

UNIDAD 5: ALGORITMOS DE MEMORIA EXTERNA

BÚSQUEDA

Gibran Fuentes Pineda

Junio 2020

BÚSQUEDA DE ELEMENTOS

- Dado un arreglo de N elementos, determinar si un elemento de consulta q está en el arreglo
- Recorriendo todos los elementos tomaría $O(N/B)$ operaciones E/S
- Si están ordenados podríamos usar búsqueda binaria y tomaría $O(\log_2 \frac{N}{B})$ operaciones E/S

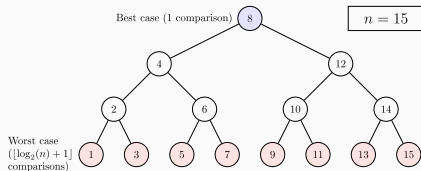


Imagen tomada de Wikipedia (Binary search algorithm)

- *Hashing*
 - Métodos con directorios
 - Métodos sin directorios
- Algoritmos de árboles
 - Árboles-B y variantes
 - Árboles de *buffer*

- Directorio
 - Consiste de una tabla con 2^d referencias, donde $d \geq 0$
 - La ubicación en la tabla de un elemento está dada por los d bits menos significativos de su valor *hash*

$$hash_d(x) = hash(x) \bmod 2^d$$

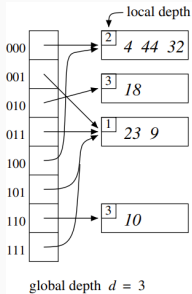
- Cada ubicación de la tabla contiene una referencia a un bloque donde se almacenan los elementos
- Una ubicación comparte el mismo bloque con las ubicaciones con los mismos k bits menos significativos

$$hash_k(x) = hash_d(x) \bmod 2^k$$

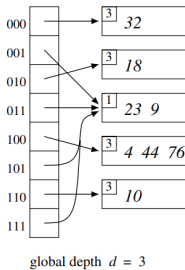
- Búsquedas requieren 2 operaciones E/S si el directorio reside en disco
 - Acceso a ubicación de la tabla
 - Acceso a bloque asignado a ubicación

- d es la profundidad global
 - Se elige el número más pequeño con el cual cada ubicación de la tabla tiene a lo mucho B elementos
- k es la profundidad local
 - Se elige el número más pequeño para que los elementos asignados entren en un solo bloque
- Cuando un bloque se desborda, d y k se recalculan
 - Se divide el bloque y se redistribuyen sus elementos
 - El tamaño del directorio crece (se duplica cuando d se incrementa en 1)
 - Permite adaptarse al crecimiento en el número de elementos N

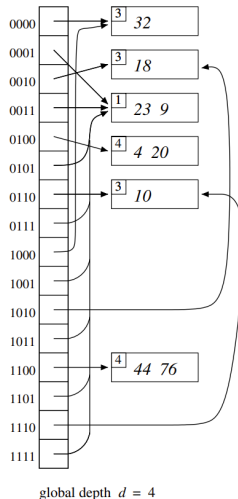
EJEMPLO DE HASHING EXTENDIBLE CON BLOQUE $B = 3$



(a)



(b)



- No necesita directorio: cada ubicación de la tabla corresponde a un bloque donde se almacenan elementos
- Cuando hay desbordamiento, un bloque predefinido de la tabla se divide y sus elementos se redistribuyen
 - El bloque que se divide no es necesariamente el que se desbordó, por lo que los bloques desbordados necesitan listas auxiliares para almacenar elementos desbordados
- Búsquedas requieren usualmente una sola operación E/S

- Árboles balanceados que mantienen sus elementos ordenados
 - Operaciones: escaneo secuencial, inserción, eliminación y búsqueda
- Definición de árbol-B de orden m ¹
 - Cada nodo tiene a lo mucho m hijos
 - Cada nodo no hoja tiene al menos $\lceil m/2 \rceil$ hijos (excepto la raíz)
 - Si no es hoja, el nodo raíz tiene al menos 2 hijos
 - Un nodo no hoja con k hijos contiene $k - 1$ valores de separación (llaves)
 - Los nodos hoja están en el mismo nivel del árbol
 - Elementos de un nodo están ordenados

¹De acuerdo a D. Knuth: *Sorting and Searching, The Art of Computer Programming*, vol. 3.

EJEMPLO DE ÁRBOL-B

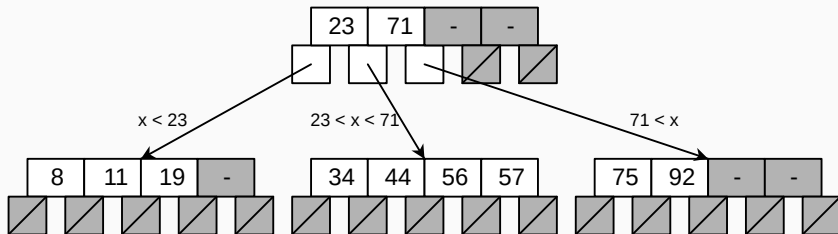


Imagen tomada de Wikipedia (Árbol-B)

INSERCIÓN EN ÁRBOLES-B

- Se hacen de nodos hoja a raíz
 1. Se encuentra el nodo hoja donde insertar el elemento
 2. Si el nodo hoja tiene menos elementos que el máximo, se inserta en ese nodo
 3. En caso contrario, se divide en 2
 - a) Escoge mediana de elementos del nodo y nuevo elemento
 - b) Valores menores a mediana se agregan en hijo izquierdo y mayores en hijo derecho
 - c) La mediana se inserta en nodo padre, si excede capacidad se divide en 2 (se repite a)–c))
 - d) Si se llega a raíz, se crea nueva

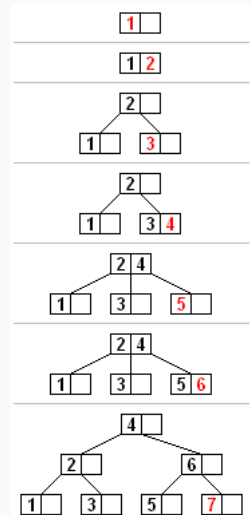


Imagen tomada de Wikipedia (B-tree)

- Eliminación en un nodo hoja
 1. Se busca el elemento a eliminar
 2. Si el valor se encuentra en un nodo hoja, se elimina
 3. Si queda con muy pocos elementos, se rebalancea
- Eliminación en un nodo no hoja (interno)
 1. Encuentra un nuevo separador que substituya al elemento
 2. Toma elemento más grande del hijo izquierdo o el más pequeño del derecho, elimínalo del sub-árbol correspondiente y reemplaza con este el elemento a eliminar
 3. Rebalancea sub-árbol, si es necesario

1. Se recorre el árbol de la raíz a las hojas
2. Si el elemento está almacenado en el nodo la búsqueda termina y si no se continúa con el sub-árbol correspondiente al valor a buscar
3. Si estamos en un nodo hoja y el elemento no está almacenado, entonces no se encontró

ÁRBOLES-B+

- Árboles-B+ son una variante en la cual todos los elementos se almacenan en las hojas y los nodos internos solo almacenan llaves y referencias

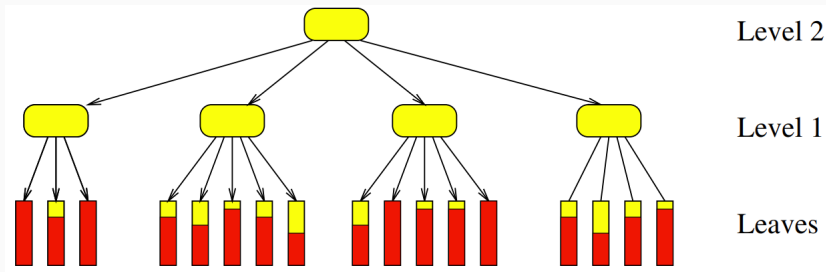


Imagen tomada de Vitter. *Algorithms and Data Structures for External Memory*, 2008.