



Chương 2-Bảng băm

Nội dung

Giới thiệu về bảng băm

Hàm băm.

Các phương pháp giải quyết đụng độ

1. Phương pháp kết nối trực tiếp
2. Phương pháp kết nối hợp nhất
3. Phương pháp dò tuyến tính
4. Phương pháp dò bậc 2
5. Phương pháp băm kép

Phân tích Phép băm



Giới thiệu Bảng băm

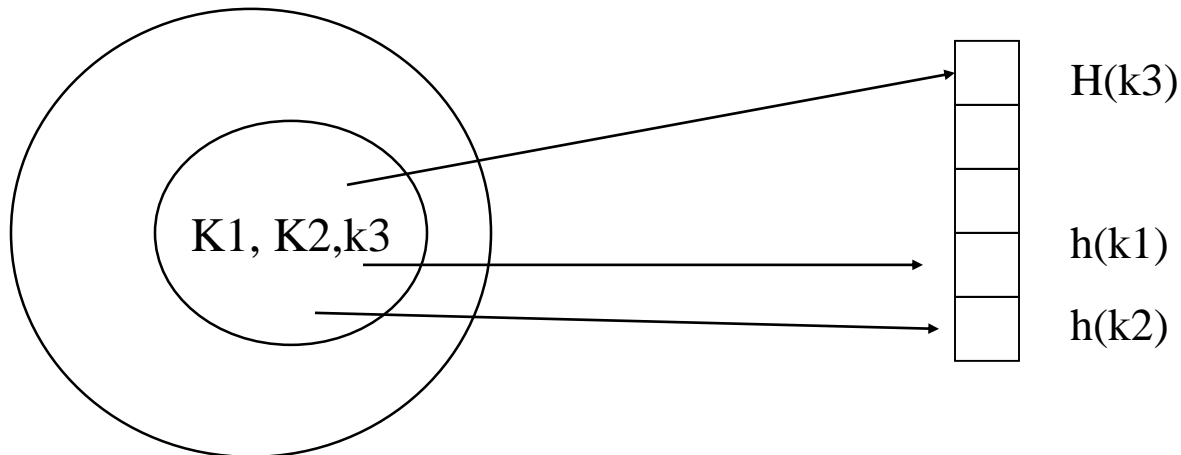
- Phép Băm (Hashing): Là quá trình ánh xạ một giá trị khóa vào một vị trí trong bảng.
- Một hàm băm ánh xạ các giá trị khóa đến các vị trí ký hiệu: h
- Bảng băm (Hash Table) là mảng lưu trữ các record, ký hiệu: HT
- HT có M vị trí được đánh chỉ mục từ 0 đến $M-1$, M là kích thước của bảng băm.
- Bảng băm thích hợp cho các ứng dụng được cài đặt trên đĩa và bộ nhớ, là một cấu trúc dung hòa giữa thời gian truy xuất và không gian lưu trữ.

HÀM BẮM (Hash functions)

- Hàm băm biến đổi một khóa vào một bảng các địa chỉ.



- Khóa có thể là dạng số hay số dạng chuỗi.
- Địa chỉ tính ra được là số nguyên trong khoảng 0 đến $M-1$ với M là số địa chỉ có trên bảng băm





Hàm băm(Hash Functions)

Hàm băm tốt thỏa mãn các điều kiện sau:

- Tính toán nhanh.

- Các khoá được phân bố đều trong bảng.

- Ít xảy ra đụng độ.

Giải quyết vấn đề băm với các khoá không phải là số nguyên:

- Tìm cách biến đổi khoá thành số nguyên

 - Trong ví dụ trên, loại bỏ dấu '-' 9635-8904 đưa về số nguyên 96358904

 - Đối với chuỗi, sử dụng giá trị các ký tự trong bảng mã ASCII

- Sau đó sử dụng các hàm băm chuẩn trên số nguyên.

HF: Phương pháp chia

Dùng số dư:

$$h(k) = k \bmod m$$

k là khoá, m là kích thước của bảng.

vấn đề chọn giá trị m

$m = 2^n$ (không tốt)

nếu chọn $m = 2^n$ thông thường k 0110010111000011010

$h(k) = k \bmod 2^n$ sẽ chọn cùng n bits cuối của k

m là nguyên tố (tốt)

Gia tăng sự phân bố đều

Thông thường m được chọn là số nguyên tố gần với 2^n

Chẳng hạn bảng ~ 4000 mục, chọn $m = 4093$

$k \bmod 2^8$ chọn các bits



HF: Phương pháp nhân

Sử dụng

$$h(k) = \lfloor m (k A \bmod 1) \rfloor$$

k là khóa, m là kích thước bảng, A là hằng số: $0 < A < 1$

Chọn m và A

M thường chọn $m = 2^p$

Sự tối ưu trong việc chọn A phụ thuộc vào đặc trưng của dữ liệu.

Theo Knuth chọn $A = \frac{1}{2}(\sqrt{5} - 1) \approx 0.618033987$ được xem là tốt.



Phép băm phổ quát

Việc chọn hàm băm không tốt có thể dẫn đến xác suất đụng độ lớn.

Giải pháp:

Lựa chọn hàm băm h ngẫu nhiên.

Chọn hàm băm độc lập với khóa.

Khởi tạo một tập các hàm băm H phổ quát và từ đó h được chọn ngẫu nhiên.

Một tập các hàm băm H là phổ quát (*universal*) nếu với mọi $\forall f, k \in H$ và 2 khóa k, l ta có xác suất: $\Pr\{f(k) = f(l)\} \leq 1/m$



Sự đụng độ (collision)

Sự đụng độ là hiện tượng các khóa khác nhau nhưng băm cùng địa chỉ như nhau.

Khi $key1 \neq key2$ mà $f(key1) = f(key2)$ chúng ta nói nút có khóa $key1$ đụng độ với nút có khóa $key2$.

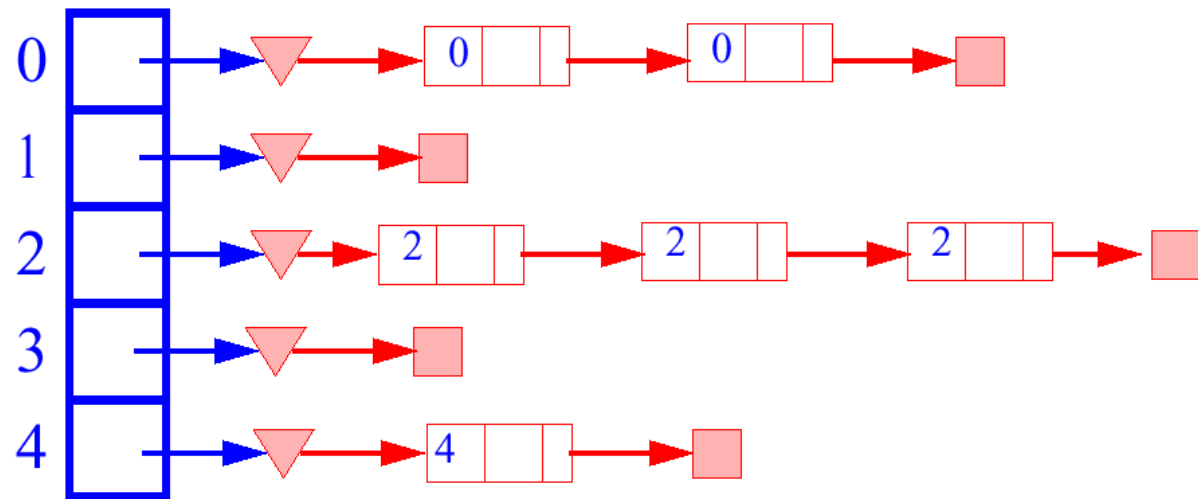
Thực tế người ta giải quyết sự đụng độ theo hai phương pháp: phương pháp nối kết và phương pháp băm lại.

Phương pháp nối kết trực tiếp

I. Phương pháp nối kết trực tiếp:

Các nút bị bấm cùng địa chỉ (các nút bị xung đột) được gom thành một danh sách liên kết.

các nút trên bảng băm được *bấm* thành các danh sách liên kết. Các nút bị xung đột tại địa chỉ i được nối kết trực tiếp với nhau qua danh sách liên kết i .



Phương pháp nối kết trực tiếp (tt)

```
#define M 100
typedef struct nodes
{
    int key;
    struct nodes *next
}NODE;
//khai bao kieu con tro chi nut
typedef NODE *PHNODE;
/*
khai bao mang HASHTABLE chua M con tro dau cua
MHASHTABLE
*/
typedef PHNODE HASHTABLE[M];
```

Phương pháp nối kết trực tiếp

Hàm băm

Giả sử chúng ta chọn hàm băm dạng %: $f(\text{key}) = \text{key} \% M$.

```
int HF (int key)
```

```
{
```

```
    return (key % M);
```

```
}
```

Phép toán khởi tạo bảng băm:

Khởi động các HASHTABLE.

```
void InitHASHTABLE( )
```

```
{
```

```
for (int i=0;<M;i++)
```

```
HASHTABLE[i]=NULL;
```

```
}
```

Phương pháp nối kết trực tiếp (*tt*)

■ Phép toán kiểm tra bảng băm rỗng HASHTABLE[i]:

```
int emptyHASHTABLE (int i)
{
    return(HASHTABLE[i] ==NULL ?1 :0);
}
```

■ Phép toán toàn bộ bảng băm rỗng:

```
int empty( )
for (int i = 0;i<M;i++)
    if(HASHTABLE[i] !=NULL)
        return 0;
return 1
}
```

Phương pháp nối kết trực tiếp (*tt*)

■ Phép toán thêm vào:

Thêm phần tử có khóa k vào bảng băm.

Giả sử các phần tử trên các HASHTABLE là có thứ tự để thêm một phần tử khóa k vào bảng băm trước tiên chúng ta xác định HASHTABLE phù hợp, sau đó dùng phép toán place của danh sách liên kết để đặt phần tử vào vị trí phù hợp trên HASHTABLE.

```
void insert(int k)
```

```
{
```

```
    int i= HF(k)
```

```
    InsertList(HASHTABLE[i],k); //phép toán thêm khoá k vào  
    danh sach lien ket HASHTABLE[i]
```

```
}
```

Phương pháp nối kết trực tiếp (*tt*)

Phép toán loại bỏ:

Xóa phần tử có khóa k trong bảng băm.

```
void remove ( int k)
{
    int b;
    PHNODE q,  p;
    b = HF(k);
    p = HASHTABLE(b);
    q=p;
```

Phương pháp nối kết trực tiếp (*tt*)

```
while(p != NULL && p->key !=k)
{
    q=p;
    p=p->next;
}
if (p == NULL) printf("\n không có nút có khóa %d" ,k);
else if (p == HASHTABLE [b])
    pop(b);
    //Tac vu pop của danh sách liên kết
else
    delafter(q);
    /*tac vu delafter của danh sách liên kết*/
}
```

Phương pháp nối kết trực tiếp (*tt*)

■ Phép toán tìm kiếm:

```
PHNODE p;  
int b;  
b = HF (k);  
p = HASHTABLE[b];  
while(k != p->key && p != NULL)  
p=p->next;  
if (p == NULL || k !=p->key)    // không tìm thấy  
return(NULL);  
else    //tìm thấy  
    return(p);  
}
```


Phương pháp nối kết trực tiếp (*tt*)

■ Phép toán duyệt HASHTABLE[i]:

Duyệt các phần tử trong HASHTABLE b.

```
void traverseHASHTABLE (int b)
```

```
{
```

```
    PHNODE p;
```

```
    p= HASHTABLE[b];
```

```
    while (p !=NULL)
```

```
    {
```

```
        printf("%3d", p->key);
```

```
    p= p->next;
```

```
    }
```

```
}
```



Phương pháp nối kết trực tiếp (*tt*)

Phép toán duyệt toàn bộ bảng băm:

Duyệt toàn bộ bảng băm.

```
void traverse( )  
{  
    int b;  
        for (b = 0; n < M; b++)  
    {  
        printf("\nButket %d:", b);  
        traverseHASHTABLE(b);  
    }  
}
```

Phương pháp nối kết trực tiếp (tt)

Nhận xét:

Bảng băm dùng phương pháp nối kết trực tiếp sẽ "băm" n nút vào M danh sách liên kết (M bucket).

Tốc độ Truy xuất phụ thuộc vào việc lựa chọn hàm băm sao cho băm đều n nút của bảng băm cho M bucket.

Nếu chọn M càng lớn thì tốc độ thực hiện các phép toán trên bảng băm càng nhanh tuy nhiên không hiệu quả về bộ nhớ. Chúng ta có thể điều chỉnh M để dung hòa giữa tốc độ truy xuất và dung lượng bộ nhớ;

Nếu chọn $M=n$ thời gian truy xuất tương đương với truy xuất trên mảng (có bậc $O(1)$), song tốn bộ nhớ.

Nếu chọn $M = n / k (k = 2, 3, 4, \dots)$ thì ít tốn bộ nhớ hơn k lần, nhưng tốc độ chậm đi k lần.

Phương pháp nối kết hợp nhất (1)

II. Bảng băm với phương pháp nối kết hợp nhất (*coalesced Chaining Method*):

Bảng băm trong trường hợp này được cài đặt bằng danh sách liên kết dùng mảng, có M nút. Các nút bị xung đột địa chỉ được nối kết nhau qua một danh sách liên kết.

Mỗi nút của bảng băm là một mẫu tin có 2 trường:

Trường key: chứa các khóa node

Trường next: con trỏ chỉ node kế tiếp nếu có xung đột.

Khi khởi động bảng băm thì tất cả trường key được gán NULLKEY, tất cả trường next được gán -1.

Key	next
NULL	-1
...	...
NULL	-1

phương pháp nối kết hợp nhất (2)

■ Khi thêm một nút có khóa key vào bảng băm, hàm băm $f(\text{key})$ sẽ xác định địa chỉ i trong khoảng từ 0 đến $M-1$.

Nếu chưa bị xung đột thì thêm nút mới vào địa chỉ này .

Nếu bị xung đột thì nút mới được cấp phát là nút trống phía cuối mảng. Cập nhật liên kết next sao cho các nút bị xung đột hình thành một danh sách liên kết.

Khi tìm một nút có khóa key trong bảng băm, hàm băm $f(\text{key})$ sẽ xác định địa chỉ i trong khoảng từ 0 đến $M-1$, tìm nút khóa key trong danh sách liên kết xuất phát từ địa chỉ i .

Phương pháp nối kết hợp nhất (3)

Minh họa cho bảng băm có tập khóa là tập số tự nhiên, tập địa chỉ có 10 địa chỉ ($M=10$) (từ địa chỉ 0 đến 9), chọn hàm băm $f(\text{key}) = \text{key} \% 10$.

Key=10, 42, 20, 109, 62

0	10	-1
1		-1
2		-1
...	null	-1
.		
8		-1
9		-1



Phương pháp nối kết hợp nhất ()

a. Khai báo cấu trúc bảng băm:

```
//Khai bao cau truc mot nut cua bang bam
typedef struct
{
    int key; //khoa cua nut tren bang bam
           int next;
//con tro chi nut ke tiep khi co xung dot
}NODE;
//Khai bao bang bam
typedef NODE HASHTABLE[M];
int avail;
```

Phương pháp nối kết hợp nhất (5)

Hàm băm:

```
int HF(int key)
{
    return(key % 10);
}
```

Phép toán khởi tạo (Initialize):

```
void initialize()
{
    int i;
    for(i = 0; i < M; i++)
    {
        HASHTABLE[i].key = NULLKEY;
        HASHTABLE[i].Next = -1;
    }
    avail = M-1; /* nút M-1 là nút ở cuối bảng chuẩn bị cấp
phát nếu có xung đột*/
}
```


Phương pháp nối kết hợp nhất (6)

Phép toán tìm kiếm (search):

```
int search(int k)
{
    int i;
    i=HF(k);
    while(k !=HASHTABLE[i].key && i !=-1)
        i=HASHTABLE[i].next;
    if(k== HASHTABLE[i]key)
        return(i);    //tim thay
    return(M);        //khong tim thay
}
```

Phương pháp nối kết hợp nhất (7)

Phép toán lấy phần tử trống (Getempty):

Chọn phần tử còn trống phía cuối bản băm để cấp phát khi xảy ra xung đột.

```
int getempty()
{
    while(HASHTABLE[avail].key != NULLKEY)
        avail - -;
    return(avail);
}
```

Phương pháp nối kết hợp nhất (8)

```
int insert(int k)
{
    int I;
    //con tro lan theo danh sach lien ket chua cac nut //bi xung dot
    int j;
    //dia chi nut trong duoc cap phat
        i = search(k);
        if(i != M)
        {
            printf("\n khoa %d bi trung,khong them nut nay duoc",k);
            return(i);
        }
        i=HF(k);
    while(HASHTABLE[i].next >=0)
        i=HASHTABLE[i].next;
```

Phương pháp nối kết hợp nhất (9)

```
if(HASHTABLE[i].key == NULLKEY)
    //Nut i con trong thi cap nhat
    j = i;
else
    //Neu nut i la nut cuoi cua DSLK
    {
        j = getempty();
        if(j < 0)
        {
            printf("\n Bang bam bi `"); return(j);
        }
        else
            HASHTABLE[i].next = j;
    }
HASHTABLE[j].key = k;
return(j);
}
```

Phương pháp dò tuyến tính (1)

III. Bảng băm với phương pháp dò tuyến tính (Linear Probing Method):

Bảng băm trong trường hợp này được cài đặt bằng danh sách kê có M nút, mỗi nút của bảng băm là một mẫu tin có một trường key để chứa khoá của nút.

Khi khởi động bảng băm thì tất cả trường key được gán NULLKEY.

Khi thêm nút có khoá key vào bảng băm, hàm băm $f(\text{key})$ sẽ xác định địa chỉ i trong khoảng từ 0 đến $M-1$:

Thêm các nút 32, 53, 22, 92, 17, 34, 24, 37, 56 vào bảng băm.

0

nullkey

1

nullkey

2

nullkey

3

nullkey


...

nullkey

$M-1$

nullkey

Thêm các nút 32, 53, 22, 92, 17, 34, 24, 37, 56 vào bảng băm.



0	NULL	0	NULL	0	NULL	0	NULL	0	56
1	NULL	1	NULL	1	NULL	1	NULL	1	NULL
2	32	2	32	2	32	2	32	2	32
3	53	3	53	3	53	3	53	3	53
4	NULL	4	22	4	22	4	22	4	22
5	NULL	5	92	5	92	5	92	5	92
6	NULL	6	NULL	6	34	6	34	6	34
7	NULL	7	NULL	7	17	7	17	7	17
8	NULL	8	NULL	8	NULL	8	24	8	24
9	NULL	9	NULL	9	NULL	9	37	9	37



Phương pháp dò tuyến tính (2)

- Nếu chưa bị xung đột thì thêm nút mới vào địa chỉ này.

Nếu bị xung đột thì hàm băm lại lần 1- f_1 sẽ xét địa chỉ kế tiếp, nếu lại bị xung đột thì hàm băm thì hàm băm lại lần 2 - f_2 sẽ xét địa chỉ kế tiếp.

Quá trình cứ thế cho đến khi nào tìm được địa chỉ trống và thêm nút vào địa chỉ này.

Khi tìm một nút có khoá key vào bảng băm, hàm băm $f(\text{key})$ sẽ xác định địa chỉ i trong khoảng từ 0 đến $M-1$, tìm nút khoá key trong khối đặt chứa các nút xuất phát từ địa chỉ i .

Phương pháp dò tuyến tính (3)

❑ Hàm băm lại của phương pháp dò tuyến tính là truy xuất địa chỉ kế tiếp. Hàm băm lại lần i được biểu diễn bằng công thức sau:

❑ $f(\text{key}) = (f(\text{key}) + i) \% M$ với $f(\text{key})$ là hàm băm chính của bảng băm.

❑ Lưu ý địa chỉ dò tìm kế tiếp là địa chỉ 0 nếu đã dò đến cuối bảng

❑ **Nhận xét:**

❑ Chúng ta thấy bảng băm này chỉ tối ưu khi băm đều, tốc độ truy xuất lúc này có bậc $O(1)$.

❑ Trường hợp xấu nhất là băm không đều hoặc bảng băm đầy, lúc này hình thành một khối đặc có n nút trên tốc độ truy xuất lúc này có bậc $O(n)$.



Phương pháp dò tuyến tính (4)

```
//khai bao cau truc mot node cua bang bam
typedef struct
{
    int key; //khoa cua nut tren bang bam
}NODE;
//Khai bao bang bam co M nut
NODE HASHTABLE[M];
int N; /*bien toan cuc chi so nut hien co tren
bang bam*/
```

Phương pháp dò tuyến tính (5)

Hàm băm:

Giả sử chúng ta chọn hàm băm dạng %: $f(\text{key}) = \text{key} \% 10$.

```
int HF(int key)
{
    return(key% 10);
}
```

Chúng ta có thể dùng một hàm băm bất kì thay cho hàm băm dạng % trên.

Phép toán khởi tạo (initialize):

Khởi tạo bảng băm.

Gán tất cả các phần tử trên bảng có trường key là NULL.

Gán biến toàn cục $N=0$.

```
void initialize( )
{
    int i;
    for(i=0;i<M;i++)
        HASHTABLE[i].key=NULLKEY;
    N=0;
    //so nut hien co khoi dong bang 0
}
```

Phương pháp dò tuyến tính (6)

■ Phép toán kiểm tra trống (empty):

Kiểm tra bảng băm có trống hay không.

```
int empty( );  
{  
    return(N==0 ? TRUE;FALSE);  
}
```

■ Phép toán kiểm tra đầy (full):

Kiểm tra bảng băm đã đầy chưa.

```
int full( )  
{  
    return (N==M-1 ? TRUE; FALSE);  
}
```

Lưu ý bảng băm đầy khi $N=M-1$, chúng ta nên dành ít nhất một phần tử trống trên bảng băm.

Phương pháp dò tuyến tính (7)

■ Phép toán search:

```
int search(int k)
{
    int i;
    i=HF(k);
    while(HASHTABLE[i].key!=k &&
HASHTABLE[i].key!=NULKEY)
    {
        i=i+1;
        if(i>=M) i=i-M;
    }
    if(HASHTABLE[i].key==k) //tim thay
return(i);
else //khong tim thay
return(M);
}
```

Phương pháp dò tuyến tính (8)

■ Phép toán insert:

Thêm phần tử có khoá k vào bảng băm.

```
int insert(int k)
```

```
{
```

```
    int i, j;
```

```
    if(search(K)<M) return M;//Trùng khoá
```

```
    if(full( ))
```

```
    {
```

```
        printf("\n Bang bam bi day khong them  
        khoa %d duoc",k);
```

nut co

```
        return;
```

```
    }
```



Phương pháp dò tuyến tính (9)

```
i=HF(k);  
    while(HASHTABLE[i].key !=NULLKEY)  
    {  
        //Bam lai (theo phuong phap do tuyen tinh)  
        i ++;  
        if(i >=M)  
            i= i-M;  
    }  
    HASHTABLE[i].key=k;  
    N=N+1;  
    return(i);  
}
```

Phương pháp dò tuyến tính (10)

Nhận xét bảng băm dùng phương pháp dò tuyến tính:

Bảng băm này chỉ tối ưu khi băm đều, nghĩa là, trên bảng băm các khối đặc chứa vài phần tử và các khối phần tử chưa sử dụng xen kẽ nhau, tốc độ truy xuất lúc này có bậc $O(1)$. Trường hợp xấu nhất là băm không đều hoặc bảng băm đầy, lúc này hình thành một khối đặc có n phần tử, nên tốc độ truy xuất lúc này có bậc $O(n)$.



Phương pháp dò bậc hai (1)

IV. Bảng băm với phương pháp dò bậc hai (Quadratic Probing Method)

Bảng băm dùng phương pháp dò tuyến tính bị hạn chế do rải các nút không đều, bảng băm với phương pháp dò bậc hai rải các nút đều hơn.

Bảng băm trong trường hợp này được cài đặt bằng danh sách kê có M nút, mỗi nút của bảng băm là một mẫu tin có một trường key để chứa khóa các nút

Khi khởi động bảng băm thì tất cả trường key bị gán NULLKEY.

Khi thêm nút có khóa key vào bảng băm, hàm băm $f(key)$ sẽ xác định địa chỉ i trong khoảng từ 0 đến $M-1$.



Phương pháp dò bậc hai (2)

- Nếu chưa bị xung đột thì thêm nút mới vào địa chỉ i này.

Nếu bị xung đột thì hàm băm lại lần 1 f_1 sẽ xét địa chỉ cách i^2 , nếu lại bị xung đột thì hàm băm lại lần 2 f_2 sẽ xét địa chỉ cách i^2 , ..., quá trình cứ thế cho đến khi nào tìm được trống và thêm nút vào địa chỉ này.

Khi tìm một nút có khóa key trong bảng băm thì xét nút tại địa chỉ $i=f(\text{key})$, nếu chưa tìm thấy thì xét nút cách $i^2, 2^2, \dots$, quá trình cứ thế cho đến khi tìm được khóa (trường hợp tìm thấy) hoặc rơi vào địa chỉ trống (trường hợp không tìm thấy).



Phương pháp dò bậc hai (3)

Hàm băm lại của phương pháp dò bậc hai là truy xuất các địa chỉ cách bậc 2.

Hàm băm lại hàm f_i được biểu diễn bằng công thức sau:

$f_i(\text{key}) = (f(\text{key}) + i^2) \% M$ với $f(\text{key})$ là hàm băm chính của bảng băm.

Nếu đã dò đến cuối bảng thì trở về dò lại từ đầu bảng.

Bảng băm với phương pháp dò bậc hai nên chọn số địa chỉ M là số nguyên tố.

Thêm vào các khóa 10, 15, 16, 20, 30, 25, ,26, 36

0	10	0	10	0	10	0	10	0	10
1	NULL	1	20	1	20	1	20	1	20
2	NULL	2	NULL	2	NULL	2	NULL	2	36
3	NULL	3	NULL	3	NULL	3	NULL	3	NULL
4	NULL	4	NULL	4	30	4	30	4	30
5	15	5	15	5	15	5	15	5	15
6	16	6	16	6	16	6	16	6	16
7	NULL	7	NULL	7	NULL	7	26	7	26
8	NULL	8	NULL	8	NULL	8	NULL	8	NULL
9	NULL	9	NULL	9	25	9	25	9	25



Phương pháp dò bậc hai (*tt*)

```
//Khai bao nut cua bang bam
typedef struct
{
    int key;           //Khoa cua nut tren bang
    bam
}NODE;
//Khai bao bang bam co M nut
NODE HASHTABLE[M];
int N;
//Bien toan cuc chi so nut hien co tren bang bam
```



Phương pháp dò bậc hai (tt)

Hàm băm:

Giả sử chúng ta chọn hàm băm dạng%: $f(\text{key}) = \text{key} \% 10$.

Tương tự các hàm băm nói trên

Chúng ta có thể dùng một hàm băm bất kì thay cho hàm băm dạng % trên.

Phép toán initialize

Gán tất cả các phần tử trên bảng có trường key là NULLKEY.

Gán biến toàn cục N=0.

```
void initialize()
```

```
{
```

```
    int i;
```

```
    for(i=0; i<M;i++)
```

```
        HASHTABLE[i].key = NULLKEY;
```

```
    N=0; //so nut hien co khoi dong bang 0
```

```
}
```

Phương pháp dò bậc hai (tt)

■ Phép toán search:

Tìm phần tử có khóa k trên bảng băm, nếu không tìm thấy hàm này trả về trị M, nếu tìm thấy hàm này trả về địa chỉ tìm thấy.

```
int search(int k)
{
    int i, d;
    i = hashfuns(k);
    d = 1;
    while(HASHTABLE[i].key!=k&&HASHTABLE[i].key!=NULLKEY)
    {
        //Bam lai (theo phuong phap bac hai)
        i = (i+d) % M;
        d = d+2;
    }
    if(HASHTABLE[i].key == k)
        return i;
    return M;
}
```

Phương pháp dò bậc hai (tt)

■ Phép toán insert:

Thêm phần tử có khoá k vào bảng băm.

```
int insert(int k)
{
    int i, d;
    i = hashfuns(k);
    d = 1;
    if(search(K)<M) return M;//Trùng khoá
    if(full( ))
    {
        printf("\n Bang bam bi day khong them nut co
                khoa %d duoc",k);
        return;
    }
}
```



Phương pháp dò bậc hai (*tt*)

```
i=HF(k);  
while(HASHTABLE[i].key !=NULLKEY)  
{  
    //Bam lai (theo phuong phap bac hai)  
    i = (i+d) % M;  
    d = d+2;  
}  
HASHTABLE[i].key=k;  
N=N+1;  
return(i);  
}
```




Phương pháp dò bậc hai (tt)

Nhận xét bảng băm dùng phương pháp dò bậc hai:

Nên chọn số địa chỉ M là số nguyên tố. Khi khởi động bảng băm thì tất cả M trường key được gán NULL, biến toàn cục N được gán 0.

Bảng băm đầy khi $N = M-1$, và nên dành ít nhất một phần tử trống trên bảng băm.

Bảng băm này tối ưu hơn bảng băm dùng phương pháp dò tuyến tính do rải rác phần tử đều hơn, nếu bảng băm chưa đầy thì tốc độ truy xuất có bậc $O(1)$. Trường hợp xấu nhất là bảng băm đầy vì lúc đó tốc độ truy xuất chậm do phải thực hiện nhiều lần so sánh.

Phương pháp băm kép

Mô tả:

- Cấu trúc dữ liệu: Bảng băm này dùng hai hàm băm khác nhau với mục đích để rải rác đều các phần tử trên bảng băm.

Chúng ta có thể dùng hai hàm băm bất kì, ví dụ chọn hai hàm băm như sau:

$$f1(key) = key \% M.$$

$$f2(key) = (M-2) - key \% (M-2).$$

bảng băm trong trường hợp này được cài đặt bằng danh sách kề có M phần tử, mỗi phần tử của bảng băm là một mẫu tin có một trường key để lưu khoá các phần tử.

Phương pháp băm kép (tt)

- Khi khởi động bảng băm, tất cả trường key được gán NULL.

- Khi thêm phần tử có khoá key vào bảng băm, thì $i=f1(key)$ và $j=f2(key)$ sẽ xác định địa chỉ i và j trong khoảng từ 0 đến $M-1$:

Nếu chưa bị xung đột thì thêm phần tử mới tại địa chỉ i này.

Nếu bị xung đột thì hàm băm lại lần 1 $f1$ sẽ xét địa chỉ mới $i+j$, nếu lại bị xung đột thì hàm băm lại lần 2 là $f2$ sẽ xét địa chỉ $i+2j$, ..., quá trình cứ thế cho đến khi nào tìm được địa chỉ trống và thêm phần tử vào địa chỉ này.

Thêm vào các khóa 10, 15, 16, 20, 30, 25, 26, 36 :

$f1(key) = key \% M.$

$f2(key) = (M-2) - key \% (M-2).$

0	10	0	10	0	10	0	10	0	10
1	NULL	1	NULL	1	NULL	1	20	1	20
2	NULL	2	NULL	2	30	2	30	2	26
3	NULL	3	NULL	3	NULL	3	NULL	3	36
4	NULL	4	20	4	NULL	4	NULL	4	20
5	15	5	15	5	15	5	15	5	15
6	16	6	16	6	16	6	16	6	16
7	NULL	7	NULL	7	NULL	7	NULL	7	NULL
8	NULL	8	NULL	8	NULL	8	26	8	NULL
9	NULL	9	NULL	9	25	9	25	9	25

Phương pháp băm kép (tt)

- Khi tìm kiếm một phần tử có khoá key trong bảng băm, hàm băm $i=f_1(key)$ và $j=f_2(key)$ sẽ xác định địa chỉ i và j trong khoảng từ 0 đến $M-1$. Xét phần tử tại địa chỉ i , nếu chưa tìm thấy thì xét tiếp phần tử $i+j$, $i+2j$, ..., quá trình cứ thế cho đến khi nào tìm được khoá (trường hợp tìm thấy) hoặc bị rơi vào địa chỉ trống (trường hợp không tìm thấy).



Phương pháp băm kép (*tt*)

```
//Khai bao phan tu cua bang bam
typedef struct
{
    int key;    //khoa cua nut tren bang bam
}NODE;
//khai bao bang bam co M nut
struct node HASHTABLE[M];
int N;
//bien toan cuc chi so nut hien co tren bang bam
```

Phương pháp băm kép (tt)

Hàm băm:

Giả sử chúng ta chọn hai hàm băm dạng %:

$f_1(\text{key}) = \text{key} \% M$ và $f_2(\text{key}) = M - 2 - \text{key} \% (M - 2)$.

//Hàm băm thu nhất

```
int HF(int key)
```

```
{
```

```
    return(key%M);
```

```
}
```

//Hàm băm thu hai

```
int HF2(int key)
```

```
{
```

```
    return(M-2 - key%(M-2));
```

```
}
```



Phương pháp băm kép (*tt*)

■ Phép toán initialize :

Khởi động bảng băm.

Gán tất cả các phần tử trên bảng có trường key là NULL.

Gán biến toàn cục $N = 0$.

```
void initialize()
```

```
{
```

```
    int i;
```

```
    for (i = 0 ; i < M ; i++ )
```

```
        HASHTABLE [i].key = NULLKEY;
```

```
    N = 0; // so nut hien co khoi dong bang 0
```

```
}
```


Phương pháp băm kép (*tt*)

■ Phép toán empty :

Kiểm tra bảng băm có rỗng không.

```
int empty() .
```

```
{
```

```
    return (N == 0 ? TRUE : FALSE) ;
```

```
}
```

■ Phép toán full :

Kiểm tra bảng băm đã đầy chưa.

```
int full() .
```

```
{
```

```
    return (N == M-1 ? TRUE : FALSE) ;
```

```
}
```

Lưu ý bảng băm đầy khi $N=M-1$, chúng ta nên dành ít nhất một phần tử trống trên bảng băm.

Phương pháp băm kép (*tt*)

■ Phép toán search :

```
int search(int k)
{
    int i, j ;
    i = HF (k);
    j = HF2 (k);
    While (HASHTABLE [i].key!=k &&HASHTABLE [i] .key !=
NULLKEY)
        i = (i+j) % M ; //băm lại (theo phương pháp băm kép)
    if (HASHTABLE [i].key == k) // tìm thấy
        return (i) ;
    else
        return (M) ; // không tìm thấy
}
```



Phương pháp băm kép (*tt*)

■ Phép toán insert :

Thêm phần tử có khoá k vào bảng băm.

```
int insert(int k)
{
    int i, j;
    if(search(k)<M) return M;//trùng khóa
    if (full () )
    {
        printf ("Bang bam bi day") ;
        return (M) ;
    }
}
```



Phương pháp băm kép (*tt*)

```
if (search (k) < M)
{
    printf ("Da co khoa nay trong bang bam") ;
    return (M) ;
}
i = HF (k) ;
j = HF 2 (k) ;
while (HASHTABLE [i].key != NULLEY)
// Bam lai (theo phuong phap bam kep)
i = (i + j) % M;
HASHTABLE [i].key = k ;
N = N+1;
return (i) ;
}
```



Tóm tắt về bảng băm

- Bảng băm đặt cơ sở trên mảng.

- Phạm vi các giá trị khóa thường lớn hơn kích thước của mảng.

- Một giá trị khóa được băm thành một chỉ mục của mảng bằng hàm băm.

- Việc băm một khóa vào vào một ô đã có dữ liệu trong mảng gọi là sự đụng độ.

- Sự đụng độ có thể được giải quyết bằng hai phương pháp chính: Phương pháp nối kết và phương pháp băm lại.



Tóm tắt về bảng băm (tt)

Trong phương pháp băm lại, các mục dữ liệu được băm vào các ô đã có dữ liệu sẽ được đưa vào ô khác trong mảng.

Trong phương pháp nối kết, mỗi phần tử trong mảng có một danh sách liên kết. Các mục dữ liệu được băm vào các ô sẽ được đưa vào danh sách ở ô đó.

Vấn đề Hàm băm

a. Hàm băm dùng phương pháp chia: $h(k) = k \bmod m$
 m là kích thước bảng băm, k là khóa.

b. Hàm băm dùng phương pháp nhân: $h(k) = \lfloor m(kA \bmod 1) \rfloor$
Knuth đề nghị $A = 0.6180339887$

Tóm tắt về bảng băm (tt)

Kích thước bảng băm	PP chia	PP Nhân
200	698	699
512	470	466
997	309	288
1024	301	292

Theo bảng trên kết quả cho thấy kích thước bảng băm tỷ lệ nghịch với số lần đụng độ. Số đụng độ còn phụ thuộc vào phương pháp sử dụng hàm băm. Hệ số tải là tỉ số giữa các mục dữ liệu trong một bảng băm với kích thước của mảng.

Hệ số tải cực đại trong phương pháp băm lại khoảng 0,5. Đối với băm kép ở Hệ số tải này (0,5), các phép tìm kiếm sẽ có chiều dài thăm dò trung bình là 2.



Tóm tắt về bảng băm (*tt*)

Trong phương pháp băm lại , thời gian tìm kiếm sẽ là vô cực khi hệ số tải đạt đến 1.

Điều quan trọng trong phương pháp băm lại là bảng băm không bao giờ được quá đầy.

Phương pháp nối kết thích hợp với hệ số tải là 1.

Với hệ số tải này, chiều dài thăm dò trung bình là 1,5 khi phép tìm thành công, và là 2.0 khi phép tìm thất bại.



Tóm tắt về bảng băm (tt)

- Chiều dài thăm dò trong phương pháp nối kết tăng tuyến tính theo hệ số tải. Kích thước của bảng băm thường là số nguyên tố. Điều này đặc biệt quan trọng trong thăm dò bậc hai và trong phương pháp nối kết. Các bảng băm có thể dùng cách lưu trữ ngoại. Một cách để thực hiện việc này là cho các phần tử trong bảng băm chứa số lượng các khối của tập tin trên đĩa