

CÁC KHÁI NIỆM

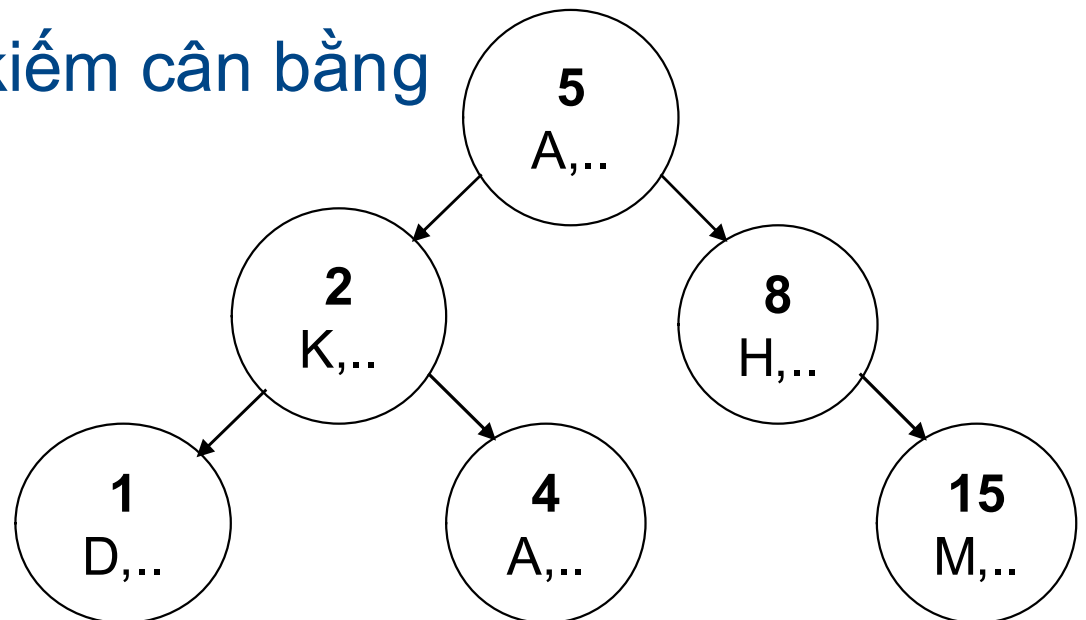
❖ BẢNG BĂM (Hashtable)

Vấn đề: cho tập dữ liệu gồm n nhân viên, mỗi nhân viên gồm thông tin mã số nhân viên, họ tên, mức lương, việc tìm kiếm được thực hiện trên trường mã số (trường khóa). Xét các cách lưu trữ sau:

1) Dùng cây nhị phân tìm kiếm cân bằng

Thời gian truy xuất:

$O(\log(n))$



CÁC KHÁI NIỆM

❖ BẢNG BĂM (Hashtable)

2) Dùng mảng sao cho chỉ số của mảng trùng với mã số nhân viên

1	2	3	4	5	6, 7	8	9,...,14	15
1, D,..	2, K,..		4, A,..	5, A,..	...	8, H,..	...	15, M,..

thời gian truy xuất $O(1)$

Nhược điểm:

- Miền giá trị của mã số lớn thì không thể tạo mảng
- Lãng phí bộ nhớ
- Chỉ áp dụng cho dữ liệu số nguyên

CÁC KHÁI NIỆM

❖ BẢNG BĂM (Hashtable)

Khái niệm bảng băm: là một cấu trúc dữ liệu tương tự mảng có kích thước m , cho phép ánh xạ một giá trị khóa thành một địa chỉ trong đoạn $[0, m-1]$ nhằm xác định nhanh chóng phần tử có khóa cần xử lý.

Ví dụ 1: để lưu trữ thông tin nhân viên trên, dùng bảng băm với hàm băm là $f(x) = x \% 9$

1	2	3	4	5	6	7	8
1, D,..	2, K,..		4, A,..	5, A,..	15, M,..		8, H,..

CÁC KHÁI NIỆM

❖ HÀM BẮM (Hash function)

Ánh xạ biến giá trị khóa thành địa chỉ trong bảng băm

Tính chất của hàm băm:

- Phải trả về giá trị trong đoạn $[0, m-1]$
- Tính toán đơn giản với độ phức tạp $O(1)$
- Không lãng phí không gian. Với mỗi địa chỉ trong bảng băm, phải có ít nhất 1 khóa có giá trị hàm băm bằng nó.
- Tối thiểu hóa đụng độ. Các khóa khác nhau sẽ có giá trị hàm băm ít trùng nhau.

CÁC KHÁI NIỆM

❖ ĐỤNG ĐỘ (Collision)

Giả sử có hàm băm $f(k)$, đụng độ là hiện tượng có hai khóa k_1 và k_2 không trùng nhau nhưng $f(k_1) = f(k_2)$

Ví dụ 2: cho bảng băm kích thước 6, các giá trị khóa là 2, 4, 6, 8, 10. Xét hàm băm $f(k) = k \% 6$:

- Có $f(2)=2$, $f(4)=4$, $f(6)=0$, $f(8)=2$, $f(10)=4$, $f(12)=0$: hàm $f(k)$ không xác định được các vị trí 1, 3, 5 \rightarrow chưa đảm bảo tính chất của hàm băm
- Có $f(2)=f(8)$, $f(4)=f(10)$, $f(6)=f(12)$: hàm $f(k)$ chưa tối thiểu hóa được các đụng độ.

CÁC KHÁI NIỆM

❖ CÁC DẠNG HÀM BẮM

- Hàm băm dạng bảng: các khóa được liệt kê tương ứng với địa chỉ.

Ví dụ 3: có $f(k)$ được cho như bảng sau:

khóa	for	long	to	do	fix	int	byte	in	if	while
địa chỉ	0	1	2	3	4	5	6	7	8	9

có: $f(\text{"if"}) = 8, f(\text{"fix"}) = 4$.

CÁC KHÁI NIỆM

❖ CÁC DẠNG HÀM BĂM

- Hàm băm dùng phương pháp chia: là hàm băm có dạng:

$$f(k) = k \% m$$

trong đó m là kích thước của bảng băm, k là khóa có giá trị nguyên.

Để tránh hiện tượng như ví dụ 2: kích thước bảng băm m sẽ được chọn là số nguyên tố nhỏ nhất và lớn hơn hoặc bằng kích thước bảng băm cần thiết.

CÁC KHÁI NIỆM

❖ CÁC DẠNG HÀM BĂM

- Hàm băm dùng phương pháp chia:

Ví dụ 4: Với yêu cầu bảng băm chứa 6 phần tử có khóa là 2, 4, 6, 8, 10, 12, kích thước cần cho bảng băm là 7, hàm băm dùng phương pháp chia là:

$$f(k) = k \% 7$$

Khi đó:

$$f(2) = 2, f(4) = 4, f(6) = 6, f(8) = 1, f(10) = 3, f(12) = 5.$$

CÁC KHÁI NIỆM

❖ CÁC DẠNG HÀM BẮM

- Hàm băm dùng phương pháp chia:

Trường hợp khóa có dạng chuỗi: dùng công thức Horner với m là kích thước bảng băm, n là chiều dài khóa k :

$$h_{n-1} = k[n-1] \% m$$

$$h_{n-2} = (k[n-2] + h_{n-1} * 32) \% m$$

$$h_{n-3} = (k[n-3] + h_{n-2} * 32) \% m$$

....

$$h_0 = (k[0] + h_1 * 32) \% m$$

$$f(k) = h_0$$

CÁC KHÁI NIỆM

❖ CÁC DẠNG HÀM BẮM

- Hàm băm dùng phương pháp chia:

ví dụ 5: cho bảng băm kích thước 7, dùng hàm băm theo công thức Horner tính địa chỉ cho chuỗi "for"

$$h_2 = 'r' \% 7 = 114 \% 7 = 2$$

$$h_1 = ('o' + 2*32) \% 7 = (111 + 64) \% 7 = 0$$

$$h_0 = ('f' + 0*32) \% 7 = 102 \% 7 = 4$$

$$\rightarrow f(\text{"for"}) = 4.$$

CÁC KHÁI NIỆM

❖ CÁC DẠNG HÀM BĂM

- Hàm băm dùng phương pháp nhân: là hàm băm có dạng:

$$f(k) = \lfloor m * \{k * A\} \rfloor$$

- m là kích thước của bảng băm, m thường chọn là 2^p
- k là khóa.
- A là hằng số tùy chọn trong khoảng (0, 1). Knuth đề nghị $A = (\sqrt{5} + 1) / 2$.
- $\{ \}$ là phép toán lấy phần thập phân.

CÁC KHÁI NIỆM

❖ CÁC DẠNG HÀM BĂM

- Hàm băm dùng phương pháp nhân:

Ví dụ 6: cho bảng băm kích thước 8, hàm băm dùng phương pháp nhân với hằng số A do Knuth đề nghị, tính địa chỉ của khóa $k = 110$

$$\begin{aligned} f(k) &= \lfloor 8 * \{110 * ((\sqrt{5} + 1) / 2)\} \rfloor \\ &= \lfloor 8 * \{177.984\} \rfloor = \lfloor 8 * 0.984 \rfloor = \lfloor 7.86991 \rfloor \\ &= 7 \end{aligned}$$

Lưu ý: trong C/C++, để lấy phần thập phân của một số thực x , dùng hàm: `fmod(x, 1)` được khai báo trong `cmath`