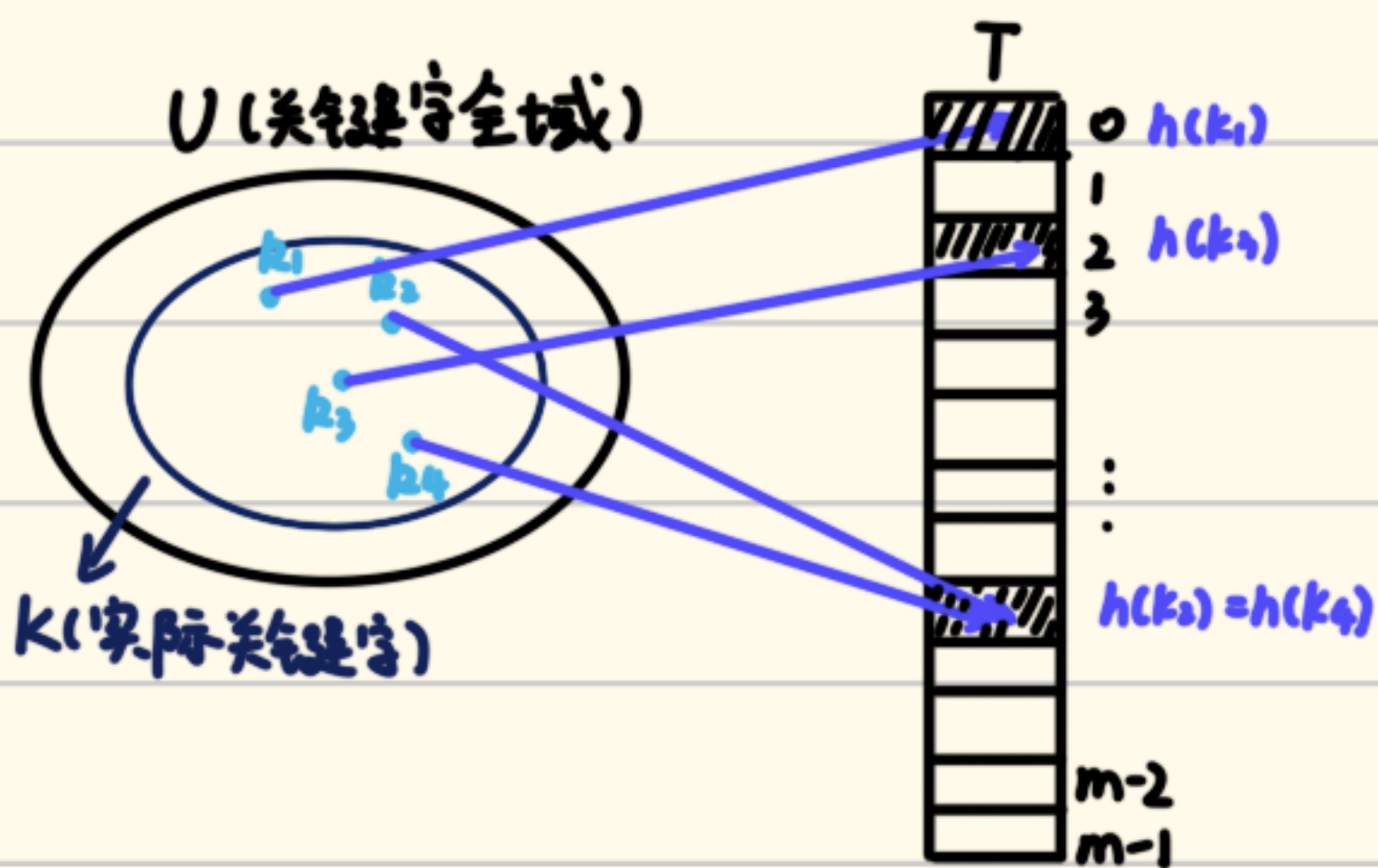


Lecture 5 Hashing

(I) Hash Table:

- Key Words Set : K
- Hash Table : $T[0, 1, \dots, m-1]$
- Hash Function: $h: U \rightarrow \{0, 1, 2, \dots, m-2, m-1\}$

Typically, $m \ll |U|$



Two methods to deal with the conflicts { 链接法(chaining)
开放寻址法(open addressing)

(II) Chaining

1. Basic Idea

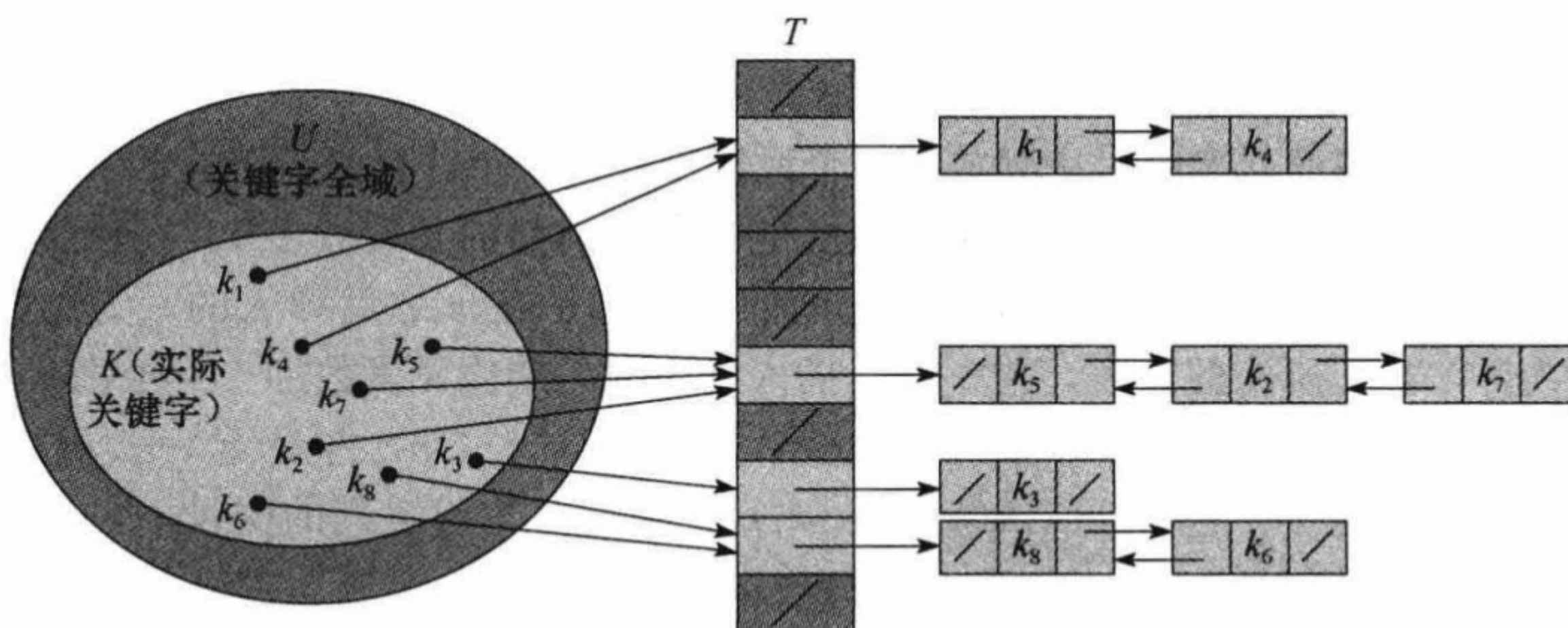


图 11-3 通过链接法解决冲突。每个散列表槽 $T[j]$ 都包含一个链表，其中所有关键字的散列值均为 j 。例如， $h(k_1)=h(k_4)$ ，还有 $h(k_5)=h(k_7)=h(k_2)$ 。这个链表可能是单链表，也可能是双向链表；图中链表画为双链，因为删除操作比较快

2. Hash Function

① 除法散列法:

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

② 乘法散列法:

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

$(kA \bmod 1)$ 表示 kA 的小数部分

3. Analysis

① Load Factor (装载因子)

散列表T有m个槽位, n个元素

$$\alpha = \frac{n}{m}$$

即一个链的平均有 α 个元素

$$n = n_0 + n_1 + n_2 + \dots + n_{m-2} + n_{m-1}$$

在链接法中, α 可以小于1/等于1/大于1

② Simple Uniform Hashing Assumption

假定任何一个给定元素等可能地散列至m个槽上

③ Time Complexity:

Under the Simple Uniform Hashing Assumption:

• Successful Search: $O(1 + \alpha)$

• Failed Search: $O(1 + \alpha)$

(III). Open Addressing

在开放寻址法中, 所有元素都有放在散列表里.

每个表项要么有元素, 要么没有元素是Nil

$\Rightarrow$ 装载因子 α 一定不会超过1

1. 探查 (probe)

- 为使用开放寻址法插入一个元素, 需要连续地检查散列表, 直至找到一个空槽来放置待插入的关键词为止
- 我们需要有一个探查序列 $\langle h(k, 0), h(k, 1), \dots, h(k, m-1) \rangle$

其中 $h(k, i)$ 中的 i 为 index

探查序列 {

- ① 线性探查
- ② 二次探查
- ③ 双重探查

2. 线性探查

① 散列函数: $h(k, i) = (h'(k) + i) \bmod m$ $i = 0, 1, \dots, m-1$

初始探查位置决定了整个序列, 只有m种不同的探查序列

② 会出现“一次群集”现象的问题

当一个空槽前有 i 个满的槽时, 该空槽为下一个被占用的概率为 $\frac{i+1}{m}$

连续被占用的槽越来越长, 平均查找时间也越来越大

3. 二次探查

① 散列函数: $h(k, i) = (h'(k) + c_1 i + c_2 i^2) \bmod m$ $i = 0, 1, \dots, m-1$

初始探查位置决定了整个序列, 只有 m 种不同的探查序列

② 避免了一次群集的问题, 但可能会造成二次群集

4. 双重散列

① 散列函数: $h(k, i) = (h_1(k) + h_2(k) \cdot i) \bmod m$ $i = 0, 1, \dots, m-1$

步长为 $h_2(k)$ 而非常数, 要求 $\forall k, h_2(k)$ 要与表的大小 m 互素

双重散列法用到了 $\Theta(m^2)$ 种探查序列, 线性/二次探查只用了 $\Theta(m)$ 种

5. Analysis

• Successful Search: $\frac{1}{\alpha} \ln \frac{1}{1-\alpha}$

• Failed Search: $\frac{1}{1-\alpha}$

2、将关键字序列 {7, 8, 30, 11, 18, 9, 14} 散列存储到散列表中, 散列表的存储空间是一个下标从 0 开始的一维数组, 散列函数为 $H(\text{key}) = (\text{key} * 3) \bmod 7$ 。处理冲突采用线性探查法, 要求装填因子为 0.7。

(1) 请画出所构造的散列表。

(2) 分别计算等概率情况下, 查找成功和查找不成功的平均查找长度。

1) $|K| = 7, \alpha = 0.7 \Rightarrow$ 散列表长度为 10

7	14		8		11	30	18	9	
0	1	2	3	4	5	6	7	8	9

散列函数: $H(\text{key}) = (\text{key} * 3) \bmod 7$

$(7 \times 3) \bmod 7 = 0$, $(8 \times 3) \bmod 7 = 3$, $(30 \times 3) \bmod 7 = 6$, $(11 \times 3) \bmod 7 = 5$

$(18 \times 3) \bmod 7 = 5$, $(9 \times 3) \bmod 7 = 6$, $(14 \times 3) \bmod 7 = 0$

(2) 查找成功

7	14		8		11	30	18	9	
0	1	2	3	4	5	6	7	8	9
1	2		1		1	1	3	3	

$$\frac{1}{7} (1 + 1 + 1 + 1 + 3 + 3 + 2) = \frac{12}{7}$$

7	14		8		11	30	18	9	
0	1	2	3	4	5	6	7	8	9
3	2	1	2	1	5	4			

$$\frac{1}{7} (3 + 2 + 1 + 2 + 1 + 5 + 4) = \frac{18}{7}$$