



1



Universidad Nacional Autónoma de México
Facultad de Ingeniería
Algoritmos y estructuras de datos
Tema 7:
MÉTODOS DE BÚSQUEDA



7 Métodos de búsqueda

Objetivo: Aplicar el método de búsqueda apropiado a conjuntos de datos residentes, tanto en la memoria principal, como en la memoria secundaria. Diseñar y aplicar algoritmos.



7 Métodos de búsqueda

7.1 Generalidades.

7.2 Definición de la operación de búsqueda.

7.3 Búsqueda por comparación de llaves.

7.4 Búsqueda por transformación de llaves.



7.1 Generalidades.



7.1 Generalidades.

En el procesamiento de información se poseen diversas estructuras de datos sobre las cuales se pueden efectuar varias operaciones. Las operaciones más frecuentes en la estructuras de datos son agregar, eliminar y consultar.

La operación de consulta, a su vez, puede ser de actualización (modificar un elemento), de lectura (obtener los datos de un elemento) o de verificación (comprobar la existencia de un elemento).



Las operaciones de actualización, lectura o verificación requieren de una operación común denominada búsqueda.

La operación de búsqueda se puede llevar a cabo sobre elementos ordenados o desordenados.



7.2 Definición de la operación de búsqueda.



7.2 Definición de la operación de búsqueda.

Los métodos de búsqueda se pueden clasificar en internos y externos.

Los métodos de búsqueda internos se realizan cuando el conjunto de elementos sobre el que se aplica la consulta se encuentran en memoria principal. Los métodos de búsqueda externos se realizan cuando el conjunto de elementos sobre el que se aplica la consulta se encuentran en memoria secundaria.



Los métodos de búsqueda se pueden realizar tanto en conjuntos que residen en memoria principal como en los que residen en memoria secundaria. Los principales métodos de búsqueda son:

- **Secuencial o lineal**
- **Binaria**
- **Por transformación de claves**
- **Árboles de búsqueda**



10



7.3 Búsqueda por comparación de llaves.



7.3 Búsqueda por comparación de llaves.

Los métodos de búsqueda que se analizarán en este tema se basan en la búsqueda directa de una llave a través de las llaves de los nodos de la estructura de datos.

Las búsquedas que se basan en la comparación de llaves son la lineal y la binaria.



7.3.1 Búsqueda lineal

La búsqueda lineal, también llamada secuencial, es el método de búsqueda más sencillo, consiste en revisar elemento tras elemento hasta encontrar el dato buscado o llegar al final del conjunto de datos.

Este método se puede aplicar a estructuras ordenadas o desordenadas. Se suele utilizar cuando el conjunto de datos es pequeño. Si los elementos residen en memoria secundaria, hay que considerar el tiempo de acceso al dispositivo, adicional al tiempo de recorrido.

Ejemplo 1



13

Sea A un conjunto de elementos contenidos en una estructura de datos lineal:

A	6	7	3	15	0	4	23	22	11	10	9	17
---	---	---	---	----	---	---	----	----	----	----	---	----

Ejemplo 1



A

6	7	3	15	0	4	23	22	11	10	9	17
---	---	---	----	---	---	----	----	----	----	---	----

- a) Para conocer si el número 23 se encuentra en la estructura se necesitan siete comparaciones.
- b) Se requieren un total de doce comparaciones para darse cuenta que un elemento no está en la estructura. Este es el peor caso de complejidad.
- c) El número promedio de comparaciones para buscar un elemento en la estructura está dado por el número de elementos entre 2, es decir: $n/2$. En este caso es 6.



El algoritmo del método de búsqueda lineal es:

```
FUNC buscarLineal (x: Entero) dev ENTERO
    i ← 1, n ← numElementos(A) : ENTERO
    MIENTRAS i < n Y A[i] <> x HACER
        i ← i + 1
    FIN_MIENTRAS
    SI (i > n) ENTONCES
        DEV -1
    FIN_SI
    EN_CASO_CONTRARIO
        DEV i
    FIN_ECC
FIN_FUNC
```



El algoritmo del método de búsqueda lineal anterior permite encontrar la primera recurrencia de un número, pero no las subsecuentes, es decir, si la clave 'x' se encuentra repetida en el conjunto, solo se devuelve la posición de la primera ocurrencia. El algoritmo se puede modificar para encontrar elementos repetidos pero el tiempo de se vuelve lineal $O(n)$.

El algoritmo anterior se puede hacer recursivo, para ello hay que considerar tres casos: cuando i es mayor a n , cuando $E[i]$ es igual a 'x' y el caso recursivo.



El análisis de complejidad del método de búsqueda lineal está dado por el número total de comparaciones.

El peor caso de complejidad se presenta cuando el elemento buscado (x) no se encuentra dentro del conjunto de n elementos, en tal caso se realizarán n comparaciones.

Si el elemento buscado (x) sí se encuentra en el conjunto en el mejor caso se realiza 1 comparación ($C_{\min} = 1$), en el peor caso se realizan n comparaciones ($C_{\max} = n$) y en el caso base se realizan i comparaciones ($C_{\text{med}} = n/2$).



7.3.2 Búsqueda binaria

El método de búsqueda binaria consiste en dividir el intervalo de búsqueda en dos partes y compara el elemento que ocupa la posición central del conjunto. Si el elemento del conjunto no es igual al elemento buscado se redefinen los extremos del intervalo, dependiendo de si el elemento central es mayor o menor que el elemento buscado, reduciendo así el espacio de búsqueda. El proceso termina cuando el elemento buscado es encontrado o cuando el intervalo de búsqueda llega a su fin.

El método de búsqueda binaria funciona, únicamente, con conjunto ordenados.

Ejemplo 2



19

Dado un conjunto de elementos contenidos en una estructura de datos lineal ordenado de manera ascendente, encontrar el elemento 325.

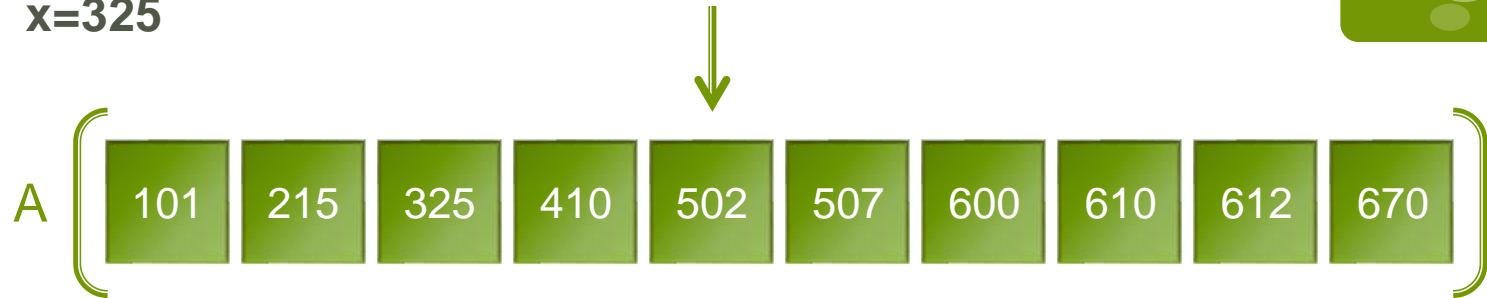
A

101	215	325	410	502	507	600	610	612	670
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Ejemplo 2



x=325





El algoritmo del método de búsqueda binaria es:

```
FUNC buscarBinaria (x: Entero) dev ENTERO
    izq  $\leftarrow$  1, der  $\leftarrow$  n, cen  $\leftarrow$  (n div 2): ENTERO
    band  $\leftarrow$  falso: BOOLEANO
    MIENTRAS izq  $\leq$  der Y band = falso HACER
        SI x = A[cen] ENTONCES
            band  $\leftarrow$  verdadero
        FIN_SI
    EN_CASO_CONTRARIO
        SI x > A[cen] ENTONCES
            izq  $\leftarrow$  cen + 1
        FIN_SI
    EN_CASO_CONTRARIO
        der  $\leftarrow$  cen - 1
    FIN_ECC
FIN_ECC
FIN_MIENTRAS
DEV (band=verdadero) ? cen : -1
FIN_FUNC
```



El análisis de complejidad del método de búsqueda binaria está dado por el número total de comparaciones.

El mejor caso de complejidad se presenta cuando el elemento buscado (x) se encuentra a la mitad del conjunto $C_{\min} = 1$.

El peor caso de complejidad se presenta cuando el algoritmo no se encuentra o se encuentra en la última comparación. En tal caso $C_{\max} = \log_2 (n)$ comparaciones.

El caso promedio está dado por el número máximo de comparaciones entre 2, es decir $C_{\text{med}} = \log_2 n / 2$.



Se podría pensar que la búsqueda binaria es más eficiente que la lineal, empero, hay que recordar que la búsqueda binaria no funciona dentro de un conjunto que no esté ordenado y, por ende, se debe ordenar antes el conjunto para que dicha búsqueda se pueda realizar.

Por lo tanto, el tiempo que se debe contabilizar para encontrar un elemento en un conjunto de datos desordenado a través de una búsqueda binaria está también en función del tiempo que se lleve ordenar el conjunto.



7.4 Búsqueda por transformación de llaves.



7.4 Búsqueda por transformación de llaves.

Las técnicas de búsqueda por transformación de llaves se basan en la idea de calcular una dirección en forma directa a partir de la llave de un nodo.

Normalmente se expresan utilizando una función de mapeo $H:K \rightarrow Y$, donde K es el dominio de la función formado por el conjunto de llaves que pueden ser transformadas y Y es el rango de la función formado por el conjunto de números enteros que representan a las direcciones de almacenamiento.



Esta técnica tiene un tiempo de búsqueda independiente del número de nodos en la estructura y su eficiencia es proporcional al tiempo utilizado para llevar a cabo la transformación.

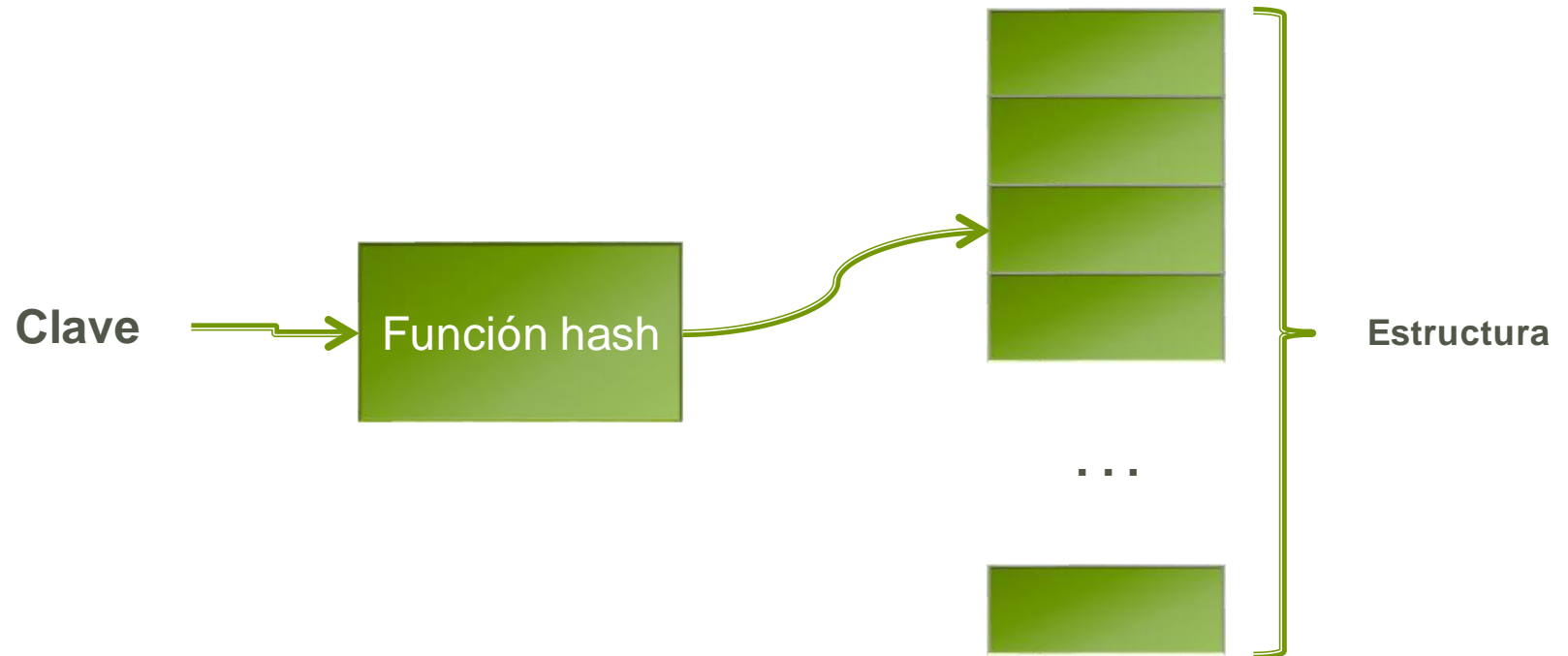
Hay que resaltar que si se quiere ocupar este tipo de técnicas se deben usar tanto en la recuperación de la información como en el almacenamiento de la misma.



7.4.1 Funciones hash

El método hash permite aumentar la velocidad de búsqueda sin necesidad de tener los elementos ordenados y, a diferencia de los métodos anteriores donde su complejidad es proporcional al número de elementos del conjunto (entre más elementos más comparaciones), el método hash es independiente del número de nodos de la estructura.

Funciones hash





El método de transformación de claves permite tener acceso directo a los elementos del conjunto utilizando una función que convierte una clave dada en una dirección (o índice) dentro del conjunto.

$$\text{dirección} \leftarrow H(\text{clave})$$

La función hash (H) aplicada a la clave genera un índice o dirección en el conjunto que permite acceder directamente al elemento.



Funciones hash por módulo (división)

La función hash por módulo consiste en tomar el residuo de la división de la clave entre el número de componentes del arreglo, es decir:

$$H(k) = (k \bmod n) + 1$$

Donde

n: es el número de elementos

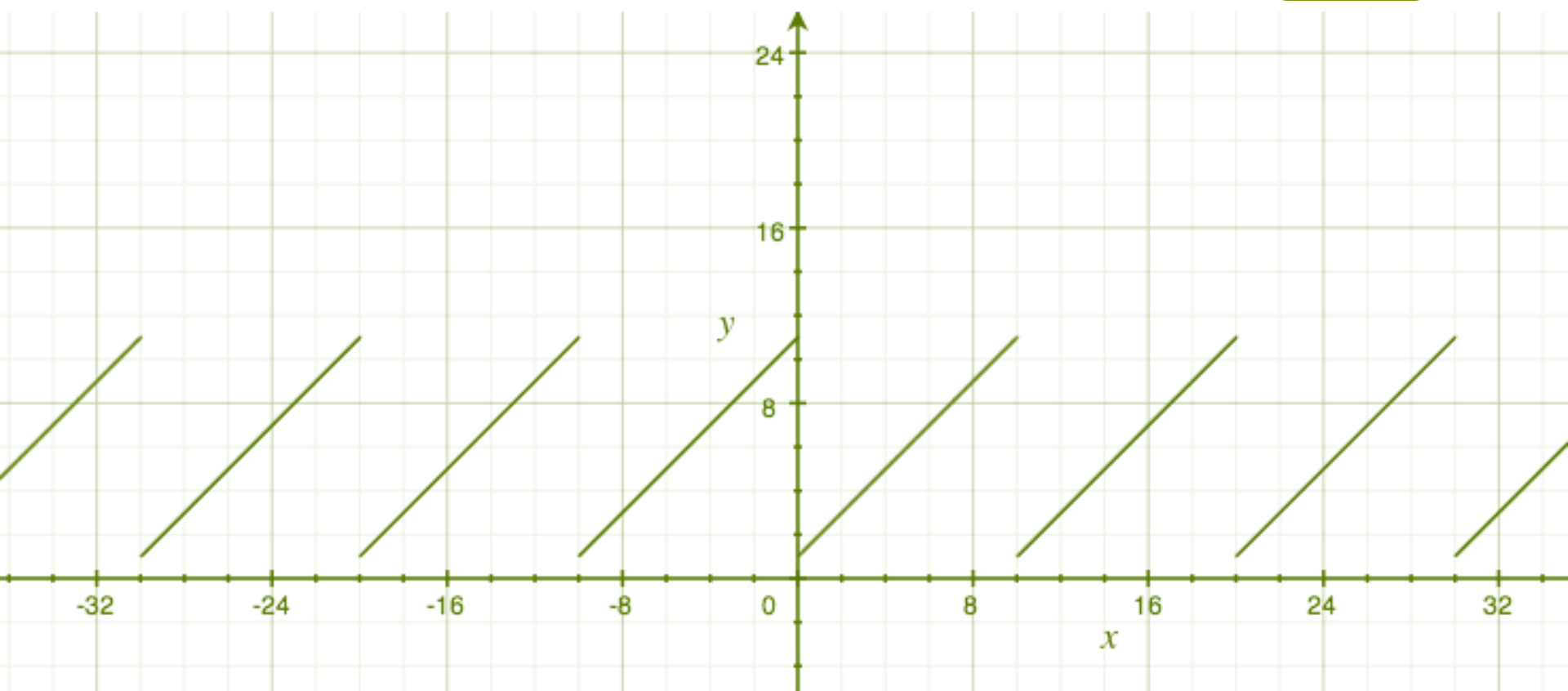
k: es la clave del dato a buscar/almacenar

El resultado del lado derecho de la ecuación genera un número en el rango $[1, n]$.



Se tiene un conjunto de 100 elementos y se deben asignar direcciones en el rango $[1, 100]$ para almacenar o recuperar un dato. Obtener la posición que se ocupa para almacenar el elemento $k = 7,259$ por medio de la función hash por módulo.

$$H(k) = (7,259 \bmod 100) + 1$$
$$H(k) = 60$$





Funciones hash cuadrado

La función hash cuadrado consiste en elevar al cuadrado la clave y tomar los dígitos centrales como dirección. El número de dígitos que se deben considerar está determinado por el rango del índice. La función hash cuadrada está definida por:

$$H(k) = \text{dígitosCentrales}(k^2) + 1$$

Donde

k: es la clave del dato a buscar/almacenar

El resultado del lado derecho de la ecuación genera un número en el rango $[1, n]$.



Se tiene un conjunto de 100 elementos y se deben asignar direcciones en el rango $[1, 100]$ para almacenar o recuperar un dato. Obtener la posición que se ocupa para almacenar los elementos $k = 7,259$ por medio de la función hash cuadrado.

$$k^2 = 52,693,081$$

$$H(k) = \text{dígitosCentrales}(52,693,081) + 1 = 94$$

Debido a que el rango de índices varía de 1 a 100, se toman solamente dos dígitos centrales del cuadrado de las claves.



Funciones hash por plegamiento

La función hash por plegamiento consiste en dividir la clave en partes, tomando igual número de dígitos y operar con ellos asignando como dirección los dígitos menos significativos. A su vez, se pueden realizar operaciones entre los diferentes dígitos.



Sea k la clave del dato a buscar. k está formada por los dígitos d_1, d_2, \dots, d_n . La función hash por plegamiento se define como:

$$H(k) = \text{dígMenosSig} ((d_1 \dots d_i) + (d_{i+1} \dots d_j) + \dots + (d_1 \dots d_n)) + 1$$

La operación entre dígitos puede ser suma (como en la función anterior) o multiplicación. Al final se genera un número en el rango $[1, n]$.



Se tiene un conjunto de 100 elementos y se deben asignar direcciones en el rango [1, 100] para almacenar o recuperar un dato. Obtener la posición que se ocupa para almacenar los elementos $k = 7,259$ por medio de la función hash plegamiento.

$$H(k) = \text{dígMenosSig}(72 + 59) + 1$$

$$H(k) = \text{dígMenosSig}(1 + 31) + 1$$

$$H(k) = 33$$

Debido a que el rango de índices varía de 1 a 100, se toman solamente dos dígitos del número (de derecha a izquierda), es decir, los menos significativos.



Funciones hash por truncamiento

La función hash por truncamiento consiste en tomar algunos dígitos de la clave y formar con ellos una dirección. Este método es de los más sencillos, pero ofrece muy poca uniformidad en la distribución de claves.

Sea k la clave de dato a buscar, k está formada por los dígitos d_1, d_2, \dots, d_n . La función hash por truncamiento está dada por:

$$H(k) = \text{elegirDígitos}(d_1, d_2, \dots, d_n) + 1$$



La elección de los dígitos es arbitraria, es decir, se pueden tomar los dígitos de las posiciones impares o de las posiciones pares. Además, se pueden unir de izquierda a derecha o viceversa.



Se tiene un conjunto de 100 elementos y se deben asignar direcciones en el rango $[1, 100]$ para almacenar o recuperar un dato. Obtener la posición que se ocupa para almacenar los elementos $k = 7,259$ por medio de la función hash cuadrado.

$$H(k) = \text{elegirDígitos}(7259) + 1$$

$$H(k) = 75 + 1$$

$$H(k) = 76$$

En este caso, se eligen los dígitos en las posiciones 1 y 3 y se unen de izquierda a derecha.



7.4.2 Colisiones

Una función hash debe ser simple de calcular y asignar direcciones de la manera más uniforme posible, es decir, generar posiciones diferentes dadas dos claves diferentes.

$$\begin{aligned}H(k_1) &= d \\ H(k_2) &= d \\ k_1 &\neq k_2\end{aligned}$$

Es decir, existe una asignación de la misma dirección a dos o más claves distintas. A este hecho se le conoce como colisión.



Se tiene un conjunto de 100 elementos y se deben asignar direcciones en el rango $[1, 100]$ para almacenar o recuperar un dato. Obtener la posición que se ocupan para almacenar los elementos $k_1 = 7,259$ y $k_2 = 9,359$.

$$H(k_1) = (7,259 \bmod 100) + 1 = 60$$

$$H(k_2) = (9,359 \bmod 100) + 1 = 60$$

Se puede observar que $H(k_1)$ y $H(k_2)$ son iguales para cuando k_1 es diferente a k_2 , por lo tanto se presenta una colisión que se debe resolver para poder almacenar el elemento.



Cuando ocurren colisiones se debe tener algún método para solventarlas.

Para el ejemplo anterior, una posible solución es aplicar la fórmula con un número primo cercano a n (100), el resultado sería:

$$H(k_1) = (7,259 \bmod 97) + 1 = 82$$

$$H(k_2) = (9,359 \bmod 97) + 1 = 48$$

Se puede observar que para $H(k_1)$ y $H(k_2)$ ya no hay colisión.



La elección de un método adecuado para resolver colisiones es tan importante como la elección de una buena función hash.

Generalmente, no importa el método elegido, resulta costoso tratar las colisiones. Por ello se debe realizar un esfuerzo importante para encontrar una función que ofrezca la mayor uniformidad posible en la distribución de las claves.



La manera más natural de resolver el problema de las colisiones es reservar una casilla por clave, pero esta solución puede tener un alto costo en memoria. Por ejemplo, si se intenta almacenar un número de cuenta, se tendrían que reservar 1 billón de registros y se ocuparían muy pocas localidades.

Existen métodos muy utilizados para corregir colisiones, los cuales se pueden clasificar en:

- Reasignación
- Arreglos anidados
- Encadenamiento



Reasignación

Esta clasificación se basa en el principio de comparación y reasignación de elementos. Algunos métodos que trabajan bajo este esquema son:

- Prueba lineal
- Prueba cuadrática
- Doble dirección hash



Reasignación: prueba lineal

El método de prueba lineal consiste en recorrer el conjunto secuencialmente a partir del punto de colisión buscando el elemento. El proceso de búsqueda concluye cuando el elemento es hallado o cuando se encuentre una posición vacía. El método trata al conjunto como una estructura circular, es decir, el siguiente elemento del último es el primero.

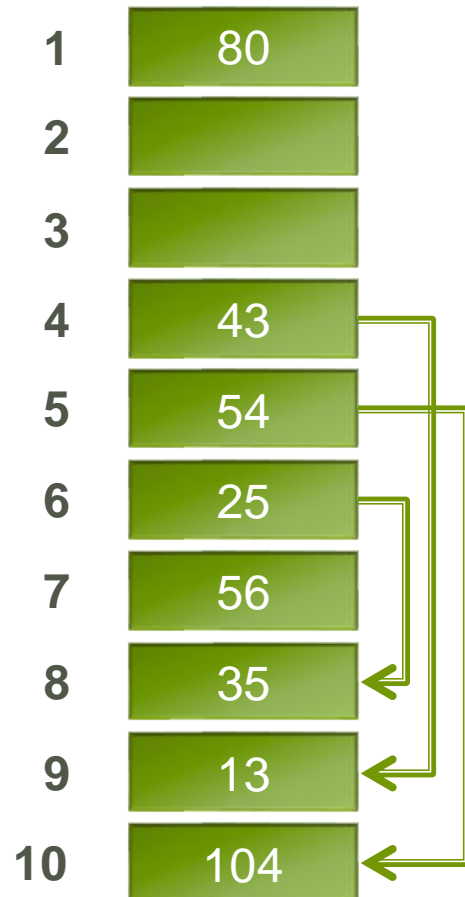
Ejemplo 8



48

Se tiene una estructura de datos lineal y se ingresan las claves 25, 43, 56, 35, 54, 13, 80 y 104, según la función hash por módulo ($H(k) = (k \bmod 10) + 1$).

Ejemplo 8



k	H(k)
25	6
43	4
56	7
35	6
54	5
13	4
80	1
104	5

Ejemplo 8



50

Si se aplica la función hash para buscar la clave 35 se obtiene una dirección igual a 6, sin embargo, en dicha dirección no se encuentra el elemento buscado, por lo que se inicia el recorrido secuencial del conjunto a partir de la siguiente posición.



Reasignación: prueba cuadrática

El método de prueba cuadrática consiste en recorrer el conjunto a partir del punto de colisión buscando el elemento con direcciones alternativas de la forma $D + i^2$. Esta variación permite una mejor distribución de las llaves que colisionan.

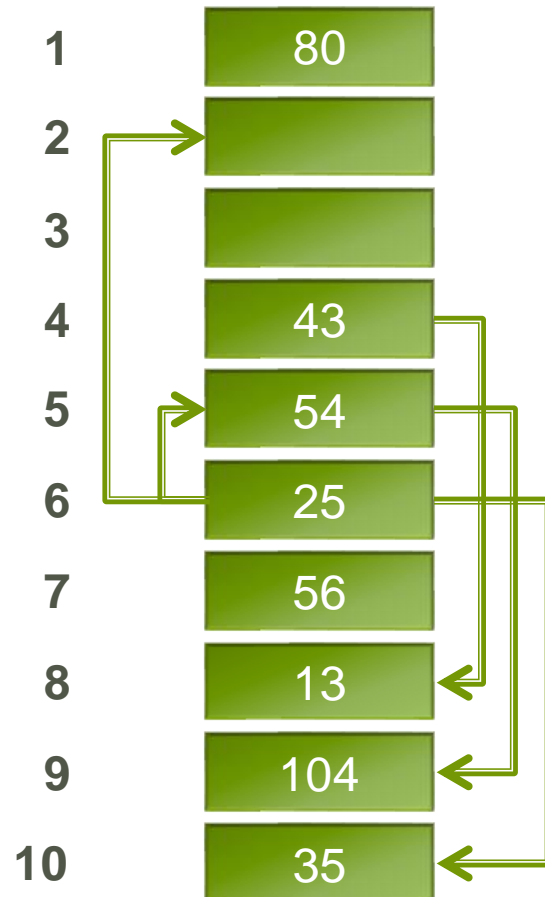
Ejemplo 9



52

Se tiene una estructura de datos lineal y se ingresan las claves 25, 43, 56, 35, 54, 13, 80, 104 y 55, según la función hash por módulo ($H(k) = (k \bmod 10) + 1$).

Ejemplo 9



k	H(k)
25	6
43	4
56	7
35	6
54	5
13	4
80	1
104	5
35	6



Si se aplica la función hash para buscar la clave 35 se obtiene una dirección igual a 6, sin embargo, en dicha dirección no se encuentra el elemento buscado, por lo que se calcula la siguiente posición.

La siguiente posición se calcula como la suma de $d + (i \cdot i)$, de donde se obtiene la dirección 7. El proceso termina cuando se encuentra el valor deseado.



Reasignación: doble dirección hash

El método de dirección hash consiste en generar otra dirección aplicando la otra función hash a la dirección obtenida en la colisión. La función hash puede ser la misma aplicada al obtener la colisión u otra diferente. El proceso termina cuando el elemento es hallado o cuando se encuentra una posición vacía.

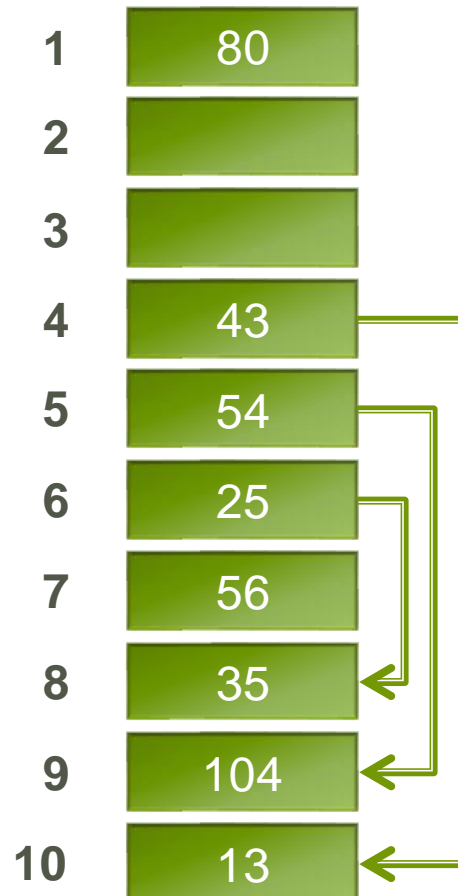


Se tiene una estructura de datos lineal y se ingresan las claves 25, 43, 56, 35, 54, 13, 80 y 104, según la función hash por módulo ($H(k) = (k \bmod 10) + 1$).

Además, se define la función H' para calcular direcciones alternativas en caso de colisión:

$$H'(d) = ((d+1) \bmod 10) + 1$$

Ejemplo 10



k	H(k)	H'(d)	H'(d')	H'(d'')
25	6	—	—	—
43	4	—	—	—
56	7	—	—	—
35	6	8	—	—
54	5	—	—	—
13	4	6	8	10
80	1	—	—	—
104	5	7	9	—



Si se aplica la función hash para buscar la clave 13 se obtiene una dirección igual a 4, sin embargo, en dicha dirección no se encuentra el elemento buscado, por lo que se aplica H' reiteradamente hasta localizar el elemento deseado.



Arreglos anidados

Este método consiste en que cada elemento del arreglo donde se almacenan las llaves posee otro arreglo en el que se almacenan los elementos que colisionan.

La desventaja de esta solución es que resulta ineficiente y/o insuficiente el espacio en memoria (dependiendo del número de elementos del arreglo).

Ejemplo 11



60

Se tiene una estructura tipo arreglo bidimensional y se ingresan las claves 25, 43, 56, 35, 54, 13, 80 y 104, según la función hash por módulo ($H(k) = (k \bmod 10) + 1$).

Ejemplo 11



1	80			
2				
3				
4	43	13		
5	54	104		
6	25	35		
7	56			
8				
9				
10				

...

k	H(k)
25	6
43	4
56	7
35	6
54	5
13	4
80	1
104	5



Encadenamiento

El método de encadenamiento consiste en que cada elemento del conjunto tiene una referencia hacia una lista ligada, la cual se va generando en tiempo real y guarda los valores que colisionan.

Las desventajas del método de encadenamiento es el uso propio de listas (si no se implementan correctamente) así como el tamaño que pueden tomar dichas listas.

Ejemplo 12



63

Se tiene una estructura tipo arreglo bidimensional y se ingresan las claves 25, 43, 56, 35, 54, 13, 80 y 104, según la función hash por módulo ($H(k) = (k \bmod 10) + 1$).

Ejemplo 12



1	80	
2		
3		
4	43	→ 13
5	54	→ 104
6	25	→ 35
7	56	
8		
9		
10		

k	H(k)
25	6
43	4
56	7
35	6
54	5
13	4
80	1
104	5



7 Métodos de búsqueda

Objetivo: Aplicar el método de búsqueda apropiado a conjuntos de datos residentes, tanto en la memoria principal, como en la memoria secundaria. Diseñar y aplicar algoritmos.

7.1 Generalidades.

7.2 Definición de la operación de búsqueda.

7.3 Búsqueda por comparación de llaves.

7.4 Búsqueda por transformación de llaves.