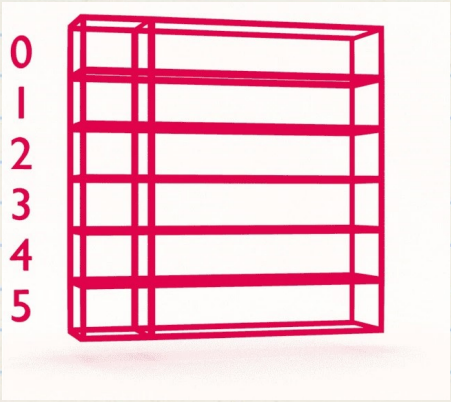
CHƯƠNG 6. BẢNG BĂM

ThS. Nguyễn Chí Hiếu

2021

NỘI DUNG

Giới thiệu bảng băm



Giới thiệu bảng băm

Định nghĩa

Bảng băm (Hash Table) là một cấu trúc dữ liệu phổ biến trong việc lưu trữ dữ liệu chưa có thứ tự. Mỗi mẩu tin của dữ liệu sẽ được đưa qua một hàm băm để lưu vào chỉ mục tương ứng trong bảng băm. Các thao tác thêm, xóa, tìm kiếm trên bảng băm được thực hiện với độ phức tạp tuyến tính trong trường hợp bảng băm được thiết kế tốt.

Định nghĩa

Hàm băm (Hash) là phép biến đổi/ánh xạ khóa của một mẩu tin thành một chỉ mục trong bảng băm.

Bảng băm thường được dùng trong bài toán tìm kiếm dữ liệu theo một chỉ mục nào đó.

Giới thiệu bảng băm

Định nghĩa

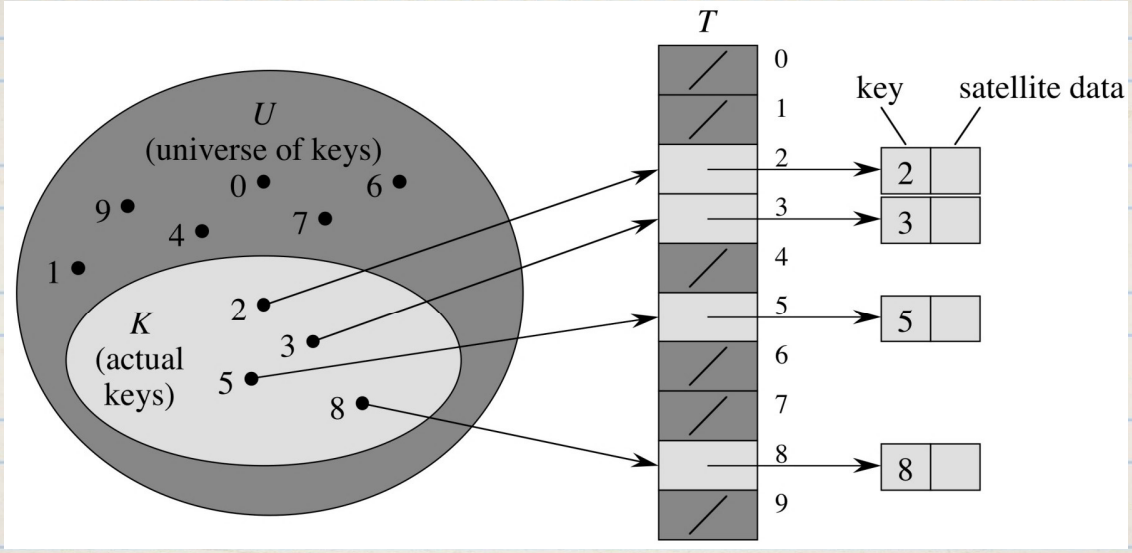
Phép biến đổi khóa là một ánh xạ từ tập hợp vũ trụ $\mathcal{U}$ của tất cả các khóa vào tập $S \in \mathcal{U}$.

$$\begin{array}{rcl} h : \mathcal{U} & \rightarrow & S = \{0, 1, \ldots, m-1\} \\ k & \mapsto & h(k) \end{array} \tag{1}$$

Trong đó,

- $\mathcal{U}$ là tập vũ trụ (chứa số lượng rất lớn các khóa)
- S là tập các chỉ mục trong bảng băm
- $h(k)$ là một hàm băm cho trước

Giới thiệu bảng băm



Hình 1: Mô tả cách lưu trữ của bảng băm.

Giới thiệu bảng băm

Khóa của hàm băm:

- Phải là một số nguyên không dấu
- Có thể dạng số (số thực cần chuyển thành số nguyên) hay chuỗi (sau khi chuyển thành mã ASCII)

Phân loại hàm băm

Một hàm băm tốt phải thỏa các điều kiện sau:

- Tính toán nhanh.
- Các khóa được phân bố đều trong bảng.
- Ít xảy ra đụng độ.
- Xử lý được các loại khóa có kiểu dữ liệu khác nhau.

Phân loại hàm băm

Hàm băm sử dụng phương pháp chia lấy phần dư (modulo)

$$h(k) = k \bmod m \tag{2}$$

- k là khóa cần lưu trữ
- m là kích thước/số lượng chỉ mục của bảng băm

Phân loại hàm băm

Chọn m sẽ ảnh hưởng đến giá trị của $h(k)$:

- Nếu chọn $m = 2^n$ thì giá trị của $h(k)$ sẽ là n bit cuối cùng của k trong biểu diễn nhị phân.
- Nếu chọn $m = 10^n$ thì giá trị của $h(k)$ sẽ là n chữ số cuối cùng của k trong biểu diễn thập phân.
- Hai cách trên giá trị của $h(k)$ chỉ phụ thuộc vào n bit (n chữ số cuối). Thường chọn m là số nguyên tố để $h(k)$ phụ thuộc vào khóa.

Phân loại hàm băm

Hàm băm sử dụng phương pháp nhân

Tác giả Knuth đã đề xuất hàm băm có công thức như sau:

$$h(k) = \lfloor m \cdot (k \cdot A \bmod 1) \rfloor \tag{3}$$

Trong đó,

- k là khóa
- m là kích thước bảng băm
- A là hằng số với $0 < A < 1$

Theo tác giả Knuth, chọn m và A như sau:

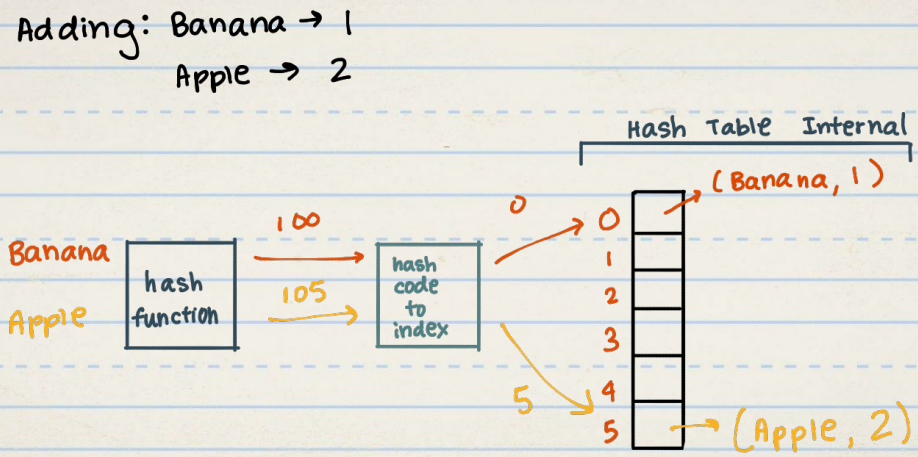
- Thường chọn $m = 2^n$
- Giá trị A thường chọn $A = \frac{\sqrt{5}-1}{2} \approx 0.61803398874989$

Các thao tác với bảng băm

Các thao tác cơ bản trong bảng băm:

- Khởi tạo bảng băm
- Thêm phần tử vào bảng băm
- Xóa phần tử khỏi bảng băm
- Tìm kiếm một phần tử trong bảng băm

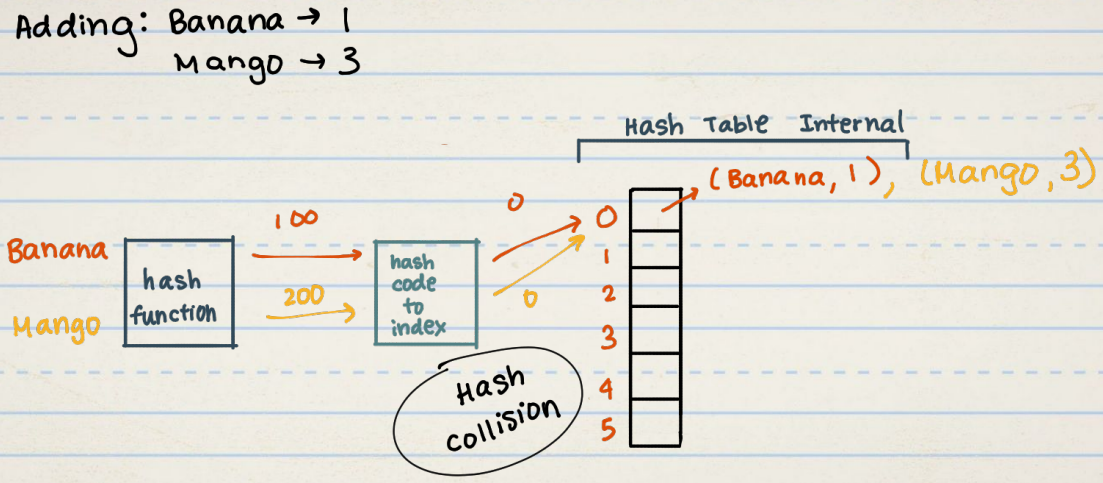
Các phương pháp xử lý đụng độ của hàm băm



Nguồn: *https://guides.codepath.com/compsci/Hash-Tables*

Hình 2: Bảng băm không xảy ra đụng độ.

Các phương pháp xử lý đụng độ của hàm băm



Nguồn: *https://guides.codepath.com/compsci/Hash-Tables*

Hình 3: Bảng băm xảy ra đụng độ khi hai khóa cùng chỉ mục.

Các phương pháp xử lý đụng độ của hàm băm

- Phương pháp tạo dây chuyền/kết nối trực tiếp (Direct chaining): mỗi chỉ mục của bảng có chứa một danh sách liên kết đơn. Các chỉ mục này lưu địa chỉ phần tử đầu tiên của danh sách.
- Phương pháp dùng địa chỉ mở (Open addressing): nếu xảy ra đụng độ, thì tìm chỉ mục kế tiếp cho đến khi tìm thấy hoặc chỉ mục trống (không tìm thấy). Các chỉ mục của bảng băm lưu một phần tử duy nhất.
  - Phương pháp dò tuyến tính (Linear probing)
  - Phương pháp dò bậc 2 (Quadratic probing)
  - Phương pháp băm kép (Double hashing)

Các phương pháp xử lý đụng độ của hàm băm

Phương pháp nối kết trực tiếp

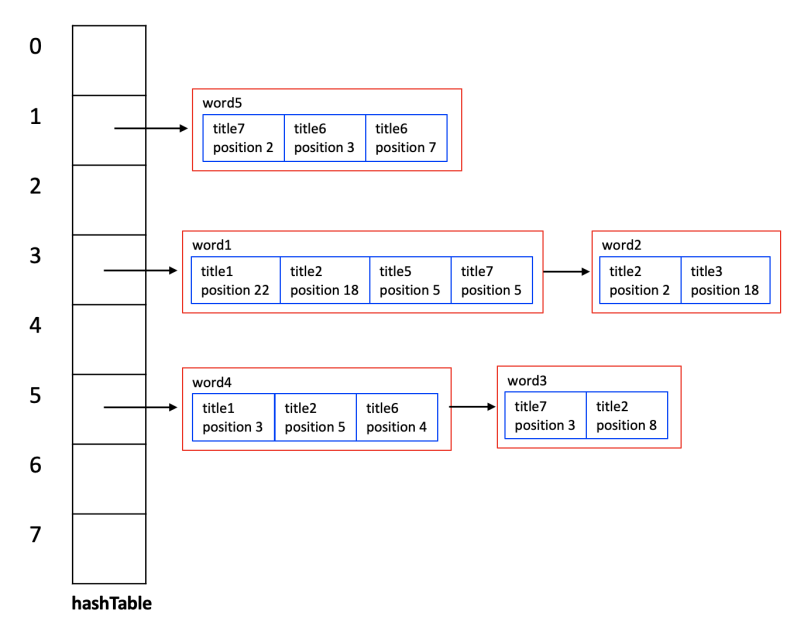
- Các chỉ mục của bảng băm sẽ được **băm** thành m danh sách liên kết. Nên bảng băm sẽ gồm m chỉ mục chứa địa chỉ đầu của các danh sách liên kết.
- Tại mỗi chỉ mục của bảng băm có một danh sách liên kết chứa các khóa khác nhau nhưng có cùng chỉ mục.

Các phương pháp xử lý đụng độ của hàm băm

Ví dụ 1

Phim được lưu trữ thông tin gồm tiêu đề phim và mô tả tóm tắt nội dung. Tiêu đề cũng như mô tả của bộ phim có thể chứa một hoặc nhiều từ. Trong bảng băm, mỗi địa chỉ là các từ trong mô tả phim và các giá trị là số lần xuất hiện của từ (mỗi bộ phim mà từ đó xuất hiện và vị trí tương ứng của nó trong mô tả).

Các phương pháp xử lý đụng độ của hàm băm



Nguồn: *https://ds.cs.rutgers.edu/assignment-rumdb/*

Hình 4: Xử lý đụng độ của hàm băm bằng phương pháp nối kết trực tiếp.

Các phương pháp xử lý đụng độ của hàm băm

Kỹ thuật địa chỉ mở sử dụng phương pháp dò tuyến tính

- Khi xảy ra đụng độ, tiếp tục dò địa chỉ phần tử kế tiếp, nếu địa chỉ trống thì thêm vào.
- Sử dụng một hàm băm tốt $h(x)$ để định nghĩa m hàm băm $h_i(x)$:

$$h_i(x) = (h(x) + i) \bmod m, 0 \leq i \leq m - 1 \quad (4)$$

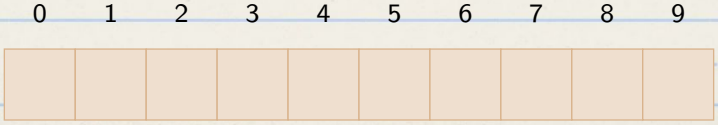
Ví dụ 2

Cho hàm băm $h(k) = k \bmod 13$. Hãy lần lượt thêm các khóa 89,18,49,58,79 vào bảng băm.

Kỹ thuật địa chỉ mở sử dụng phương pháp dò tuyến tính

Ví dụ 3

Cho bảng băm chứa 10 chỉ mục. Hãy lần lượt thêm các khóa 89, 18, 49, 58, 79 vào bảng băm *(trường hợp đụng độ xử lý bằng phương pháp dò tuyến tính)*.



Kỹ thuật địa chỉ mở sử dụng phương pháp dò tuyến tính

Các phương pháp xử lý đụng độ của hàm băm

Kỹ thuật địa chỉ mở sử dụng phương pháp dò bậc 2

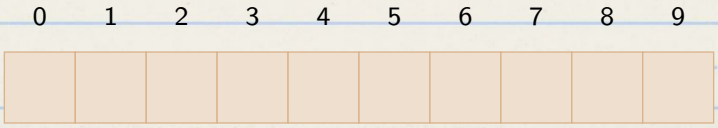
- Tương tự phương pháp dò tuyến tính, nhưng sử dụng hàm băm là một hàm bậc 2:

$$h_i(x) = (h(x) + i^2) \bmod m, 0 \leq i \leq m-1 \quad (5)$$

Kỹ thuật địa chỉ mở sử dụng phương pháp dò bậc 2

Ví dụ 4

Cho bảng băm chứa 10 chỉ mục. Hãy lần lượt thêm các khóa 89, 18, 49, 58, 79 vào bảng băm *(trường hợp đụng độ xử lý bằng phương pháp dò bậc 2)*.



Kỹ thuật địa chỉ mở sử dụng phương pháp dò bậc 2

Các phương pháp xử lý đụng độ của hàm băm

Kỹ thuật địa chỉ mở sử dụng phương pháp băm kép

- Kết hợp hai hàm băm $h(k), g(k)$ độc lập để định nghĩa m hàm băm:

$$h_1(k) = (h(k) + ig(k)) \bmod m, 0 \leq m \leq m - 1. \quad (6)$$

- Khóa được phân bố đều hơn phương pháp dò tuyến tính.

Kỹ thuật địa chỉ mở sử dụng phương pháp băm kép

Ví dụ 5

Cho bảng băm chứa 10 chỉ mục và hai hàm băm $h_1(k) = k \bmod 10$ và $h_2(k) = 7 - (k \bmod 7)$. Hãy lần lượt thêm các khóa 89, 18, 49, 58, 79 vào bảng băm.

Kỹ thuật địa chỉ mở sử dụng phương pháp băm kép

Độ phức tạp của thuật toán

Đối với thao tác tìm kiếm trong một bảng băm hoàn hảo *(hàm băm phân phối đều các khoá vào các vị trí trong bảng băm)*, thì độ phức tạp thuật toán

- Trường hợp xấu nhất: $O(n)$
- Trường hợp trung bình: hiệu quả phụ thuộc trên mức độ đầy α với $\alpha = n/m$, là tỷ số giữa số chỉ mục đã sử dụng và kích thước của bảng băm.
  - Tìm thành công:

$$\frac{1}{2}\left(1 + \frac{1}{1-\alpha}\right) \tag{7}$$

  - Tìm thất bại:

$$\frac{1}{2}\left(1 + \frac{1}{(1-\alpha)^2}\right) \tag{8}$$

Bài tập

1. Cho bảng băm kích thước 11 chỉ mục và hàm băm $h(k) = (5k + 4) \bmod 11$. Thêm lần lượt các khóa 3, 9, 2, 1, 14, 6, 25 vào bảng băm. Nếu xảy ra đụng độ, xử lý bằng phương pháp nối kết trực tiếp.
2. Cho bảng băm kích thước 11 chỉ mục và hàm băm $h(k) = k \bmod 11$. Thêm lần lượt các khóa 0, 1, 8, 9, 52, 44, 56, 53, 61, 64 vào bảng băm. Xử lý trường hợp đụng độ theo các cách sau:
   a) Phương pháp dò tuyến tính
   b) Phương pháp dò bậc 2
   c) Phương pháp sử dụng băm kép

Bài tập

3. Trong bảng băm, số lượng khóa có thể ít hơn chỉ mục hay không? Trường hợp nào?
4. Cho bảng băm lưu trữ các khóa như hình. Hãy cho biết mức độ đầy α của bảng băm là bao nhiêu?

0	1	2	3	4	5	6	7	8	9
49		58	79					18	89

5. Viết mã giả đoạn chương trình thực hiện thao tác xác định chỉ mục còn trống trong bảng băm tương ứng các cách sau:
   a) Phương pháp dò tuyến tính
   b) Phương pháp dò bậc 2
   c) Phương pháp sử dụng băm kép

Tài liệu tham khảo


Donald E. Knuth.
The Art of Computer Programming, Volume 3.
Addison-Wesley, 1998.


Dương Anh Đức, Trần Hạnh Nhi.
Nhập môn Cấu trúc dữ liệu và Thuật toán.
Đại học Khoa học tự nhiên TP Hồ Chí Minh, 2003.


Niklaus Wirth.
Algorithms + Data Structures = Programs.
Prentice-Hall, 1976.


Robert Sedgewick.
Algorithms in C.
Addison-Wesley, 1990.