

$$\mathbb{E}[R_n] \sim \sum_{w \in (0+1)^K} \mathbb{E}[P_{n,w}].$$

Finalmente, para la distribución uniforme y puesto que los Δ_i's son suficientemente "pequeños" tenemos $\mathrm{Vol}(\hat{Q}_w) \sim \left(\prod_{i:w_i=0} \Delta_i\right) \cdot \left(\prod_{i:w_i=1}(1-\Delta_i)\right)$. ■

Es importante resaltar que el término $\Delta_0 \cdots \Delta_{K-1} \cdot n$ en $\mathbb{E}[R_n]$ es el número esperado de puntos que la búsqueda por rango reportará y en consecuencia el *overhead* en el coste es $\mathcal{O}(n^{\alpha(1/K)})$.

3 Otras estructuras de datos multidimensionales

Las caracterizaciones combinatorias obtenidas en la sección previa no hacían ninguna hipótesis particular sobre la partición inducida por el árbol K-d, de manera que es bastante inmediato generalizar el análisis a todo tipo de estructuras de datos multidimensionales, siempre y cuando sean jerárquicas e induzcan particiones del espacio $[0,1]^K$ en hiperrectángulos ortogonales a los ejes de coordenadas. Ejemplos concretos incluyen los árboles K-d estándar [Ben75], los *squarish K-d trees* [DJZC00], los árboles K-d-t [CLF89], los quadtrees [BF74], los K-d tries [Riv76], los K-d tries relajados [MPP01], etc.

Las diferencias radicarán en los costes de las búsquedas parciales en una u otra estructura de datos. En general, si $w \neq 00\ldots0$ y $w \neq 11\ldots1$ tendremos

$$\mathbb{E}[P_{n,w}] = \beta_w \cdot \mathrm{Vol}(\hat{Q}_w) \cdot n^{\alpha(\rho)} + \mathcal{O}(1),$$

donde $\rho = \mathrm{orden}(w)/K$, $\alpha(x) = 1 - x + \phi(x)$ y β_w es una constante dependiente de w. Y entonces el coste esperado de las búsquedas por rango, cuando $\Delta_i \to 0$, será de la forma

$$\mathbb{E}[R_n] = \Delta_0 \cdots \Delta_{K-1} \cdot n + \sum_{1 \le j < K} c_j \cdot n^{\alpha(j/K)} + 2 \cdot (1-\Delta_0) \cdots (1-\Delta_{K-1}) \cdot \log n + \mathcal{O}(1), \quad (1)$$

donde $c_j = \sum_{w:\mathrm{orden}(w)=j} \beta_w \cdot \mathrm{Vol}(\hat{Q}_w)$,

Las diferentes estructuras de datos estarán caracterizadas por distintos valores de α y de las constantes β_w. Por ejemplo, para los árboles K-d estándar el análisis de la búsqueda parcial aparece en [FP86] y tenemos que $\alpha(x) = 1 - x + \phi(x)$ donde $\phi(x) < 0.07$ es la única solución real de la ecuación

$$(\phi(x) + 3 - x)^x (\phi(x) + 2 - x)^{1-x} - 2 = 0.$$

No se conoce una expresión cerrada simple para los valores de β_w, pero se pueden calcular de manera explícita con algo de esfuerzo (mayor cuanto mayor es K). Los valores de las β's para todos los posibles patrones con $K \le 4$ se dan en [FP86].

Referencias

[Ben75] J.L. Bentley. Multidimensional binary search trees used for associative retrieval. *Comm. ACM*, 18(9):509–517, 1975.

[BF74] J.L. Bentley y R.A. Finkel. Quad trees: A data structure for retrieval on composites keys. *Acta Informatica*, 4:1–9, 1974.

[BF79] J.L. Bentley y J.H. Friedman. Data structures for range searching. *ACM Computing Surveys*, 11(4):397–409, 1979.

[CDZC01] P. Chanzy, L. Devroye y C. Zamora-Cura. Analysis of range search for random k-d trees. *Acta Informatica*, 37:355–383, 2001.

[CLF89] W. Cunto, G. Lau y Ph. Flajolet. Analysis of kdt-trees: kd-trees improved by local reorganisations. En F. Dehne, J.-R. Sack y N. Santoro, editores, *Workshop on Algorithms and Data Structures (WADS'89)*, volumen 382 de *Lecture Notes in Computer Science*, páginas 24–38. Springer-Verlag, 1989.

[DECM98] A. Duch, V. Estivill-Castro y C. Martínez. Randomized k-dimensional binary search trees. En K.-Y. Chwa y O. H. Ibarra, editores, *Int. Symposium on Algorithms and Computation (ISAAC'98)*, volumen 1533 de *Lecture Notes in Computer Science*, páginas 199–208. Springer, 1998.

[DJZC00] L. Devroye, J. Jabbour y C. Zamora-Cura. Squarish k-d trees. *SIAM J. Comput.*, 30:1678–1700, 2000.

[DM02a] A. Duch y C. Martínez. On the average performance of orthogonal range search in multidimensional data structures. En *Actas de 29th Int. Col. on Automata, Languages and Programming (ICALP)*, Lecture Notes in Computer Science. Springer, 2002. Aceptado para publicación.

[DM02b] A. Duch y C. Martínez. On the average performance of orthogonal range search in multidimensional data structures. *J. Algorithms*, 2002. Aceptado para publicación.

[FP86] Ph. Flajolet y C. Puech. Partial match retrieval of multidimensional data. *J. Assoc. Comput. Mach.*, 33(2):371–407, 1986.

[KP94] P. Kirschenhofer y H. Prodinger. Multidimensional digital searching–alternative data structures. *Random Structures & Algorithms*, 5(1):123–134, 1994.

[MPP01] C. Martínez, A. Panholzer y H. Prodinger. Partial match queries in relaxed multidimensional search trees. *Algorithmica*, 29(1–2):181–204, 2001.

[Riv76] R. L. Rivest. Partial-match retrieval algorithms. *SIAM J. Comput.*, 5(1):19–50, 1976.

[Sam90] H. Samet. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley, 1990.