Universidad Nacional Mayor de San Mar os
Facultad de Ingeniería de Sistemas

Algorítmica III

Guía 4

Técnicas de dispersión - Hashing

Mg Augusto Cortez Vásquez

Objetivos

- **Conocer las técnicas de Dispersión abierta**
- **Conocer las técnicas de dispersión cerrada**
- **Conocer la técnica de resolución de colisiones**

Pensar es el diálogo del alma consigo misma

Desde su perspectiva dualista, la vida mental pertenece a un plano de la realidad distinto al de la materia.
Platon

Roger **Penrose**, OM, FRS (Reino Unido: /ˈɹɒdʒə ˈpenˌɹoz/; Colchester, 8 de agosto de 1931). Es un físico matemático oriundo de Inglaterra y profesor emérito de Matemáticas de la Universidad de Oxford. Es reconocido por su trabajo en física matemática, en particular por sus contribuciones a la teoría de la relatividad general y a la cosmología

Introducción

La ejecución de un programa de software requiere que tanto éste como los datos que necesite estén cargados en memoria principal del computador.

Esto presenta dos inconvenientes:

1. La memoria principal del computador es limitada, lo cual provoca que los programas no puedan utilizar una cantidad de datos muy grande.
2. La existencia de los datos en la memoria principal está limitada por la ejecución del programa, es decir, los datos desaparecen de la memoria principal cuando el programa concluye su ejecución, lo cual impide su reutilización por él mismo o por otros programas

· ARCHIVOS

Para superar estos inconvenientes los datos se almacenan en dispositivos de almacenamiento secundario, en recipientes denominados archivos (cintas, discos etc.), los cuales a diferencia de la memoria principal, no es volátil, sino permanente.

Ventajas de utilizar archivos

1. Independencia de los datos respecto de los programas
2. Facilidad de acceso por distintos programas en diferentes momentos
3. La información almacenada es permanente
4. Posee una gran capacidad de almacenamiento

Definición

Un ***archivo*** es una secuencia de datos que contienen información relacionada. Los datos de un archivo se encuentran estructurados en registros.

Definición

Un ***registro*** es una estructura de datos formada por uno o mas elementos denominados campos que pueden ser tratados como una unidad desde el punto de vista conceptual.

Definición

Un ***campo*** es una unidad básica de un registro. Puede ser simple(entero, real o carácter) , o compuesto (consta de subcampos que a su vez pueden ser simples o compuestos.

Definición

Una ***Clave*** es un conjunto de campos que identifica a un registro de los demás registros del archivo. La clave debe ser diferente para cada registro del archivo. Pueden existir varias claves dentro de un archivo. Una de ellas es denominada clave principal. Las demás se denominan claves secundarias.

Diseño de base de datos

El proceso de **normalización de base de datos** consiste en aplicar una serie de reglas a las relaciones obtenidas tras el paso del modelo entidad-relacion al modelo relacional.

Las bases de datos relacionales se normalizan para:

- Evitar la redundancia de los datos.
- Evitar problemas de actualización de los datos en las tablas.
- Proteger la integridad de los datos.

En el modelo relacional es frecuente llamar tabla a una relación, aunque para que una tabla sea considerada como una relación tiene que cumplir con algunas restricciones:

- Cada columna debe tener su nombre único.
- No puede haber dos filas iguales. No se permiten los duplicados.
- Todos los datos en una columna deben ser del mismo tipo.

En general, las primeras tres formas normales son suficientes para cubrir las necesidades de la mayoría de las bases de datos. El creador de estas 3 primeras formas normales (o reglas) fue Edgar F.Codd

Primera Forma Normal (1FN)]

Una tabla está en Primera Forma Normal sólo si

- Todos los atributos son atómicos. Un atributo es atómico si los elementos del dominio son indivisibles, mínimos.
- La tabla contiene una clave primaria.
- La tabla no contiene atributos nulos.
- Si no posee ciclos repetitivos.

Una columna no puede tener múltiples valores. Los datos son atómicos. (Si a cada valor de X le pertenece un valor de Y, entonces a cada valor de Y le pertenece un valor de X)

Esta forma normal elimina los valores repetidos dentro de una BD

Segunda Forma Normal (2FN)

Dependencia Funcional. Una relación está en 2FN si está en 1FN y si los atributos que no forman parte de ninguna clave dependen de forma completa de la clave principal. Es decir que no existen dependencias parciales.

En otras palabras podríamos decir que la segunda forma normal está basada en el concepto de dependencia completamente funcional. Una dependencia funcional $x \rightarrow y$es completamente funcional si al eliminar los atributos A de X significa que la dependencia no es mantenida, esto es que A ϵ X, (X – {A}) -x-> Y. Una dependencia funcional $x \rightarrow y$es una dependencia parcial si hay algunos atributos $A \in X$que pueden ser removidos de X y la dependencia todavía se mantiene, esto es A ϵ X, (X – {A}) -> Y

.

Tercera Forma Normal (3FN)

La tabla se encuentra en 3FN si es 2FN y cada atributo que no forma parte de ninguna clave, depende directamente y no transitivamente, de la clave primaria.

Un ejemplo de este concepto sería que, una dependencia funcional X->Y en un esquema de relación R es una dependencia transitiva si hay un conjunto de atributos Z que no es un subconjunto de alguna clave de R, donde se mantiene X->Z y Z->Y.

Ejemplo 1

```
Registro Alumno

                Codigo
                Nombre
                Facultad
FinReg
```

FinReg

```
Registro Direccion
                Carácter        calle
                Entero          numero
                Carácter        distrito[30]
FinReg
```

```
Registro Fecha
                Entero          dia
                Entero          mes
                Entero          anio
FinReg
```

registro con tres campos simples
Empleado

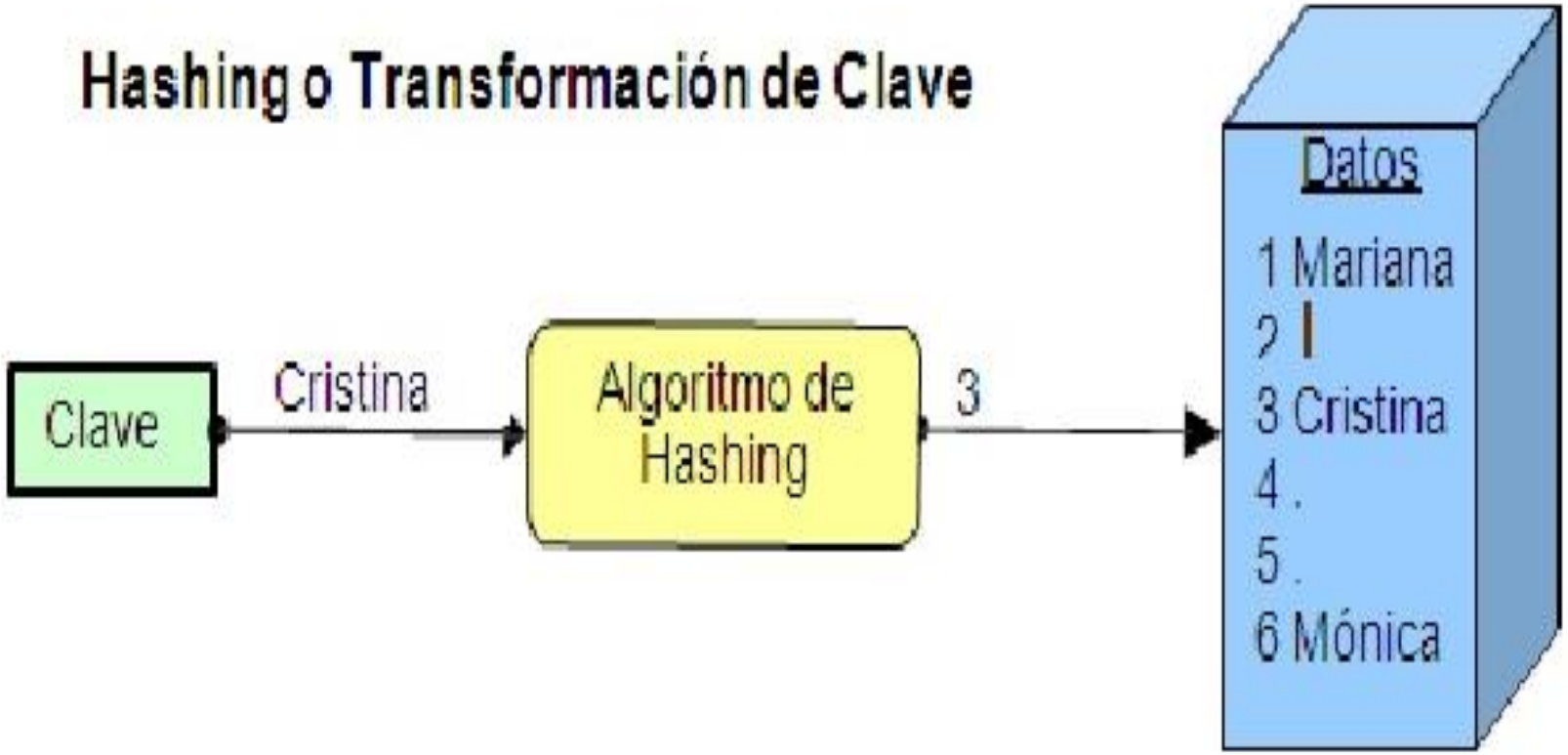
Carácter	Codigo[6]
Carácter	Nombre[39]
Fecha	Fec_Nac
Direc	Domicilio

ARCHIVOS DISPERSOS

La dispersión es una técnica empleada para realizar inserciones, eliminaciones y búsquedas en un tiempo promedio constante. Aquí no son eficientes las operaciones sobre arboles que requieren cualquier información de ordenamiento de datos. No se cuenta con operaciones como en arboles como Buscar_Max, Buscar_Min, ni la visualización de la tabla completa en orden.

Las tablas Hash o tablas de dispersión solucionan satisfactoriamente nuestro problema: nos permiten acceder asociativamente a la información y, además, lo hacen en un tiempo medio constante, es decir, que el tiempo necesario para acceder a un elemento, no va a depender del número de elementos que almacene la estructura.

Hashing o Transformación de Clave



Una estructura hash se construye con **tres elementos básicos** :

Un vector direccionable mediante número de posición (un array) capaz de almacenar N elementos.

Una función de dispersión que nos permita a partir de la clave obtener el índice donde estará el dato asociado a esa clave. Es frecuente que existan dos claves distintas para las que la función de dispersión produzca el mismo índice. Esto se denomina colisión, y las dos claves distintas que dieron lugar al mismo índice, se dicen sinónimas respecto a la función de dispersión utilizada.

Una función de resolución de colisiones

Problemas que se presentan al aplicar la técnica de dispersión son

a) Elegir el tamaño de la tabla

b) Elegir una función de dispersión apropiada

En los archivos dispersos, la dirección de cada registro se calcula aplicando una función de dispersión $\mathbf{F_d}$ sobre uno o varios de sus campos, denominados *campos de dispersión*. Los registros de un archivo disperso se encuentran distribuidos en forma aleatoria, es por eso que también se les llama archivos aleatorios o archivos de acceso directo. El acceso a los datos es muy rápido sólo si se busca con la condición de igualdad sobre el campo de dispersión.

La función de dispersión se debe escoger de modo que los registros queden distribuidos uniformemente en todo el archivo.

Ejemplo 2

Consideremos el ejemplo de una universidad con un archivo de alumnos con código de 7 dígitos. Suponga que la Universidad tiene menos de 10,000 alumnos y que solo hay un registro por alumno. Entonces un arreglo de 10,000 alumnos es suficiente para contener a los 10,000 el archivo completo. El vector esta indexado por enteros del 0 al 9999 inclusive

Los últimos 4 dígitos del código del alumno se utilizan como índice para registro de cada alumno en el vector.
Así si se quiere acceder al registro con llave 4561561 , la función de dispersión obtendrá los 4 últimos dígitos 1561 como índice en el vector

Si F_d es una función de dispersión y K es una llave, entonces $F_d(K)$ se llama la dispersión de llave y es el índice en el cual un registro con llave K debe ser colocado.

En el ejemplo $F_d(K) = K \text{ modulo } 10000$

Sin embargo se presenta el problema cuando se quiere acceder a dos llaves 5141561 y 78951561. Ambas corresponden al mismo índice 1561

0	4562610			
1	5861451			
2	1452952			
3	1579683			
1560	1461560			
1561	4561561			
1562	5591562			
9997	1859997			
9998	8179998			
9999	0029999			

ALGUNAS FUNCIONES DE DISPERSIÓN

Ejemplo 3

Consideremos La Facultad de Ingeniería de Sistemas numero de alumnos es 800 se escoge como clave para identificación su código de alumno (CODALU) de 8 dígitos. El intervalo de variación del número de este documento es de millones. Por lo que no es práctico tratar de manejar un archivo de acceso directo en el que se reserve espacio para cada código de 8 dígitos.

Para solucionar el problema es necesario, a partir del código, generar un número seudoaleatorio de 1 a 800 tal que permita encontrar la información del alumno.

I = ALEATORIO(CODALU)
Esto indica que a partir de CODALU se genera un numero seudoaleatorio I tal que $1 \leq I \leq 800$

Método de residuo

Se toma la clave y se divide por el tamaño de la tabla o del archivo, y el residuo determina la posición relativa en la tabla o el archivo.
Normalmente para disminuir el número de colisiones es conveniente usar un numero primo igual o ligeramente menor que el tamaño de la tabla.

M	Tamaño de la tabla
I =	RESIDUO(CLAVE/M) + 1 Genera valores entre 0 y M-1
I	Permite determinar la posición en la tabla

Método de cuadrados

Se eleva la clave al cuadrado, se toman los números centrales y se multiplican por el factor de conversión, con el objeto de ajustarla el tamaño de la tabla

Ejemplo 4

Dada la clave de un producto 8254

El factor de conversión es 0.4

8254 * 8254 =68128516

128*0.4=51.2

luego Aleatorio(8254) es 51

Método de desfasamiento

Se Suman los números de ambos extremos de la clave sobre los números centrales. Se suman y se multiplican por un factor de conversión. Esta técnica se puede emplear cuando la clave es bastante grande con respecto a tamaño de la tabla.

Ejemplo 5

Dada la clave de un producto 483259782
El factor de conversión es 0.4

259
483
782

524

la suma se multiplica por el factor de conversión
524* 0.4 = 209.6
luego Aleatorio(483259782) es 209

Método del doblaje(folding)

Cuando la clave es muy grande, se divide en tres partes; los extremos se giran sobre la ter
parte central, se suman y se multiplican por el factor de conversión

Ejemplo 6

Dada la clave de un producto 170863519
El factor de conversión es 0.7

863
915
071

849

La suma se multiplica por el factor de conversión
849* 0.9 = 764.1

luego Aleatorio(17086335519) es 764

Método de cambio de base de los números

Es posible utilizar el cambio de base de los números para generar un numero seudoaleatorio. Uno de los cambios de base mas empleados debido a la facilidad de conversión es expresar la clave en base 11.

Ejemplo 7

Dada la clave de un producto 59582

$5 \times 11^{4} + 9 \times 11^{3} + 5 \times 11^{2} + 8 \times 11 + 2$

Este número se reduce mediante cualquiera de los métodos anteriores

Método de división de polinomios

A partir de la clave numérica se genera un polinomio P(x), este polinomio se divide por otro polinomio Q(x) previamente establecido, el cual genera un polinomio R(x) ; los coeficientes de R(x) se multiplican por el factor de conversión para generar la posición

Ejemplo 8

Dada la clave de un producto 359849 se quiere generar una posición en un archivo de 700 registros

$P(x) = 3x^5 + 5x^4 + 9x^3 + 8x^2 + 4x + 9$

Se divide entre un polinomio Q(x) que siempre va a ser el mismo. Si consideramos $Q(x) = 3x^3 + 2x^2 + x + 2$, entonces $R(x) = x^2 + 5x$

Con base al factor de conversión 0.7 la posición de almacenamiento es 73

Colisión

En este caso hay que resolver las colisiones
La resolución dependerá de la técnica de dispersión

Dispersión cerrada:

Resolución por examen lineal
Resolución por examen cuadrático
Resolución por redispersion

Dispersión abierta

Resolución mediante área de desborde

Hay varias técnicas para gestionar las colisiones:

- **Direccionamiento abierto**.
- **Encadenamiento.**
- **Dispersión múltiple**..

- **Direccionamiento abierto**. Cuando se produce una colisión, el sistema hace una búsqueda lineal a partir del bloque al que iba destinado el registro para encontrar un agujero donde insertarlo. Si se llega al final del archivo sin encontrar un agujero, se continúa la búsqueda desde el principio.

:

- **Encadenamiento.** En lugar de buscar un agujero libre, lo que se hace es disponer de una serie de bloques como área de desborde. Cuando se produce una colisión, el registro se sitúa en el área de desborde y mediante un puntero en el bloque colisionado, se apunta a la dirección del bloque de desborde y la posición relativa del registro dentro del bloque. Además, todos los registros que han colisionado en un mismo bloque se van encadenando mediante punteros.

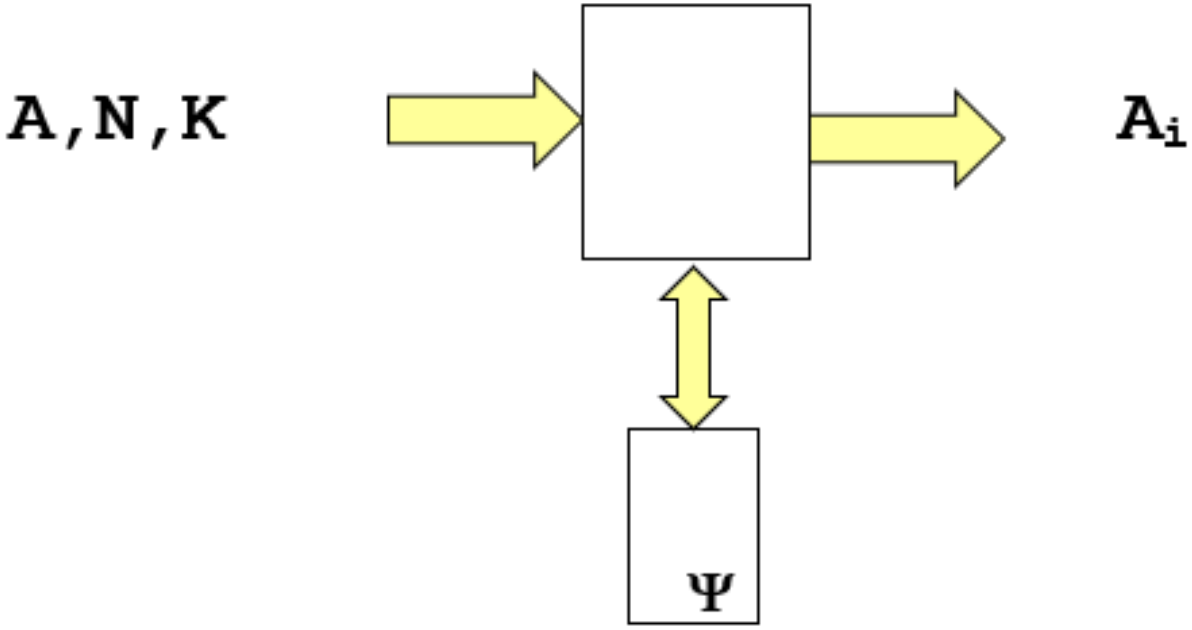
- **Dispersión múltiple**. Esta técnica de resolución de colisiones consiste en utilizar una segunda función de dispersión cuando la primera ha producido una colisión. El objetivo es producir una nueva dirección que no provoque colisión. Normalmente, la segunda función da una dirección de bloque situada en un área de desborde.

DISPERSION CERRADA (ENCADENAMIENTO EN LA MISMA AREA DE DIRECCIONAMIENTO)

Cuando ocurre una colisión se busca una celda en la misma área de direccionamiento (vector) y para ello se utiliza cualquiera de las técnicas de examen: lineal, cuadrático o redispersion.

Entrada : A: vector de dispersión de orden N
K: clave a insertar

Salida : A_i : vector de dispersión de orden N



```
Accion InsertarClaveLineal( A[ ] : TABLA_DISP, K:clave, N : entero)
Inicio
        i = Ψ (K )
        Si (A[i] = K )
                Escribir "la clave esta en la posicion" , i
        Sino
                Si (A[i] esta vacio )
                        A[i] = K
                Sino
                        di= i+1
                        Mientras (A[di ] ≠ K  ∧ A[di] ≠ ∅ ∧ di ≠ i  )
                                di = di +1
                                Si (di = N+1 )
                                        di = 1
                                FinSi
                        FinMientras
                        Si (A[di] = K )
                                Escribir "la clave esta en la posicion" , di
                        Sino
                                Si (di = i )
                                        escribir "el arreglo no tiene casillas vacias"
                                Sino
                                        A[di] = K
                                FinSi
                        FinSi
                Finsi
        FinSi
Fin
```

Desventajas de la dispersión

Aunque la dispersión es el método de acceso directo más rápido a través del campo de dispersión, no es muy útil cuando también se quiere acceder al archivo a través de otro campo. Ya que la mayoría de las funciones de dispersión que se utilizan no mantienen el orden entre los registros, tampoco es útil cuando se quiere leer los registros ordenadamente.

Ejercicios propuestos

1. Cuáles son los elementos básicos para construir una estructura HASH
2. Cuáles son las técnicas para resolver colisiones
3. Cuáles son las desventajas de utilizar dispersión
4. Defina primera forma normal, segunda forma normal y tercera forma normal
5. Proponga un ejemplo en el que ilustre la primera forma normal, la segunda tercera forma normal
6. Cuáles son las ventajas y desventajas de utilizar archivos
7. Que funciones de dispersión conoce. Proporciones un ejemplo de cada una de ellas

Ejercicios propuestos

8 **Implementar las operaciones básicas para dispersión abierta**

9 **Implementar las operaciones básicas para dispersión cerrada**