Estructuras de datos: Tablas de dispersión

Algoritmos

Facultad de Informática Universidad de A Coruña



Tablas de dispersión

- Objetivo: realizar inserciones, eliminaciones y búsquedas en tiempo promedio constante.
- Estructura de datos ideal: un vector con tamaño fijo N.
 - Una función de dispersión establece la correspondencia de cada clave con algún número en el intervalo [0...N-1].
 - Esta función tiene que ser fácil de calcular, y asegurar que dos claves distintas se correspondan con celdas diferentes.
 - Como esto último es imposible, buscamos una función que distribuya homogéneamente las claves entre las celdas.
 - Resta escoger una función y decidir el tamaño de la tabla y qué hacer cuando dos claves caen en la misma celda.
 - Si al insertar un elemento, éste se dispersa en el mismo valor que un elemento ya insertado tenemos una colisión y hay que resolverla.
- Las tablas de dispersión se usan para representar diccionarios en los que se busca una clave y se devuelve su definición.



Índice

- Funciones de dispersión
- Dispersión abierta
- Oispersión cerrada
 - Exploración lineal
 - Exploración cuadrática
 - Exploración doble

Funciones de dispersión (i)

- Toda función de dispersión debe:
 - calcularse de forma sencilla, y
 - · distribuir uniformemente las claves.
- Por ejemplo, si las claves son números enteros, clave mod N es una función buena, salvo que haya propiedades indeseables:
 - Si N fuese 100 y todas las claves terminasen en cero, esta función de dispersión sería una mala opción.
 - Es buena idea asegurarse de que el tamaño de la tabla sea un número primo.
 - Si las claves fuesen enteros aleatorios, esta función sería muy simple y distribuiría las claves con uniformidad.
- Por lo regular, las claves son cadenas de caracteres.
 - La longitud y las propiedades de las claves influirán en la elección de una buena función de dispersión.



Funciones de dispersión (ii)

Una opción es sumar los valores ASCII de los caracteres.

```
función Dispersión1 (Clave, TamañoClave): Índice
  valor := ascii(Clave[1]);
  para i := 2 hasta TamañoClave hacer
    valor := valor + acii(Clave[i])
  fin para
  devolver valor mod N
fin función
```

- Es una función fácil de implementar, y se ejecuta con rapidez.
- Pero si el tamaño de la tabla es grande, esta función no distribuye bien las claves.
 - Por ejemplo, si N = 10007 y las claves tienen 8 caracteres, la función sólo toma valores entre 0 y 1016 = 127 · 8.



Funciones de dispersión (iii)

fin función

- En esta función de dispersión se supone que la clave tiene al menos tres caracteres.
- Si los primeros caracteres son aleatorios y el tamaño de la tabla es 10007, esperaríamos una distribución bastante homogénea.
- Desafortunadamente, los lenguajes naturales no son aleatorios.
 - Aunque hay 27³ = 17576 combinaciones posibles, en un diccionario el número de combinaciones diferentes que nos encontramos es menor que 3000.
 - Sólo un porcentaje bajo de la tabla puede ser aprovechada por la dispersión.

Funciones de dispersión (iv)

 En la función de dispersión que sigue intervienen todos los caracteres en la clave y se puede esperar una buena distribución.

```
función Dispersión3 (Clave, TamañoClave): Índice
  valor := ascii(Clave[1]);
  para i := 2 hasta TamañoClave hacer
    valor := (32*valor + acii(Clave[i])) mod N
  fin para
  devolver valor
fin función
```

- El código calcula una función polinómica con base en la regla: $\sum_{i=0}^{LongitudClave-1} 32^i \cdot ascii(clave[LongitudClave-i])$
 - Para "abcd" p. ej. $(32^3 \cdot ascii(a) + 32^2 \cdot ascii(b) + 32 \cdot ascii(c) + ascii(d)) \bmod N$

Funciones de dispersión (v)

- La multiplicación por 32 es el desplazamiento de cinco bits.
- Con lenguajes que permitan el desbordamiento se aplicaría mod una sola vez justo antes de volver.
- Si las claves son muy grandes, no se usan todos los caracteres.

Índice

- Funciones de dispersión
- 2 Dispersión abierta
- O Dispersión cerrada
 - Exploración lineal
 - Exploración cuadrática
 - Exploración doble

Resolución de colisiones: dispersión abierta (i)

- La solución consiste en tener una lista de todos los elementos que se dispersan en un mismo valor.
- Al buscar, usamos la función de dispersión para determinar qué lista recorrer.
- Al insertar, recorremos la lista adecuada.
 - Si el elemento resulta ser nuevo, se inserta al frente o al final de la lista.
- Además de listas, se podría usar cualquier otra estructura para resolver las colisiones, como un árbol binario de búsqueda.



Resolución de colisiones: dispersión abierta (ii)

 El factor de carga, λ, de una tabla de dispersión es la relación entre el número de elementos en la tabla y su tamaño.

$$\lambda = \frac{\textit{N\'umero de claves en la tabla}}{\textit{N}}$$

- La longitud media de una lista es λ.
 - En dispersión abierta, la regla es igualar el tamaño de la tabla al número de elementos esperados, (λ = 1).
- El esfuerzo al realizar una búsqueda es:
 - el tiempo constante necesario para evaluar la función de dispersión, O(1), más
 - 2 el tiempo necesario para recorrer la lista, $O(\lambda)$.
 - En una búsqueda infructuosa el promedio de nodos recorridos es O(λ)
 - $\bullet~$ En una búsqueda con éxito, $O(\lambda/2)$



Ejemplo de dispersión abierta (i)

Valores de la función de dispersión:

Tabla después de insertar Ana:

```
0 1 2 3 4 5 6 7 8 9 10
[] [] [] [] [] [] [] [Ana] [] [] []
```

Ejemplo de dispersión abierta (i)

Valores de la función de dispersión:

hash(José,11)=7 hash(Iván,11)=6

Tabla después de insertar Ana:

```
0 1 2 3 4 5 6 7 8 9 10
[] [] [] [] [] [] [] [Ana] [] [] []
```

Tabla después de insertar Luis:

```
0 1 2 3 4 5 6 7 8 9 10
[] [] [] [] [] [] [Luis][Ana][] [] []
```

Ejemplo de dispersión abierta (i)

Valores de la función de dispersión:

Tabla después de insertar Ana:

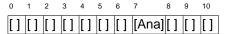


Tabla después de insertar Luis:

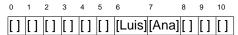


Tabla después de insertar José:

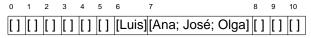
```
0 1 2 3 4 5 6 7 8 9 10

[] [] [] [] [] [] [Luis][Ana; José] [] [] []
```



Ejemplo de dispersión abierta (ii)

Tabla después de insertar Olga:



Ejemplo de dispersión abierta (ii)

Tabla después de insertar Olga:

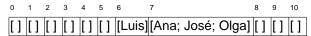
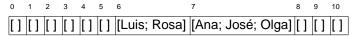


Tabla después de insertar Rosa:



Ejemplo de dispersión abierta (ii)

Tabla después de insertar Olga:

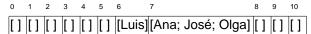


Tabla después de insertar Rosa:

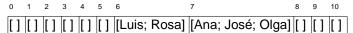


Tabla después de insertar Iván:

```
0 1 2 3 4 5 6 7 8 9 10 

[] [] [] [] [] [] [Luis; Rosa; Iván] [Ana; José; Olga] [] [] []
```

Tablas de dispersión abiertas: pseudocódigo (i)

```
tipo
  Índice = 0..N-1
 Posición = Nodo
 Lista = Posición
 Nodo = registro
   Elemento: TipoElemento
   Siquiente : Posición
 fin registro
  TablaDispersión = vector [Índice] de Lista
procedimiento InicializarTabla (T)
 para i := 0 hasta N-1 hacer
   CrearLista(T[i])
 fin para
fin procedimiento
```

Tablas de dispersión abiertas: pseudocódigo (i)

```
función Buscar (Elem, Tabla): Posición
  i := Dispersión (Elem);
 devolver BuscarLista(Elem, Tabla[i])
fin función
procedimiento Insertar (Elem, Tabla)
  pos := Buscar(Elem, Tabla); {No inserta repetidos}
  si pos = nil entonces
    i := Dispersión (Elem);
    InsertarLista(Elem, Tabla[i])
fin procedimiento
```

Índice

- Funciones de dispersión
- Dispersión abierta
- 3 Dispersión cerrada
 - Exploración lineal
 - Exploración cuadrática
 - Exploración doble

Dispersión cerrada

- En un sistema de dispersión cerrada, si ocurre una colisión, se buscan celdas alternativas hasta encontrar una vacía.
 - Se busca en sucesión en las celdas: d₀(x), d₁(x), d₂(x)... donde:

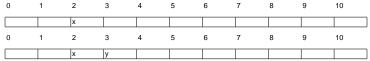
$$d_i(x) = (dispersion(x) + f(i)) \mod N, con f(0) = 0$$

- La función *f* es la estrategia de resolución de las colisiones.
 - Determinará si la dispersión cerrada es lineal, cuadrática o doble.
- Como todos los datos se guardan en la tabla, ésta tiene que ser más grande para la dispersión cerrada que para la abierta.
- En general, para la dispersión cerrada el factor de carga λ debe estar por debajo de 0,5

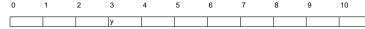


Eliminación perezosa en la dispersión cerrada

- La eliminación estándar no es realizable con dispersión cerrada.
 - La celda ocupada pudo haber causado una colisión en el pasado.
- Por ejemplo, con exploración lineal, si las claves "x" e "y" se dispersan a la misma posición (p. ej. 2)
 - Insertamos ambas claves.



Borramos la primera.



- 3 Si buscásemos ahora "y", no la encontraríamos.
- Las tablas de dispersión cerrada requieren eliminación perezosa, aunque no haya realmente "pereza".



Pseudocódigo (i)

```
tipo
 ClaseDeEntrada = (legítima, vacía, eliminada)
  Índice = 0..N-1
 Posición = Índice
 Entrada = registro
   Elemento: TipoElemento
   Información : ClaseDeEntrada
 fin registro
 TablaDispersión = vector [Índice] de Entrada
procedimiento InicializarTabla (D)
 para i := 0 hasta N-1 hacer
   D[i].Información := vacía
  fin para
fin procedimiento
```

Pseudocódigo (ii)

```
función Buscar (Elem, D): Posición
  i := 0;
  x = Dispersión(Elem);
 PosActual = x;
 mientras D[PosActual]. Elemento <> Elem y
               D[PosActual].Información <> vacía hacer
    i := i + 1;
    PosActual := (x + FunResoluciónColisión(x, i)) mod N
  fin mientras;
  devolver PosActual
fin función
   {La búsqueda finaliza
    al encontrar el elemento (legítimo o borrado),
    o al caer en una celda vacía) }
```

Pseudocódigo (iii)

```
procedimiento Insertar (Elem, D)
 pos = Buscar(Elem, D);
 si D[pos].Información <> legítima
                   {Bueno para insertar}
 entonces
   D[pos].Elemento := Elem;
   D[pos].Información := legítima
fin procedimiento
procedimiento Eliminar (Elem, D)
 pos = Buscar(Elem, D);
 si D[pos].Información = legítima
 entonces
   D[pos].Información := eliminada
fin procedimiento
```

Dispersión cerrada con exploración lineal

- Aquí la estrategia de resolución de las colisiones es una función lineal de i, por lo general f(i) = i.
 - Esto equivale a buscar secuencialmente en el vector (con circularidad) una posición vacía.
 - Si la tabla es suficientemente grande, siempre se encontrará una celda vacía.
 - Pero ello puede tomar demasiado tiempo.
- Suponiendo independencia entre intentos, el número medio de celdas examinadas en una inserción es ¹/_{1-λ}
 - En una tabla con factor de carga λ, la probabilidad de que una celda esté vacía es 1 – λ.
 - Por tanto, el número medio de intentos independientes realizados hasta encontrar una celda vacía es ¹/_{1-λ}.
 - P. ej., con $\lambda = 0.5, \frac{1}{1-\lambda} = 2$



Agrupamiento primario

- La supuesta independencia entre intentos, no se cumple.
- Aunque la tabla esté relativamente vacía, se pueden formar bloques de celdas ocupadas (agrupamiento primario).
 - Cualquier clave dispersada en el agrupamiento necesitará varios intentos para resolver la colisión, y después se agregará al agrupamiento.
- Teniendo esto en cuenta, el número medio de celdas examinadas en una inserción con exploración lineal es cercano a $\frac{1}{2}(1+\frac{1}{(1-\lambda)^2})$

λ	No medio de celdas examinadas
0,5	2,5
0,75	8,5
0,9	50,5

- El coste de una búsqueda:
 - sin éxito: es el mismo que el de una inserción;
 - con éxito: es la media de los costes de las inserciones en tablas con factores de carga más pequeños, $\frac{1}{2}(1+\frac{1}{1-\lambda})$.

Ejemplo de dispersión cerrada con exploración lineal(i)

Valores de la función de dispersión:

Tabla de dispersión después de insertar Ana:

0	1	2	3	4	5	6	7	8	9	10
							Ana			

Ejemplo de dispersión cerrada con exploración lineal(i)

Valores de la función de dispersión:

Tabla de dispersión después de insertar Ana:

0	1	2	3	4	5	6	7	8	9	10
							Ana			

Tabla de dispersión después de insertar Luis:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana			

Ejemplo de dispersión cerrada con exploración lineal(i)

Valores de la función de dispersión:

Tabla de dispersión después de insertar Ana:



Tabla de dispersión después de insertar Luis:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana			

Tabla de dispersión después de insertar José:

```
0 1 2 3 4 5 6 7 8 9 10 Luis Ana José
```



Ejemplo de dispersión cerrada con exploración lineal (ii)

Valores de la función de dispersión: hash(Ana,11) = 7 hash(Luis,11)=6 hash(José,11)=7 hash(Olga,11)=7 hash(Rosa,11)=6 hash(Iván,11)=6 hash(Iván,11)=6

Tabla de dispersión después de insertar Olga:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana	José	Olga	

Se está formando un agrupamiento primario

hash(Ana.11) = 7

hash(Luis.11)=6

Ejemplo de dispersión cerrada con exploración lineal (ii)

Valores de la función de dispersión: hash(Olga,11)=7 hash(Rosa,11)=6 hash(Nasa,11)=6 hash(Nasa

Tabla de dispersión después de insertar Rosa:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana	José	Olga	Rosa

hash(José.11)=7

hash(Ana.11) = 7

hash(Olga,11)=7

Valores de la función de dispersión:

hash(Luis.11)=6

hash(Rosa,11)=6

Ejemplo de dispersión cerrada con exploración lineal (ii)

•	Tabla	de dis	spersi	ón de	spués	de in	serta	r Olga	:		
	0	1	2	3	4	5	6	7	8	9	10
							Luis	Ana	José	Olga	
	Se está fo	rmando ur	n agrupami	iento prima	ario						
•	Tabla	de di	spersi	ón de	spués	de in	serta	r Rosa	a:		
	0	1	2	3	4	5	6	7	8	9	10
							Luis	Ana	José	Olga	Rosa
•	Tabla	de di	spersi	ón de	spués	de in	serta	r Iván:			
	0	1	2	3	4	5	6	7	8	9	10
	Iván						Luis	Ana	José	Olga	Rosa

hash(José,11)=7

hash(lván.11)=6

Dispersión cerrada con exploración cuadrática

- La función de resolución de colisiones es cuadrática, por lo general: $f(i) = i^2$.
- Con exploración lineal es malo llenar la tabla, porque se degrada el rendimiento. Para la exploración cuadrática, la situación es más drástica:
 - Con más de la mitad de la tabla ocupada, no hay garantías de encontrar una celda vacía (antes si el tamaño no es primo).
- Se demuestra que con la mitad de la tabla vacía, siendo su tamaño primo, siempre encontraremos una celda vacía al insertar.

Exploración cuadrática y agrupamientos

- La exploración cuadrática elimina el problema del agrupamiento primario que padece la exploración lineal.
 - Pero los elementos dispersados a la misma posición probarán en las mismas celdas alternas (agrupamiento secundario).
 - Se trata de un defecto teórico.
 - Los resultados prácticos indican que se producen, en general, menos de medio intento adicional por búsqueda.

Ejemplo de dispersión cerrada con exploración cuadrática (i)

Valores de la función de dispersión:

$$\begin{array}{lll} hash(Ana,11) = 7 & hash(Luis,11) = 6 & hash(José,11) = 7 \\ hash(Olga,11) = 7 & hash(Rosa,11) = 6 & hash(Iván,11) = 6 \end{array}$$

Tabla de dispersión después de insertar Ana:

0	1	2	3	4	5	6	7	8	9	10
							Ana			

Ejemplo de dispersión cerrada con exploración cuadrática (i)

Valores de la función de dispersión:

Tabla de dispersión después de insertar Ana:

0	1	2	3	4	5	6	7	8	9	10
							Ana			

Tabla de dispersión después de insertar Luis:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana			

Ejemplo de dispersión cerrada con exploración cuadrática (i)

Valores de la función de dispersión:

Tabla de dispersión después de insertar Ana:

0	1	2	3	4	5	6	7	8	9	10
							Ana	a		

Tabla de dispersión después de insertar Luis:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana			

Tabla de dispersión después de insertar José:

```
0 1 2 3 4 5 6 7 8 9 10 Luis Ana José
```



Ejemplo de dispersión cerrada con exploración cuadrática(ii)

Valores de la función de dispersión: hash(Ana,11) =7 hash(Olga,11)=7

hash(Luis,11)=6 hash(Rosa,11)=6 hash(José,11)=7 hash(Iván,11)=6

Tabla de dispersión después de insertar Olga:

0 1 2 3 4 5 6 7 8 9 10

Olga | Luis Ana José |

hash(Ana.11) = 7

hash(Olga,11)=7

Valores de la función de dispersión:

0

Olga

hash(José.11)=7

hash(Iván,11)=6

9

10

Rosa

8

José

Ana

Ejemplo de dispersión cerrada con exploración cuadrática(ii)

hash(Luis.11)=6

hash(Rosa,11)=6

Luis

Tabla de dispersión después de insertar Olga:

Olga | Luis | Ana | José |

Tabla de dispersión después de insertar Rosa:

5



Ejemplo de dispersión cerrada con exploración cuadrática(ii)

hash(Ana.11) = 7hash(Luis.11)=6 hash(José.11)=7 Valores de la función de dispersión: hash(Olga,11)=7 hash(Rosa,11)=6 hash(lván.11)=6 Tabla de dispersión después de insertar Olga: 0 7 8 9 10 Olga Luis Ana José Tabla de dispersión después de insertar Rosa: 0 5 8 9 10 Olga Ana Luis José Rosa

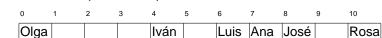


Tabla de dispersión después de insertar *lván*:

Dispersión cerrada con exploración doble

- Para la resolución de colisiones aplicamos una segunda función de dispersión, en general: $f(i) = i \cdot h_2(x)$.
 - La función nunca debe evaluarse a cero.
 - Y es importante que todas las celdas puedan ser intentadas.
 - Una función como $h_2(x) = R (x \mod R)$, con R un número primo menor que el tamaño de la tabla, funcionará bien.
 - Hay que asegurarse de que el tamaño de la tabla sea primo.

Ejemplo de dispersión cerrada con exploración doble (i)

Valores de la función de dispersión:

$$h_1(x,11)$$

 $h_2(x,11) = 5 - h_1(x,11) \%5$

	Ana	Luis	José	Olga	Rosa	Iván
	7	6	7	7	6	6
5	3	4	3	3	4	4

Tabla de dispersión después de insertar Ana:

0	1	2	3	4	5	6	7	8	9	10
							Ana			

Ejemplo de dispersión cerrada con exploración doble (i)

Valores de la función de dispersión:

$$h_1(x,11)$$

 $h_2(x,11) = 5 - h_1(x,11) \%5$

	Ana	Luis	José	Olga	Rosa	Iván
	7	6	7	7	6	6
5	3	4	3	3	4	4

Tabla de dispersión después de insertar Ana:

0	1	2	3	4	5	6	7	8	9	10
							Ana	a		

Tabla de dispersión después de insertar Luis:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana			

Ejemplo de dispersión cerrada con exploración doble (i)

Valores de la función de dispersión:

$$h_1(x,11)$$

 $h_2(x,11) = 5 - h_1(x,11) \%5$

	Ana	Luis	José	Olga	Rosa	Iván
	7	6	7	7	6	6
5	3	4	3	3	4	4

Tabla de dispersión después de insertar Ana:

0	1	2	3	4	5	6	7	8	9	10
							Ana			

Tabla de dispersión después de insertar Luis:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana			

Tabla de dispersión después de insertar José:

0	1	2	3	4	5	6	7	8	9	10
						Luis	Ana			José



Ejemplo de dispersión cerrada con exploración doble (ii)

Valores de la función de dispersión:

	Ana	Luis	José	Olga	Rosa	Iván
$h_1(x,11)$	7	6	7	7	6	6
$h_2(x,11) = 5 - h_1(x,11) \%5$	3	4	3	3	4	4

Tabla de dispersión después de insertar Olga:

0	1	2	3	4	5	6	7	8	9	10
		Olga				Luis	Ana			José

Ejemplo de dispersión cerrada con exploración doble (ii)

Valores de la función de dispersión:

	Ana	Luis	José	Olga	Rosa	Iván
$h_1(x,11)$	7	6	7	7	6	6
$h_2(x,11) = 5 - h_1(x,11)\%5$	3	4	3	3	4	4

• Tabla de dispersión después de insertar Olga:

0	1	2	3	4	5	6	7	8	9	10
		Olga				Luis	Ana			José

Tabla de dispersión después de insertar Rosa:

C	1	1	2	3	4	5	6	7	8	9	10
ſ			Olga	Rosa			Luis	Ana			José

Ejemplo de dispersión cerrada con exploración doble (ii)

Valores de la función de dispersión:

	Ana	Luis	José	Olga	Rosa	Iván
$h_1(x,11)$	7	6	7	7	6	6
$h_2(x,11) = 5 - h_1(x,11)\%5$	3	4	3	3	4	4

Tabla de dispersión después de insertar Olga:

0	1	2	3	4	5	6	7	8	9	10
		Olga				Luis	Ana			José

Tabla de dispersión después de insertar Rosa:

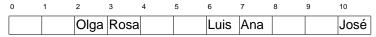


Tabla de dispersión después de insertar Iván:

