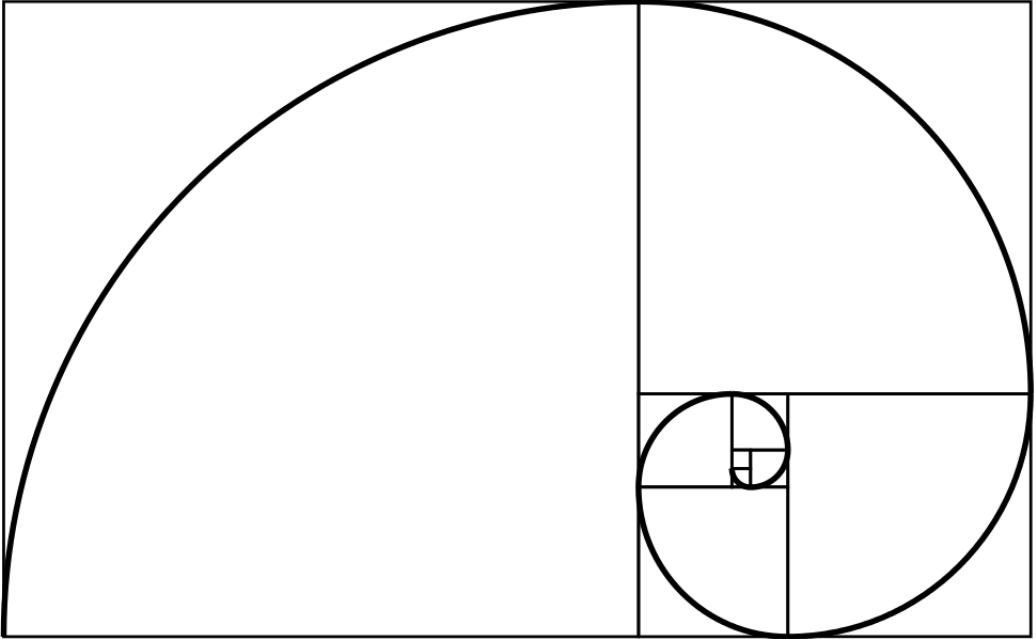


# 基本函数式算法



刘新宇<sup>1</sup>

January 2, 2025

<sup>1</sup>刘新宇

Version:  $e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + \frac{1}{1} + \frac{1}{1 \cdot 2} + \frac{1}{1 \cdot 2 \cdot 3} + \dots = 2.718283$

<https://github.com/liuxinyu95>



## 第二版前言

很多人的工作在身后才得以出版,如费马、伽罗瓦、吴敬梓、曹雪芹……很多人不肯发表不完美的作品,即使在后人看来已经很优秀、很重要了,如高斯;很多人的一些工作甚至毁于天灾人祸,遗失了,如黎曼。我们应该感谢萨缪尔·费马(整理了父亲散落的手稿和页边笔记)、刘维尔(发掘整理了伽罗瓦的遗稿)、戴德金(抢救了黎曼部分被焚毁的笔记),以及那些佚名资助、整理、出版这些作品的人。我们应该感谢那些不计经济盈亏,坚持保存、传播人类知识的出版社。今天,科技和互联网的进步大大改善了工作、记录、整理、出版的漫长过程,我得以锱铢积累地从 2009 年开始把一些习作、心得、发现用中英文分享到互联网上的版本控制系统中。到 2015 年,这些内容大致汇集成一本关于基本函数式算法和数据结构的初稿,并在 2017 年出版,名为《算法新解》。

对比前人,这真是太幸运了。我的工作远远比不上他们哪怕一丝的光辉,而且还有众多国内外优秀同行的作品尚未出版、引进给读者呢。我只有怀着感恩的心情,感谢编辑的信任,感谢所有读者的匆匆一瞥和容忍,然后听取、收集大家的意见,尽最大的努力修改、订正。我从 2020 年底开始重写此书,到 2023 年 5 月完成。这一版的主要变化有:

1. 重写了所有章节,尽量用符号和形式化的表达替换掉繁冗的叙述。
2. 调整了内容结构。(a) 将列表提前为第一章,方便不熟悉函数式编程的读者入门;(b) 增加了 B-树的列表对实现。在附录中增加了红黑树和 AVL 树的命令式删除算法;(c) 去掉了后缀树一章,在命令式字符串匹配部分,集中介绍最具代表性的 KMP 算法,去掉了 Boyer-Moore 匹配算法;(d) 增加了全部 120 道习题的答案。
3. 根据第一版读者反馈,统一了例子程序语言。函数式用 Haskell 作为例子,命令式用虚拟语言 Bourbaki (布尔巴基是 20 世纪一群顶级数学家组成的学派的笔名,包括安德烈·韦伊、亨利·嘉当等。虚拟人物尼古拉·布尔巴基被称为最伟大却并不存在的数学家,暗示 Bourbaki 语言不存在,但却能从中看到一些熟悉的影子,如 Python、Java、C 等)。并尽量把例子程序置于各章后的附录。
4. 修正了一些错误。

冈崎(Chris Okasaki)和伯德(Richard Bird)的优秀工作奠定了基本函数式数据结构和算法的主要内容。得益于此,本书可以说是对他们所建框架的剪裁与组织。尽

管函数式编程的应用和前沿技术发展日新月异,基本算法部分却相对稳定。这使得我可以用十多年时间逐渐修改、完善,并把本书<sup>1</sup>呈现到读者面前。

二零二三年十月于北京

---

<sup>1</sup>本书电子版可以在 [github](#) 上获得,如果希望获得纸质版,请在 [github](#) 上联系我。



# 第一版前言

尽管我们在课堂上学习基本算法,但除了编程竞赛、求职面试,在实际工作中很少用到它们。很多算法和数据结构已经在程序库中实现好了,我们只需要了解如何使用。谈到人工智能和机器学习算法时,实际上说的是数学模型而非基本算法和数据结构。基本算法在解决一些“有趣”的问题时,会扮演关键角色。让我们看看下面两个例子。

## 最小可用数

理查德·伯德提出过一个问题:找出不在一个列表中出现的最小数字([1] 第一章)。我们经常用数字标识实体,如身份证号、银行账户、电话号码等。我们希望找到一个最小的没有被占用数字。假设都是自然数,所有正在使用的数字记录在一个列表中:  
[18, 4, 8, 9, 16, 1, 14, 7, 19, 3, 0, 5, 2, 11, 6]

不在这个列表中的最小自然数是 10。这个题目看上去是如此简单,我们可以立即写出下面的解法:

```
1: function MIN-FREE( $A$ )
2:    $x \leftarrow 0$ 
3:   loop
4:     if  $x \notin A$  then
5:       return  $x$ 
6:     else
7:        $x \leftarrow x + 1$ 
```

其中符号  $\notin$  的实现如下:

```
1: function ' $\notin$ '( $x, X$ )
2:   for  $i \leftarrow 1$  to  $|X|$  do
3:     if  $x = X[i]$  then
4:       return False
5:   return True
```

其中  $|X| = n$  表示序列  $X$  的长度。有些编程语言内置了这一线性查找的实现。当列表包含几百万个数字时,这个方法的性能很快变差。它消耗的时间和  $n$  的平方

成正比。在一台双核 2.10GHz 处理器, 2G 内存的计算机上, 其 C 语言实现需要 5.4 秒才能在十万个数字中找到答案。当数量上升到一百万时, 则需要 8 分钟。

## 改进

对  $n$  个自然数  $x_1, x_2, \dots, x_n$ , 如果存在小于  $n$  的可用数, 必然存在某个  $x_i$  不在区间  $[0, n)$ <sup>2</sup> 内。否则这些数一定是  $0, 1, \dots, n-1$  的某个排列, 这种情况下, 最小的可用自然数是  $n$ 。

$$\text{minfree}(x_1, x_2, \dots, x_n) \leq n \quad (1)$$

我们用一个长度为  $n+1$  的数组  $F$ , 来标记区间  $[0, n]$  内的某个整数是否可用。

```

1: function MIN-FREE( $A$ )
2:    $F \leftarrow [\text{False}, \text{False}, \dots, \text{False}]$  ▷  $n+1$  个
3:   for  $x$  in  $A$  do
4:     if  $x < n$  then
5:        $F[x] \leftarrow \text{True}$ 
6:   for  $i \leftarrow 0$  to  $n$  do
7:     if  $F[i] = \text{False}$  then
8:       return  $i$ 

```

$F$  中的元素都初始化为假。遍历  $A$  中的数字, 如果  $x < n$ , 就将标志  $F[x]$  置为真。最后再扫描  $F$ , 找到第一个值为假的位置。算法用时和  $n$  成正比。我们使用了  $n+1$  个而不是  $n$  个标志, 用以处理  $\text{sort}(A) = [0, 1, 2, \dots, n-1]$  的情况。这个方法需要  $O(n)$  的空间来存储标志  $F$ 。查找结束后,  $F$  又被释放了。反复的申请和释放也会消耗时间。我们可以预先准备好足够长的数组, 每次查找复用。另外可以使用二进制位保存标志节约空间。在相同的计算机上, 相应的 C 语言程序仅用 0.023 秒就可以处理一百万个数字。

## 分而治之

分而治之的策略将问题分解为若干规模较小的子问题, 然后逐步解决以得到最终的结果。将所有满足  $x_i \leq \lfloor n/2 \rfloor$  的自然数放入子序列  $A'$ ; 其它放入子序列  $A''$ 。根据式 (1), 如果序列  $A'$  的长度正好是  $\lfloor n/2 \rfloor$ , 说明前半  $A'$  已经“满了”, 答案一定在  $A''$  中, 否则答案在  $A'$  中。通过这一划分, 问题的规模减小了。在子序列  $A''$  中递归查找时, 边界情况发生了变化。查找下界从 0 变成了  $\lfloor n/2 \rfloor + 1$ 。我们将定义修改为  $\text{search}(A, l, u)$ , 其中  $l$  是下界,  $u$  是上界。起始时:  $\text{minfree}(A) = \text{search}(A, l = 0, u = |A| - 1)$ 。

<sup>2</sup>左闭右开区间  $[a, b)$  包括  $a$  但不包括  $b$ 。

$$\begin{aligned} \text{search}(\emptyset, l, u) &= l \\ \text{search}(A, l, u) &= \begin{cases} |A'| = m - l + 1 : & \text{search}(A'', m + 1, u) \\ \text{否则} : & \text{search}(A', l, m) \end{cases} \end{aligned}$$

其中：

$$\begin{cases} m = \lfloor \frac{l+u}{2} \rfloor \\ A' = [x \leq m, x \in A], A'' = [x > m, x \in A] \end{cases}$$

这一方法不需要额外空间<sup>3</sup>。每次需要  $O(|A|)$  次比较划分  $A'$  和  $A''$ ，使问题规模减半。算法用时为  $T(n) = T(n/2) + O(n)$ ，通过主定理得到结果  $O(n)$ <sup>4</sup>。下面的例子代码实现了这一算法。

```
minFree xs = bsearch xs 0 (length xs - 1)

bsearch xs l u | xs == [] = l
               | length as == m - l + 1 = bsearch bs (m + 1) u
               | otherwise = bsearch as l m

where
  m = (l + u) `div` 2
  as = [x | x <- xs, x <= m]
  bs = [x | x <- xs, x > m]
```

递归的深度和调用栈都是  $O(\lg n)$ 。我们可以消除递归用迭代实现：

```
1: function MIN-FREE(A)
2:    $l \leftarrow 0, u \leftarrow |A|$ 
3:   while  $u - l > 0$  do
4:      $m \leftarrow \lfloor \frac{u+l}{2} \rfloor$ 
5:      $left \leftarrow l$ 
6:     for  $right \leftarrow l$  to  $u - 1$  do
7:       if  $A[right] \leq m$  then
8:         交换  $A[left] \leftrightarrow A[right]$ 
9:          $left \leftarrow left + 1$ 
10:    if  $left < m + 1$  then
11:       $u \leftarrow left$ 
12:    else
13:       $l \leftarrow left$ 
```

如图 1 所示，这段程序对数组中的元素进行划分。 $left$  之前的元素都不大于  $m$ ，而  $left$  和  $right$  之间的元素都大于  $m$ 。

<sup>3</sup>递归需要  $O(\lg n)$  的栈空间，但可以通过尾递归优化消除。

<sup>4</sup>我们也可以这样分析：第一次需要  $O(n)$  次比较划分，第二次需要比较  $O(n/2)$  次，第三次需要比较  $O(n/4)$  次……总时间为  $O(n + n/2 + n/4 + \dots) = O(2n) = O(n)$ 。

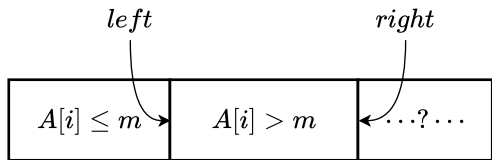


图 1: 数组划分。位于  $0 \leq i < left$  的元素满足  $A[i] \leq m$ , 位于  $left \leq i < right$  的元素满足  $A[i] > m$ , 剩余的元素尚未处理。

## 正规数

第二道趣题是如何寻找第 1500 个正规数。正规数是只含有 2、3、5 这三种因子的自然数<sup>5</sup>。2、3、5 本身也是正规数。60 = 2<sup>2</sup>3<sup>1</sup>5<sup>1</sup> 是第 25 个正规数。数字 21 = 2<sup>0</sup>3<sup>1</sup>7<sup>1</sup> 含有因子 7, 所以不是正规数。定义 1 = 2<sup>0</sup>3<sup>0</sup>5<sup>0</sup> 是第 0 个正规数。前 10 个正规数如下:

1, 2, 3, 4, 5, 6, 8, 9, 10, 12, ...

## 穷举法

我们可以从 1 开始, 逐一检查所有自然数, 对每个数把 2、3、5 这些因子不断去掉, 检查最终结果是否为 1:

```

1: function REGULAR-NUMBER(n)
2:   x ← 1
3:   while n > 0 do
4:     x ← x + 1
5:     if VALID?(x) then
6:       n ← n - 1
7:   return x

8: function VALID?(x)
9:   while x mod 2 = 0 do
10:    x ← x/2
11:  while x mod 3 = 0 do
12:    x ← x/3
13:  while x mod 5 = 0 do
14:    x ← x/5
15:  return x = 1 ?

```

随着  $n$  增大, 穷举法用时迅速增加。在同样的计算机上, 其 C 语言实现需要 40.39 秒才找到第 1500 个正规数(860934420)。

<sup>5</sup>在数论中叫作 5-光滑数。在计算机科学中又叫作哈明数以纪念理查德·哈明。

## 构造性解法

取模和除法是耗时的计算<sup>[2]</sup>。如果不再检查每一个数的因子,而是从 2、3、5 构造正规数,这样问题就转换为如何从小到大依次产生正规数序列。我们可以使用队列,从一侧加入,另一侧取出元素。先放入的先被取出,称为先进先出 FIFO(First In First Out)。把第 0 个正规数 1 加入队列,不断取出正规数,分别乘以 2、3、5,产生 3 个正规数并按大小顺序加入队列。丢弃重复的数,如图 2 所示。

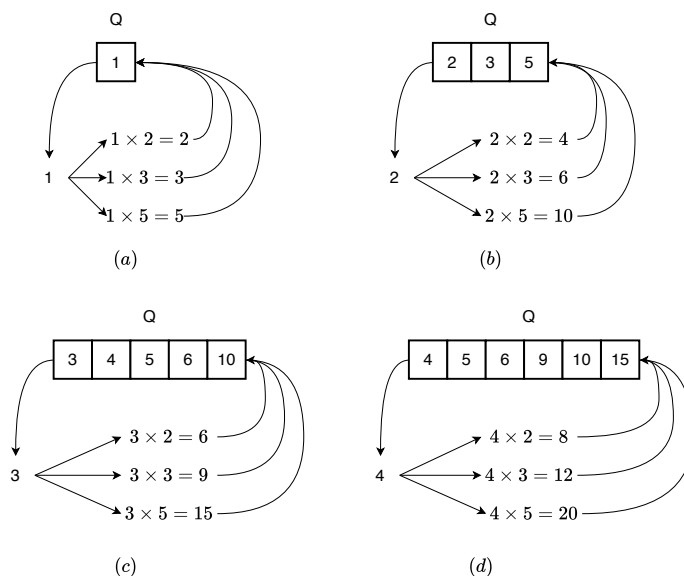


图 2: 前 4 步

```

1: function REGULAR-NUMBER( $n$ )
2:    $Q \leftarrow [1]$ 
3:   while  $n > 0$  do
4:      $x \leftarrow$  DEQUEUE( $Q$ )
5:     UNIQUE-ENQUEUE( $Q, 2x$ )
6:     UNIQUE-ENQUEUE( $Q, 3x$ )
7:     UNIQUE-ENQUEUE( $Q, 5x$ )
8:      $n \leftarrow n - 1$ 
9:   return  $x$ 

10: function UNIQUE-ENQUEUE( $Q, x$ )
11:    $i \leftarrow 0, m \leftarrow |Q|$ 
12:   while  $i < m$  且  $Q[i] < x$  do
13:      $i \leftarrow i + 1$ 
14:   if  $i \geq m$  或  $x \neq Q[i]$  then

```

15: INSERT( $Q, i, x$ )

对于长度为  $m$  的队列, UNIQUE-ENQUEUE 需要  $O(m)$  时间按序、无重复地插入新元素。队列的长度随着  $n$  线性增加 (每取出一个后最多插入三个新元素, 增加的比率  $\leq 2$ ), 总运行时间为  $O(1 + 2 + 3 + \dots + n) = O(n^2)$ 。图 3 显示了队列的访问次数和  $n$  之间的关系, 形状为二次曲线, 反映出  $O(n^2)$  的复杂度。

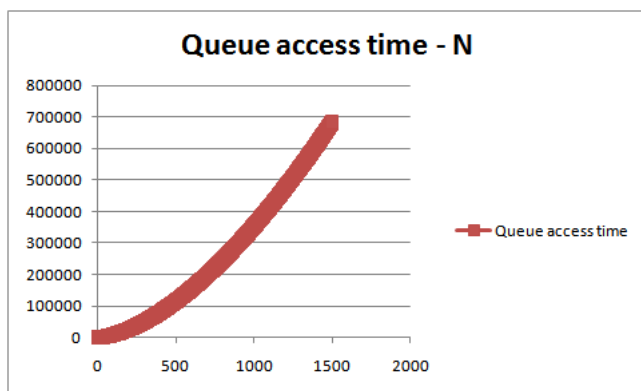


图 3: 队列访问次数和  $n$  的关系

在同样的计算机上, 其 C 语言实现仅用 0.016 秒就输出了答案, 比穷举法快 2500 倍。我们也可以用递归实现, 令  $xs$  为包含所有正规数的无穷序列  $[x_1, x_2, x_3, \dots]$ 。对每个正规数, 将其乘以 2 得到的仍然是无穷正规数列:  $[2x_1, 2x_2, 2x_3, \dots]$ 。同样依次乘以 3、5 会得到另外两个无穷正规数列。如果将这 3 个无穷数列合并, 去除重复并将 1 添加到开头, 就又得到了  $xs$ :

$$xs = 1 : [2x|x \leftarrow xs] \cup [3x|x \leftarrow xs] \cup [5x|x \leftarrow xs] \quad (2)$$

其中  $x:xs$  表示将元素  $x$  连接到列表  $xs$  的前面, 在 Lisp 中这个操作称为 cons。1 是第 0 个正规数, 在最前面。 $\cup$  合并两个列表。

$$(a:as) \cup (b:bs) = \begin{cases} a < b: & a : as \cup (b:bs) \\ a = b: & a : as \cup bs \\ a > b: & b : (a:as) \cup bs \end{cases}$$

由于  $(:)$  运算的优先级低于  $\cup$ , 我们省略了很多括弧, 例如:  $a : (as \cup (b:bs))$  可以简化为  $a : as \cup (b:bs)$ 。下面是例子程序

```
xs = 1 : [2*x | x ← xs] `merge` [3*x | x ← xs] `merge` [5*x | x ← xs]

merge (a:as) (b:bs) | a < b = a : merge as (b:bs)
                  | a = b = a : merge as bs
                  | otherwise = b : merge (a:as) bs
```

这一例子程序: `xs !! 1500`, 在同样的计算机上用 0.03 秒内得到第 1500 个正规数。

## 队列

上面的解法需要排除掉重复元素、扫描队列以保持有序。把所有正规数分成三个类： $Q_2 = \{2^i | i > 0\}$  仅包含被 2 整除的数； $Q_{23} = \{2^i 3^j | i \geq 0, j > 0\}$ 、 $Q_{235} = \{2^i 3^j 5^k | i, j \geq 0, k > 0\}$ 。其中  $Q_{23}$  要求  $j \neq 0$ ， $Q_{235}$  要求  $k \neq 0$ ，保证了三类数间没有重复。每类数都用一个队列来产生。初始化  $Q_2 = \{2\}$ ， $Q_{23} = \{3\}$  和  $Q_{235} = \{5\}$ 。每次取出三个队列头部的最小元素  $x$ ，然后进行下面的检查：

- 如果  $x$  来自  $Q_2$ ，将  $2x$  加入  $Q_2$ ， $3x$  加入  $Q_{23}$ ， $5x$  加入  $Q_{235}$ ；
- 如果  $x$  来自  $Q_{23}$ ，将  $3x$  加入  $Q_{23}$ ， $5x$  加入  $Q_{235}$ 。我们不应将  $2x$  加入  $Q_2$ ，因为  $Q_2$  中不包含被 3 整除的数。
- 如果  $x$  来自  $Q_{235}$ ，将  $5x$  加入  $Q_{235}$ 。我们不应将  $2x$  加入  $Q_2$ ， $3x$  加入  $Q_{23}$ ，因为它们不包含被 5 整除的数。

不断从三个队列中取出最小的  $n$  个元素。图 4 给出了前 4 步。

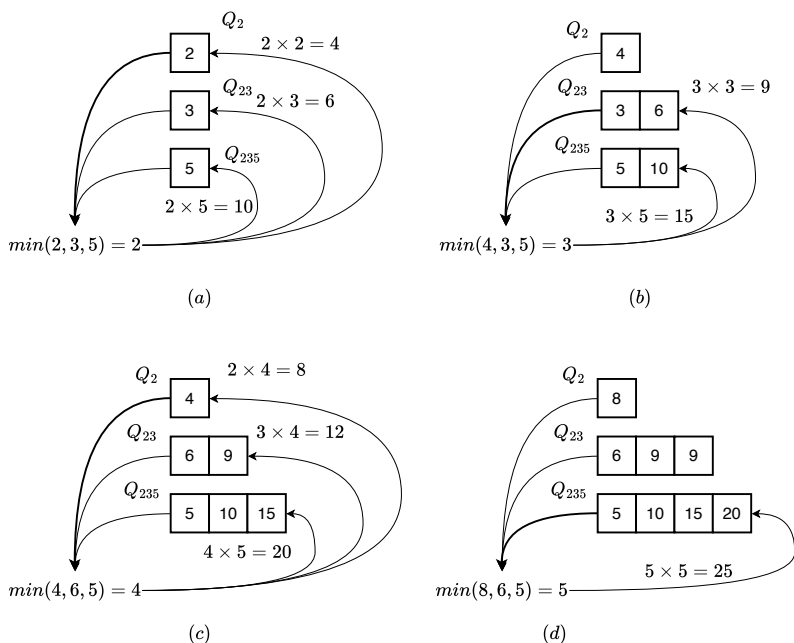


图 4: 使用三个队列  $Q_2$ 、 $Q_{23}$ 、 $Q_{235}$  构造正规数的前 4 步。

```

1: function REGULAR-NUMBER( $n$ )
2:    $x \leftarrow 1$ 
3:    $Q_2 \leftarrow \{2\}$ ,  $Q_{23} \leftarrow \{3\}$ ,  $Q_{235} \leftarrow \{5\}$ 
4:   while  $n > 0$  do
5:      $x \leftarrow \min(\text{HEAD}(Q_2), \text{HEAD}(Q_{23}), \text{HEAD}(Q_{235}))$ 
6:     if  $x = \text{HEAD}(Q_2)$  then

```

```

7:         DEQUEUE( $Q_2$ )
8:         ENQUEUE( $Q_2, 2x$ )
9:         ENQUEUE( $Q_{23}, 3x$ )
10:        ENQUEUE( $Q_{235}, 5x$ )
11:        else if  $x = \text{HEAD}(Q_{23})$  then
12:            DEQUEUE( $Q_{23}$ )
13:            ENQUEUE( $Q_{23}, 3x$ )
14:            ENQUEUE( $Q_{235}, 5x$ )
15:        else
16:            DEQUEUE( $Q_{235}$ )
17:            ENQUEUE( $Q_{235}, 5x$ )
18:         $n \leftarrow n - 1$ 
19:        return  $x$ 

```

算法循环  $n$  次。每次从三个队列中取出最小元素,这一步需要常数时间。接着根据取出元素所在的队列,产生一到三个新数加入队列,这一步也是常数时间。整个算法复杂度为  $O(n)$ 。

## 小结

尽管两道趣题都能用穷举法解决,但随着规模增加,我们不得不寻求更好的解法。本书**不是**一本算法竞赛或面试的“刷题手册”。本书介绍常见的基本算法和数据结构,对比给出纯函数式和命令式实现,展示函数式算法和数据结构的特有思路。主要参考了冈崎的著作<sup>[3]</sup>和经典的算法教材<sup>[4]</sup>。本书尽量避免依赖于特定的编程语言。一方面读者会有自己的偏好,另一方面编程语言也在不断变化。我们主要使用伪代码和数学记法对算法进行定义,并附以一些例子代码。函数式的示例类似 Haskell,命令式的示例是几种语言的混合体。

## 练习 1

- 1.1. 最小可用数趣题中,所有数都是非负整数。我们可以利用正负号来标记一个数字是否存在。首先扫描一遍列表,若  $|x| < n$ , 其中  $n$  为列表长度,将位置  $|x|$  上的数字置为负数。之后再扫描一遍列表,找到第一个正数所在的位置就是答案。编程实现这一算法。
- 1.2.  $n$  个数字  $1, 2, \dots, n$ , 经过某一处理后,它们的顺序被打乱了,并且某一个数  $x$  被改成了  $y$ 。假设  $1 \leq y \leq n$ , 设计一个方法能够在线性时间、常数空间内找出  $x$  和  $y$ 。
- 1.3. 下面是一段求正规数的代码。它是一种队列解法么?

```
Int regularNum(Int m) {
```



```
[Int] nums(m + 1)
Int n = 0, i = 0, j = 0, k = 0
nums[0] = 1
Int x2 = 2 * nums[i]
Int x3 = 3 * nums[j]
Int x5 = 5 * nums[k]
while n < m {
    n = n + 1
    nums[n] = min(x2, x3, x5)
    if x2 == nums[n] {
        i = i + 1
        x2 = 2 * nums[i]
    }
    if x3 == nums[n] {
        j = j + 1
        x3 = 3 * nums[j]
    }
    if x5 == nums[n] {
        k = k + 1
        x5 = 5 * nums[k]
    }
}
return nums[m]
}
```

答案: 345 页

# 目录

第二版前言	i
第一版前言	iii
<b>第一章 列表</b>	<b>1</b>
1.1 简介	1
1.2 定义	1
1.2.1 分解	2
1.2.2 列表的基本操作	2
1.2.3 索引	2
1.2.4 末尾元素	3
1.2.5 反向索引	4
1.2.6 更改	5
插入	6
删除	8
连接	9
1.2.7 和与积	10
1.2.8 最大值和最小值	13
1.3 变换	14
1.3.1 逐一映射	14
1.3.2 反转	18
1.4 子列表	19
1.4.1 切分和分组	20
1.5 叠加	23
1.6 查找和过滤	26
1.7 关联和分解	29
<b>第二章 二叉搜索树</b>	<b>33</b>
2.1 定义	33
2.2 插入	34

2.3	遍历	36
2.4	查找	38
2.5	删除	40
2.6	附录:例子代码	43
<b>第三章</b>	<b>插入排序</b>	<b>47</b>
3.1	简介	47
3.2	插入	48
3.3	二分查找	49
3.4	列表	50
3.5	二叉搜索树	51
<b>第四章</b>	<b>红黑树</b>	<b>53</b>
4.1	平衡	54
4.2	定义	56
4.3	插入	57
4.4	删除	59
4.5	命令式算法★	64
4.6	附录:例子程序	65
<b>第五章</b>	<b>AVL 树</b>	<b>69</b>
5.1	定义	69
5.2	插入	71
5.2.1	平衡调整	72
5.2.2	检验	74
5.3	命令式算法★	74
5.4	附录:例子程序	77
<b>第六章</b>	<b>基数树</b>	<b>79</b>
6.1	整数 trie	79
6.1.1	定义	80
6.1.2	插入	80
6.1.3	查找	82
6.2	整数前缀树	83
6.2.1	定义	83
6.2.2	插入	83
6.2.3	查找	88
6.3	Trie	88
6.3.1	插入	89

6.3.2	查找	90
6.4	前缀树	91
6.4.1	插入	91
6.4.2	查找	95
6.5	Trie 和前缀树的应用	95
6.5.1	词典和自动补齐	95
6.5.2	数字键盘输入法	98
6.6	附录:例子程序	101
<b>第七章</b>	<b>B 树</b>	<b>107</b>
7.1	简介	107
7.2	插入	109
7.2.1	先插入再分拆	109
7.2.2	先分拆再插入	113
7.2.3	列表对	114
7.3	查找	118
7.4	删除	119
7.4.1	先删除再修复	119
7.4.2	先合并再删除	123
7.5	小结	126
7.6	附录:例子程序	127
<b>第八章</b>	<b>二叉堆</b>	<b>131</b>
8.1	定义	131
8.2	用数组实现的二叉堆	131
8.2.1	堆调整	132
8.2.2	构造堆	133
8.2.3	堆的基本操作	134
8.2.4	堆排序	137
8.3	左偏堆和斜堆	138
8.3.1	左偏堆	138
8.3.2	斜堆	141
8.4	伸展堆	141
8.4.1	伸展操作	142
8.4.2	弹出顶部	145
8.4.3	合并	146
8.5	小结	146
8.6	附录:例子程序	146

<b>第九章 选择排序</b>	<b>151</b>
9.1 查找最小元素	152
9.1.1 选择排序的性能	153
9.2 改进	153
9.2.1 鸡尾酒排序	154
9.3 继续改进	157
9.3.1 锦标赛淘汰法	157
9.3.2 改进为堆排序	161
9.4 附录:例子程序	161
<b>第十章 二项式堆、斐波那契堆、配对堆</b>	<b>165</b>
10.1 简介	165
10.2 二项式堆	165
10.2.1 二项式树	165
10.2.2 树的链接	168
10.2.3 插入	169
10.2.4 堆合并	170
10.2.5 弹出	171
10.3 斐波那契堆	173
10.3.1 插入	174
10.3.2 合并	174
10.3.3 弹出	175
10.3.4 提升优先级	180
10.3.5 斐波那契堆的命名	181
10.4 配对堆	183
10.4.1 定义	183
10.4.2 合并、插入、获取堆顶	183
10.4.3 提升优先级	184
10.4.4 弹出	184
10.4.5 删除	187
10.5 附录:例子程序	187
<b>第十一章 队列</b>	<b>193</b>
11.1 列表实现	193
11.2 循环缓冲区	194
11.3 双列表队列	196
11.4 平衡队列	197
11.5 实时队列	198
11.6 惰性实时队列	201

11.7 附录:例子程序	202
<b>第十二章 序列</b>	<b>205</b>
12.1 二叉随机访问列表	205
12.2 数字表示	209
12.3 双数组序列	211
12.4 可连接列表	212
12.5 手指树	215
12.5.1 插入	215
12.5.2 删除	217
12.5.3 尾部操作	217
12.5.4 连接	219
12.5.5 随机访问	220
12.6 附录:例子程序	222
<b>第十三章 快速排序和归并排序</b>	<b>227</b>
13.1 快速排序	227
13.1.1 划分	228
13.1.2 原地排序	229
13.1.3 性能分析	231
平均情况 ★	232
13.1.4 改进	234
特殊情况	238
13.1.5 快速排序与树排序	241
13.2 归并排序	242
13.2.1 性能分析	244
13.2.2 改进	245
13.2.3 原地归并排序	246
13.2.4 自然归并排序	249
13.2.5 自底向上归并排序	253
13.3 并行处理	254
13.4 小结	254
13.5 附录:例子程序	255
<b>第十四章 搜索 I</b>	<b>259</b>
14.1 $k$ 选择问题	259
14.2 二分查找	260
14.3 二维查找	262
14.4 众数问题	270

14.5 最大子序列和 . . . . .	272
14.6 字符串搜索 . . . . .	274
14.7 附录:例子程序 . . . . .	276
<b>第十五章 搜索 II</b>	<b>279</b>
15.1 深度优先和广度优先搜索 . . . . .	279
15.1.1 迷宫 . . . . .	279
15.1.2 八皇后问题 . . . . .	282
15.1.3 跳棋趣题 . . . . .	284
15.1.4 狼、羊、白菜过河问题 . . . . .	288
15.1.5 倒水问题 . . . . .	291
15.1.6 华容道 . . . . .	295
15.2 贪心算法 . . . . .	301
15.2.1 哈夫曼编码 . . . . .	301
15.2.2 换零钱问题 . . . . .	305
15.3 动态规划 . . . . .	306
15.3.1 最长公共子序列 . . . . .	308
15.3.2 子集和问题 . . . . .	310
15.4 附录:例子程序 . . . . .	314
<b>附录 A 红黑树的命令式删除算法</b>	<b>323</b>
<b>附录 B AVL 树——证明和删除算法</b>	<b>333</b>
B.1 插入后的高度变化 . . . . .	333
B.2 插入后的平衡调整 . . . . .	334
B.3 删除算法 . . . . .	337
B.3.1 函数式删除 . . . . .	337
B.3.2 命令式删除 . . . . .	339
B.4 例子程序 . . . . .	342
<b>附录 C 参考答案</b>	<b>345</b>
<b>参考文献</b>	<b>407</b>
<b>索引</b>	<b>417</b>
<b>GNU Free Documentation License</b>	<b>421</b>
1. APPLICABILITY AND DEFINITIONS . . . . .	421
2. VERBATIM COPYING . . . . .	423
3. COPYING IN QUANTITY . . . . .	423

4. MODIFICATIONS . . . . . 424

5. COMBINING DOCUMENTS . . . . . 426

6. COLLECTIONS OF DOCUMENTS . . . . . 427

7. AGGREGATION WITH INDEPENDENT WORKS . . . . . 427

8. TRANSLATION . . . . . 427

9. TERMINATION . . . . . 428

10. FUTURE REVISIONS OF THIS LICENSE . . . . . 428

11. RELICENSING . . . . . 429

ADDENDUM: How to use this License for your documents . . . . . 429



# 第一章 列表

## 1.1 简介

列表和数组是构建其它复杂数据结构的基石。它们都是可容纳若干元素的容器。数组通常是一组连续的存储区域,每个存储单元由一个数字索引。这个数字叫作地址或者位置。数组的大小通常需要在使用前确定,而列表则无此限制,可以随时加入元素。我们可以从头到尾依次遍历列表中的元素。在函数式环境中,列表算法对计算和逻辑控制起着关键作用<sup>1</sup>。对于已经熟悉映射 (map),过滤 (filter),叠加 (fold) 等算法的读者,可以跳过这一章,直接从第二章开始阅读。

## 1.2 定义

列表又称单向链表,是一种递归的数据结构。一个列表或者为空,记为 [] 或 NIL;或者包含一个元素(也称作 key)和一个子列表。图 1.1 描述了一个由若干节点组成的列表。指向子列表的引用通常叫作 next。每个节点要么链接到下一个节点上,要么指向 NIL。通常使用复合数据结构<sup>2</sup>定义列表,例如:

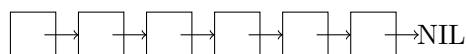


图 1.1: 由节点组成的列表

```
data List<A> {  
  A key  
  List<A> next  
}
```

很多编程环境支持空引用 NIL 概念,存在两种不同的方法表示空列表:一种直接使用 NIL(或 null 或  $\emptyset$ );另一种创建一个列表,但不填入任何元素,表示为 []。在实现上,空引用不占用内存,但 [] 则需要分配内存。

<sup>1</sup>在更底层,lambda 演算作为和图灵机等价的计算模型更为基础 [93], [99]。

<sup>2</sup>多数情况下,列表中元素有着共同的类型。有些环境(如 Lisp)支持包含不同数据类型的列表。

### 1.2.1 分解

给定一个非空列表  $X$ , 定义两个函数<sup>3</sup> 来获取头部元素和子列表。它们通常叫做  $first X$  和  $rest X$ , 或者  $head X$  和  $tail X$ <sup>4</sup>。反之, 我们可以从一个元素  $x$  和列表  $xs$  (可为空) 构造出另一个列表, 记为  $x:xs$ 。这一构造过程叫作  $cons$ 。我们有如下关系:

$$\begin{cases} head(x:xs) = x \\ tail(x:xs) = xs \end{cases} \quad (1.1)$$

我们也用  $x_1$  表示第一个元素, 用  $X'$  表示剩余列表, 例如  $X = [x_1, x_2, x_3, \dots]$ ,  $X' = [x_2, x_3, \dots]$ 。

### 练习 1.2

1.2.1. 对于元素类型为  $A$  的列表, 如果能够判断任何两个元素  $x, y \in A$  是否相等, 定义一个算法来判断两个列表是否相等。

答案: 348 页

### 1.2.2 列表的基本操作

根据定义, 我们可以递归地计算列表的长度: 空列表的长度为 0, 而非空列表的长度是除去第一个元素的子列表长度加一。

$$\begin{aligned} length [] &= 0 \\ length(x:xs) &= 1 + length xs \end{aligned} \quad (1.2)$$

因为遍历列表计算长度, 时间复杂度是  $O(n)$ , 其中  $n$  是元素个数。在不和绝对值混淆的情况下, 我们使用  $|X|$  来表示列表  $X$  的长度。为了避免反复计数, 我们可以将长度存储在一个变量中, 并在增加或删除元素时更新这一变量。下面是计算长度的迭代实现:

```

1: function LENGTH(X)
2:   n ← 0
3:   while X ≠ NIL do
4:     n ← n + 1
5:     X ← NEXT(X)
6:   return n

```

### 1.2.3 索引

数组支持以常数时间随机访问任意位置  $i$  的元素。但列表需要前进  $i$  步才能到达元素所在位置。

<sup>3</sup>本书通常将函数调用  $f(x)$  记为  $f x$ , 多元函数  $f(x, y, \dots, z)$  记为  $f x y \dots z$ 。

<sup>4</sup>在 Lisp 中, 由于历史原因, 它们被命名为  $car$  和  $cdr$  用以代表当时机器中的寄存器<sup>[63]</sup>



```

10:     X' ← CONS(FIRST(X), X')
11:     X ← REST(X)
12:     return REVERSE(X')

```

INIT 通过 CONS 累积结果。这样产生的列表是逆序的,最后需要将结果再倒转过来(见 1.3.2)。

### 1.2.5 反向索引

*last* 是反向索引的一种特例。更一般的形式是获取列表中的倒数第  $i$  个元素。最直接的思路是遍历两次:第一次获取列表长度  $n$ ,第二次获取第  $n - i - 1$  个元素:

$$\text{lastAt } i \ X = \text{getAt } (|X| - i - 1) \ L$$

更好的解法是使用两个指针  $p_1$  和  $p_2$ , 它们相距  $i$  步, 即  $\text{rest}^i(p_2) = p_1$ , 其中  $\text{rest}^i(p_2)$  表示重复执行函数 *rest* 共  $i$  次。也就是说,从  $p_2$  前进  $i$  步就可到达  $p_1$ 。 $p_2$  一开始指向链表的头部,然后同时向前移动它们,直到  $p_1$  到达链表的尾部。此时指针  $p_2$  恰好指向倒数第  $i$  个元素。图 1.2 描述了这一方法。由于  $p_1, p_2$  框出一个窗口,这一方法也称作滑动窗口法。

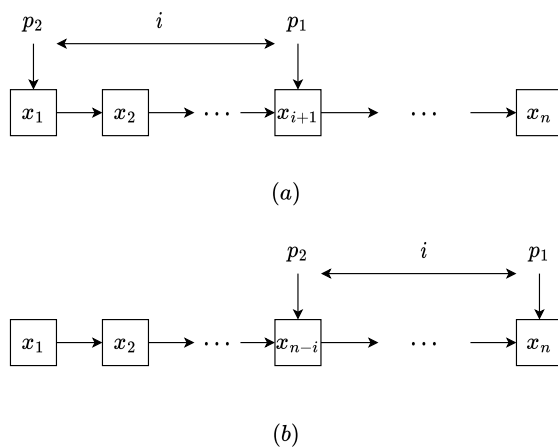


图 1.2: 双指针滑动窗口。(a)  $p_2$  指向表头,距离  $p_1$  之后  $i$  步;(b) 当  $p_1$  到达表尾时, $p_2$  指向右数第  $i$  个。

```

1: function LAST-AT( $i, X$ )
2:    $p \leftarrow X$ 
3:   while  $i > 0$  do
4:      $X \leftarrow \text{REST}(X)$ 
5:      $i \leftarrow i - 1$ 
6:   while  $\text{REST}(X) \neq \text{NIL}$  do

```

▷ 越界时出错

```

7:      X ← REST(X)
8:      p ← REST(p)
9:      return FIRST(p)

```

纯函数实现时不能直接更新指针,为此我们可以同时遍历  $X = [x_1, x_2, \dots, x_n]$  和  $Y = [x_i, x_{i+1}, \dots, x_n]$ , 其中  $Y$  是除去前  $i - 1$  个元素后的子列表。

$$\text{lastAt } i \ X = \text{slide } X \ (\text{drop } i \ X) \quad (1.5)$$

其中:

$$\begin{aligned} \text{slide } (x:xs) \ [y] &= x \\ \text{slide } (x:xs) \ (y:ys) &= \text{slide } xs \ ys \end{aligned} \quad (1.6)$$

函数  $\text{drop } m \ X$  丢弃前  $m$  个元素:

$$\begin{aligned} \text{drop } 0 \ xs &= xs \\ \text{drop } m \ [] &= [] \\ \text{drop } m \ (x:xs) &= \text{drop } (m - 1) \ xs \end{aligned} \quad (1.7)$$

## 练习 1.4

1.4.1. 在 INIT 中,可以用  $\text{APPEND}(X', \text{FIRST}(X))$  来替换 CONS 么?

1.4.2. 在 LAST-AT 中,如何处理空列表和越界的情况?

答案: 348 页

### 1.2.6 更改

更改操作包括添加、插入、更新、删除。某些函数式环境在实现时创建新列表,而原列表保持 (persist) 不变,并在适当的时候释放原始列表 ([3], 第 2 章)。添加称为 `append`, 和 `cons` 对称, `cons` 在表头增加, `append` 在末尾增加。因此也被称作 `snoc` (将 `cons` 反过来拼写)。由于要遍历到列表尾部,其复杂度为  $O(n)$ 。为了避免反复遍历,我们可以将尾部位置存储下来,并随着列表变化进行更新。

$$\begin{aligned} \text{append } [] \ x &= [x] \\ \text{append } (y:ys) \ x &= y : \text{append } ys \ x \end{aligned} \quad (1.8)$$

对应的迭代实现如下<sup>5</sup>:

```

1: function APPEND(X, x)
2:   if X = NIL then
3:     return CONS(x, NIL)

```

<sup>5</sup>参数的顺序也是对称的: `cons x xs, append xs x`。

```

4:  H ← X
5:  while REST(X) ≠ NIL do
6:    X ← REST(X)
7:    REST(X) ← CONS(x, NIL)
8:  return H

```

▷ 保存表头

更新 REST 的过程通常实现为对 next 引用的改写, 如下面的例子代码:

```

List<A> append(List<A> xs, A x) {
  if xs == null then return cons(x, null)
  var head = xs
  while xs.next ≠ null {
    xs = xs.next
  }
  xs.next = cons(x, null)
  return head
}

```

和 *getAt* 类似, 我们需要移动到列表中的指定位置以修改元素。时间复杂度为  $O(i)$ , 其中  $i$  是要修改的位置。

$$\begin{aligned}
 \text{setAt } 0 \ x \ (y:ys) &= x : ys \\
 \text{setAt } i \ x \ (y:ys) &= y : \text{setAt } (i - 1) \ x \ ys
 \end{aligned}
 \tag{1.9}$$

## 练习 1.5

- 1.5.1. 在列表的定义中增加一个尾部变量 *tail*, 将添加算法优化为常数时间。
- 1.5.2. 何时应该更新 *tail* 变量? 对性能有何影响?
- 1.5.3. 在 *setAt* 中, 如何处理空列表和越界的情况?

答案: 349 页

### 插入

列表插入有两种不同的含义:(1)在指定位置插入一个元素, 记为 *insert i x X*, 其实现和 *setAt* 类似;(2)在已序列表中按序插入一个元素。

$$\begin{aligned}
 \text{insert } 0 \ x \ ys &= x : ys \\
 \text{insert } i \ x \ (y:ys) &= y : \text{insert } (i - 1) \ x \ ys
 \end{aligned}
 \tag{1.10}$$

当  $i$  超过列表的长度时, 我们可以将其视作添加, 见本节习题。下面是相应的迭代实现:

```

1: function INSERT(i, x, X)
2:   if i = 0 then
3:     return CONS(x, X)

```

```

4:    $H \leftarrow X$ 
5:    $p \leftarrow X$ 
6:   while  $i > 0$  and  $X \neq \text{NIL}$  do
7:      $p \leftarrow X$ 
8:      $X \leftarrow \text{REST}(X)$ 
9:      $i \leftarrow i - 1$ 
10:   $\text{REST}(p) \leftarrow \text{CONS}(x, X)$ 
11:  return  $H$ 

```

令列表  $X = [x_1, x_2, \dots, x_n]$  已序, 即对任何位置  $1 \leq i \leq j \leq n$ , 有  $x_i \leq x_j$ 。这里  $\leq$  的含义是抽象的, 它可以代表任何有序的比较, 包括  $\geq$  (降序)、集合的包含关系等。我们定义有序插入, 使得列表保持有序。

$$\begin{aligned}
 \text{insert } x \ [] &= [x] \\
 \text{insert } x \ (y:ys) &= \begin{cases} x \leq y : & x:y:ys \\ \text{否则} : & y : \text{insert } x \ ys \end{cases} \quad (1.11)
 \end{aligned}$$

由于要逐一比较元素, 插入的时间复杂度为  $O(n)$ 。对应的迭代实现如下:

```

1: function INSERT( $x, X$ )
2:   if  $X = \text{NIL}$  或  $x < \text{FIRST}(X)$  then
3:     return  $\text{CONS}(x, X)$ 
4:    $H \leftarrow X$ 
5:   while  $\text{REST}(X) \neq \text{NIL}$  且  $\text{FIRST}(\text{REST}(X)) < x$  do
6:      $X \leftarrow \text{REST}(X)$ 
7:    $\text{REST}(X) \leftarrow \text{CONS}(x, \text{REST}(X))$ 
8:   return  $H$ 

```

我们可以进一步实现插入排序: 逐一将元素按序插入到列表中。由于每次插入都需要线性时间, 这一排序的复杂度为  $O(n^2)$ 。

$$\begin{aligned}
 \text{sort} \ [] &= [] \\
 \text{sort} \ (x:xs) &= \text{insert } x \ (\text{sort } xs) \quad (1.12)
 \end{aligned}$$

我们可以消除递归, 实现迭代算法: 逐一从列表中取出元素并按序插入到结果中:

```

1: function SORT( $X$ )
2:    $S \leftarrow \text{NIL}$ 
3:   while  $X \neq \text{NIL}$  do
4:      $S \leftarrow \text{INSERT}(\text{FIRST}(X), S)$ 
5:      $X \leftarrow \text{REST}(X)$ 
6:   return  $S$ 

```

在循环中的任何时刻,结果列表都是已序的。和递归实现相比,它们有一个本质不同:前者从右向左处理列表,而后者从左向右处理。我们稍后将在“尾递归”(1.2.7)中讲述如何消除这一差异。第3章详细介绍插入排序,包括性能分析和优化。

### 练习 1.6

1.6.1. 当插入位置越界时,将其按照添加来处理。

1.6.2. 针对数组实现插入算法,插入位置  $i$  后的所有元素需要向后移动一个位置。

答案: 349 页

### 删除

和插入类似,删除也有两种含义:(1)在指定位置删除元素  $delAt\ i\ X$ ; (2)查找某个值并删除  $delete\ x\ X$ 。为了删除位置  $i$  上的元素,先前进  $i$  步,然后跳过一个元素,将剩余部分连接起来。

$$\begin{aligned} delAt\ i\ [] &= [] \\ delAt\ 0\ (x:xs) &= xs \\ delAt\ i\ (x:xs) &= x : delAt\ (i - 1)\ xs \end{aligned} \tag{1.13}$$

由于需要前进  $i$  步,其时间复杂度为  $O(i)$ 。下面是相应的迭代实现:

```

1: function DEL-AT( $i, X$ )
2:    $S \leftarrow CONS(\perp, X)$  ▷ 辅助节点
3:    $p \leftarrow S$ 
4:   while  $i > 0$  且  $X \neq NIL$  do
5:      $i \leftarrow i - 1$ 
6:      $p \leftarrow X$ 
7:      $X \leftarrow REST(X)$ 
8:   if  $X \neq NIL$  then
9:      $REST(p) \leftarrow REST(X)$ 
10:  return  $REST(S)$ 

```

为了简化边界情况处理,我们引入辅助节点  $S$ ,它包含特殊值  $\perp$ ,并指向  $X$ 。我们可以安全地切除  $X$  中的任何节点,包括头节点。最后将  $S$  后继的部分返回,并丢弃  $S$ 。“查找并删除”又可以进一步细分为两种情况:(1)找到第一个元素并删除;(2)找到所有等于某值的元素全部删除。后者更加一般,见本节练习。

$$\begin{aligned} delete\ x\ [] &= [] \\ delete\ x\ (y:ys) &= \begin{cases} x = y : ys \\ x \neq y : y : delete\ x\ ys \end{cases} \end{aligned} \tag{1.14}$$



由于遍历查找待删除元素,其复杂度为  $O(n)$ 。在迭代实现中,我们同样使用辅助节点来简化处理:

```

1: function DELETE( $x, X$ )
2:    $S \leftarrow \text{CONS}(\perp, X)$ 
3:    $p \leftarrow X$ 
4:   while  $X \neq \text{NIL}$  且  $\text{FIRST}(X) \neq x$  do
5:      $p \leftarrow X$ 
6:      $X \leftarrow \text{REST}(X)$ 
7:   if  $X \neq \text{NIL}$  then
8:      $\text{REST}(p) \leftarrow \text{REST}(X)$ 
9:   return  $\text{REST}(S)$ 

```

### 练习 1.7

1.7.1. 设计算法将等于给定值的所有元素删除。

1.7.2. 设计数组的删除算法,被删除位置后的所有元素需要向前移动一个位置。

答案: 350 页

### 连接

连接是添加的更一般形式。添加向列表尾部加入一个元素,连接向尾部加入多个元素。但如果通过多次添加来实现,则整体性能会下降为平方级别。令  $|xs| = n$ ,  $|ys| = m$ , 分别是两个列表的长度。每次添加都需要前进到尾部,总共添加  $m$  次。总时间复杂度为  $O(n + (n + 1) + \dots + (n + m)) = O(nm + m^2)$ 。

$$\begin{aligned} xs \# [] &= xs \\ xs \# (y:ys) &= \text{append } xs \ y \# ys \end{aligned}$$

而 cons 的速度很快(常数时间),我们可以前进到  $xs$  的尾部再链接  $ys$ :

$$\begin{aligned} [] \# ys &= ys \\ xs \# [] &= xs \\ (x:xs) \# ys &= x : (xs \# ys) \end{aligned} \tag{1.15}$$

改进的复杂度为  $O(n)$ 。在命令式环境中,通过使用尾部引用,可实现常数时间的连接操作(见本节习题)。

```

1: function CONCAT( $X, Y$ )
2:   if  $X = \text{NIL}$  then
3:     return  $Y$ 
4:   if  $Y = \text{NIL}$  then
5:     return  $X$ 

```

```

6:  H ← X
7:  while REST(X) ≠ NIL do
8:    X ← REST(X)
9:  REST(X) ← Y
10: return H

```

### 1.2.7 和与积

我们常常需要计算列表的和与积。它们有着共同的计算结构。我们在 1.5 中对它们进行抽象。定义空列表的和为 0、积为 1。

$$\begin{aligned}
 \text{sum } [] &= 0 & \text{product } [] &= 1 \\
 \text{sum } (x:xs) &= x + \text{sum } xs & \text{product}(x:xs) &= x \cdot \text{product } xs
 \end{aligned} \tag{1.16}$$

两个计算都需要遍历列表,复杂度为  $O(n)$ 。计算顺序从右向左。我们可将其改成从左向右**累积计算**。

$$\begin{aligned}
 \text{sum}' a [] &= a & \text{prod}' a [] &= a \\
 \text{sum}' a (x:xs) &= \text{sum } (x + a) xs & \text{prod}' a (x:xs) &= \text{prod}' (x \cdot a) xs
 \end{aligned} \tag{1.17}$$

求和时从 0 开始累加。求积时从 1 开始累乘:

$$\text{sum } xs = \text{sum}' 0 xs \quad \text{product } xs = \text{prod}' 1 xs \tag{1.18}$$

或使用柯里化形式:

$$\text{sum} = \text{sum}' 0 \quad \text{product} = \text{prod}' 1$$

**柯里化**是肖芬格尔 (Schönfinkel, 1889 - 1942) 在 1924 年提出的。哈斯克爾·柯里在 1958 年后广泛使用<sup>[73]</sup>。考虑二元函数  $f(x, y)$ , 如果只传入  $x$ , 它就转换为一个关于  $y$  的一元函数:  $g(y) = f(x, y)$ , 记为  $g = f x$ 。推广到多元函数  $f(x, y, \dots, z)$ , 通过依次传入参数, 可以转换为一系列函数:  $f, f x, f x y, \dots$ 。我们称这样的转换为柯里化。它可以把多元函数转化为一系列一元函数, 即:  $f(x, y, \dots, z) = f(x)(y)\dots(z) = f x y \dots z$ 。

采用累积法后, 计算顺序变为从左向右, 并且无需记录中间结果、状态用于递归。所有状态或作为参数 (例如  $a$ ) 传入, 或丢弃不用 (例如已处理过的元素)。这样的递归可进一步转化为循环。我们称这样的函数为“尾递归”(或“尾调用”), 称这种消除递归的优化为“尾递归优化”<sup>[61]</sup>。顾名思义, 在这类函数中, 递归发生在计算尾部。尾递归优化可以极大地提高性能, 并避免由于递归过深造成的调用栈溢出。在式 (1.12) 关于插入排序的部分, 其递归实现从右向左对元素排序。我们也可以将其转化为尾递归:

$$\begin{aligned}
 \text{sort}' a [] &= a \\
 \text{sort}' a (x:xs) &= \text{sort}' (\text{insert } x a) xs
 \end{aligned} \tag{1.19}$$

传入  $[]$  启动排序 (柯里化形式):  $sort = sort' []$ 。作为尾递归的典型例子, 我们考虑如何高效地计算幂  $b^n$  ([63], 1.16 节)。最直接的方法是从 1 开始重复乘以  $b$  共  $n$  次, 这是一个  $O(n)$  时间的方法:

```

1: function POW(b, n)
2:   x ← 1
3:   loop n 次
4:     x ← x · b
5:   return x

```

考虑计算  $b^8$  的过程, 上述算法经过前两次迭代, 可以得到  $x = b^2$  的结果。此时, 我们无需用  $x$  乘以  $b$  得到  $b^3$ , 可以直接再次乘以  $b^2$  得到  $b^4$ 。然后再次乘方得到  $(b^4)^2 = b^8$ 。只要循环 4 次, 而不是 8 次。若  $n$  恰好为 2 的整数次幂, 即  $n = 2^m$ , 其中  $m$  是非负整数, 我们可以用下面的方法快速计算  $b^n$ :

$$\begin{aligned} b^1 &= b \\ b^n &= (b^{\frac{n}{2}})^2 \end{aligned}$$

我们将  $n$  推广到任意的非负整数: 若  $n = 0$ , 定义  $b^0 = 1$ ; 若  $n$  为偶数, 将  $n$  减半, 先计算  $b^{\frac{n}{2}}$ , 然后再将结果平方; 若  $n$  为奇数, 因为  $n - 1$  是偶数, 可以先递归计算  $b^{n-1}$ , 然后再将结果乘以  $b$ 。

$$\begin{aligned} b^0 &= 1 \\ b^n &= \begin{cases} 2|n: & (b^{\frac{n}{2}})^2 \\ \text{否则}: & b \cdot b^{n-1} \end{cases} \end{aligned} \quad (1.20)$$

但是  $(b^{\frac{n}{2}})^2$  无法直接转换为尾递归。为此, 我们可以先将底数平方, 然后再将指数减半。

$$\begin{aligned} b^0 &= 1 \\ b^n &= \begin{cases} 2|n: & (b^2)^{\frac{n}{2}} \\ \text{否则}: & b \cdot b^{n-1} \end{cases} \end{aligned} \quad (1.21)$$

经过这一修改, 就可以将算法转换为尾递归。我们通过等式  $b^n = pow(b, n, 1)$  计算幂。

$$\begin{aligned} pow(b, 0, a) &= a \\ pow(b, n, a) &= \begin{cases} 2|n: & pow(b^2, \frac{n}{2}, a) \\ \text{否则}: & pow(b, n-1, ab) \end{cases} \end{aligned} \quad (1.22)$$

这一方法的复杂度为  $O(\lg n)$ 。将  $n$  表示成二进制数  $n = (a_m a_{m-1} \dots a_1 a_0)_2$ , 如果  $a_i = 1$ , 我们清楚地知道, 需要计算  $b^{2^i}$ 。这和二项式堆的情况很类似 (10.2)。将所有二进制位为 1 的幂计算出, 再累乘到一起就得到最后的结果。例如, 当计算  $b^{11}$  时, 11 写

成二进制为  $11 = (1011)_2 = 2^3 + 2 + 1$ , 因此  $b^{11} = b^{2^3} \times b^2 \times b$ 。我们可以通过以下的步骤进行计算:

1. 计算  $b^1$ , 得  $b$ ;
2. 从这一结果进而得到  $b^2$ ;
3. 将第 2 步的结果平方, 从而得到  $b^{2^2}$ ;
4. 将第 3 步的结果平方, 得到  $b^{2^3}$ 。

最后, 我们将第 1、2、和第 4 步的结果乘到一起, 得到  $b^{11}$ 。综上, 我们可以改进如下:

$$\begin{aligned} \text{pow}(b, 0, a) &= a \\ \text{pow}(b, n, a) &= \begin{cases} 2|n: & \text{pow}(b^2, \frac{n}{2}, a) \\ \text{否则}: & \text{pow}(b^2, \lfloor \frac{n}{2} \rfloor, ab) \end{cases} \end{aligned} \quad (1.23)$$

这一算法本质上每次将  $n$  向右移动一个二进制位 (通过将  $n$  除以 2)。若最低位 (LSB: Least Significant Bit) 为 0,  $n$  为偶数。我们将底数平方, 继续递归, 无需改变累积结果  $a$ 。这对应上面例子的第 3 步; 若 LSB 为 1,  $n$  为奇数。除了将底数平方, 还要将  $b$  乘到累积结果  $a$  上; 当  $n$  为 0 时, 我们已处理完  $n$  中的所有位, 最终结果就是累积值  $a$ 。在任何时候, 更新的底数  $b'$ 、移位后的指数  $n'$ 、累积结果  $a$  总满足不变条件  $b^n = a(b')^{n'}$ 。此前的算法当  $n$  为奇数时, 仅仅将其减一转化为偶数进行处理; 这一改进每次都  $n$  减半。若  $n$  的二进制表示中有  $m$  位, 这一算法只计算  $m$  轮。当然它的复杂度仍然为  $O(\lg n)$ 。我们将其命令式实现作为练习。

回到求和、求积问题。在迭代实现中, 我们一边遍历, 一边累积结果:

```

1: function SUM( $X$ )
2:    $s \leftarrow 0$ 
3:   while  $X \neq \text{NIL}$  do
4:      $s \leftarrow s + \text{FIRST}(X)$ 
5:      $X \leftarrow \text{REST}(X)$ 
6:   return  $s$ 

7: function PRODUCT( $X$ )
8:    $p \leftarrow 1$ 
9:   while  $X \neq \text{NIL}$  do
10:     $p \leftarrow p \cdot \text{FIRST}(X)$ 
11:     $X \leftarrow \text{REST}(X)$ 
12:   return  $p$ 

```

利用求积算法, 我们可以将阶乘实现为递推方式:  $n! = \text{product}[1..n]$ 。

### 1.2.8 最大值和最小值

如果非空有限列表中的元素可进行比较,则存在最大、最小值。 $\max/\min$  的计算结构相同:若列表中只有一个元素  $[x_1]$ ,结果为  $x_1$ ;否则,递归地在子列表中寻找最大、最小值,并和表头元素比较得到最终结果。

$$\begin{aligned} \min [x] &= x & \max [x] &= x \\ \min (x:xs) &= \begin{cases} x < \min xs : x \\ \text{否则} : & \min xs \end{cases} & \max (x:xs) &= \begin{cases} x > \max xs : x \\ \text{否则} : & \max xs \end{cases} \end{aligned} \quad (1.24)$$

计算从右向左进行,我们可将其变为尾递归。这样就具备了“在线”处理能力:任何时候,累积结果都是已处理部分中的最大、最小值。以  $\min$  为例:

$$\begin{aligned} \min' a [] &= a \\ \min' a (x:xs) &= \begin{cases} x < a : \min' x xs \\ \text{否则} : \min' a xs \end{cases} \end{aligned} \quad (1.25)$$

与  $\text{sum}'/\text{prod}'$  不同,我们不能向  $\min'/\max'$  传入一个常数作为起始值,除非使用  $\pm\infty$ (柯里化形式):

$$\min = \min' \infty \quad \max = \max' -\infty$$

考虑到最大、最小值仅对非空列表定义,可以将表头元素传入作为起始值:

$$\min (x:xs) = \min' x xs \quad \max (x:xs) = \max' x xs \quad (1.26)$$

尾递归的最大、最小值算法可以进一步转化为迭代实现。以  $\text{MIN}$  为例:

```

1: function MIN(X)
2:    $m \leftarrow \text{FIRST}(X)$ 
3:    $X \leftarrow \text{REST}(X)$ 
4:   while  $X \neq \text{NIL}$  do
5:     if  $\text{FIRST}(X) < m$  then
6:        $m \leftarrow \text{FIRST}(X)$ 
7:      $X \leftarrow \text{REST}(X)$ 
8:   return  $m$ 

```

还有一种尾递归实现,可以复用表头元素作为累积器。递归时,由于列表中至少有两个元素,我们每次拿出前两个比较,丢弃一个,然后继续处理剩余的元素。以  $\min$  为例:

$$\begin{aligned} \min [x] &= x \\ \min (x_1:x_2:xs) &= \begin{cases} x_1 < x_2 : \min (x_1:xs) \\ \text{否则} : \min (x_2:xs) \end{cases} \end{aligned} \quad (1.27)$$

## 练习 1.8

1.8.1. 使用尾递归实现 *length*

1.8.2. 使用  $n$  的二进制形式, 实现幂  $b^n$  的快速计算, 使得复杂度为  $O(\lg n)$

答案: 350 页

## 1.3 变换

从代数角度看, 有两种不同的变换: 一种保持列表结构, 仅仅改变元素; 另一种改变列表结构, 变换结果和原列表不再同构。我们称保持列表结构的变换为**逐一映射**。

### 1.3.1 逐一映射

我们通过一些例子来认识映射。第一个例子将一系列数字映射为代表它们的字符串, 如把  $[3, 1, 2, 4, 5]$  转换为  $["three", "one", "two", "four", "five"]$ 。

$$\begin{aligned} \text{toStr } [] &= [] \\ \text{toStr } (x:xs) &= (\text{str } x) : \text{toStr } xs \end{aligned} \quad (1.28)$$

第二个例子是关于单词统计的。考虑一个字典, 包含若干单词, 并以它们的首字母分组, 例如:

```
[[a, an, another, ... ],
 [bat, bath, bool, bus, ...],
 ...,
 [zero, zoo, ...]]
```

接下来我们处理一段文章, 例如《王子复仇记》, 统计各单词在其中的出现次数:

```
[[ (a, 1041), (an, 432), (another, 802), ... ],
 [ (bat, 5), (bath, 34), (bool, 11), (bus, 0), ...],
 ...,
 [ (zero 12), (zoo, 0), ...]]
```

现在我们要找出, 对应每个首字母, 哪个单词被使用的次数最多? 输出结果是一个单词列表, 表中每个单词都是在各自首字母组中出现最多的一个, 形如:  $[a, but, can, \dots]$ 。我们需要设计一个程序, 将一组单词/次数对的列表转换成一个单词列表。首先定义一个函数, 接受一个单词/次数对列表, 找出次数最多的单词。我们无需排序, 只需要实现某种特殊的  $\text{maxBy cmp } xs$ , 其中  $\text{cmp}$  是抽象的比较函数。

$$\begin{aligned} \text{maxBy cmp } [x] &= x \\ \text{maxBy cmp } (x_1:x_2:xs) &= \begin{cases} \text{cmp } x_1 \ x_2 : \text{maxBy cmp } (x_2:xs) \\ \text{否则} : \text{maxBy cmp } (x_1:xs) \end{cases} \end{aligned} \quad (1.29)$$

对于一对值(也叫做数偶)  $p = (a, b)$ , 定义函数:

$$\begin{cases} fst(a, b) = a \\ snd(a, b) = b \end{cases} \quad (1.30)$$

定义单词/次数对的比较函数:

$$less\ p_1\ p_2 = snd\ p_1 < snd\ p_2 \quad (1.31)$$

将 *less* 传入 *maxBy* 就可选出出现次数最多的单词(柯里化):  $max'' = maxBy\ less$ 。最后,调用  $max''$  处理单词统计列表:

$$\begin{aligned} solve\ [] &= [] \\ solve\ (x:xs) &= (fst\ (max''\ x)) : solve\ xs \end{aligned} \quad (1.32)$$

尽管解决的问题不同, *solve* 和 *toStr* 反映出同样的计算结构。我们将这样的结构抽象为**逐一映射**。

$$\begin{aligned} map\ f\ [] &= [] \\ map\ f\ (x:xs) &= (f\ x) : map\ f\ xs \end{aligned} \quad (1.33)$$

*map* 接受一个函数 *f* 作为参数,然后将其应用到列表中的每个元素上。我们称将其它函数作为计算对象的函数为“高阶函数”。如果 *f* 的类型为  $A \rightarrow B$ ,即把类型 *A* 的元素映射为类型 *B* 的元素,则 *map* 的类型为:

$$map :: (A \rightarrow B) \rightarrow [A] \rightarrow [B] \quad (1.34)$$

读作:*map* 接受一个类型为  $A \rightarrow B$  的函数,然后将一个类型为  $[A]$  的列表变换为另一个类型为  $[B]$  的列表。上面的两个例子可以使用映射定义如下(柯里化):

$$toStr = map\ str \quad solve = map\ (fst \circ max'')$$

其中  $f \circ g$  表示函数组合,它首先应用函数 *g*,然后再应用函数 *f*,即  $(f \circ g)\ x = f(g(x))$ 。读作 *f* 作用于 *g* 之后。我们也可以从集合论的角度来定义映射。函数  $y = f(x)$  定义了一个从集合 *X* 中的元素 *x* 到集合 *Y* 中的元素 *y* 的映射:

$$Y = \{f(x) | x \in X\} \quad (1.35)$$

这种形式定义出的集合称为“策梅罗——弗兰克尔”集合抽象(简称 **ZF** 表达式)<sup>[72]</sup>。不同之处在于,我们定义的是从列表 *X* 到列表 *Y* 的映射  $Y = [f(x) | x \leftarrow X]$ ,而不是集合的映射。其中可以含有重复元素。列表的 **ZF** 表达式被称作“列表解析”,它是一个强大的工具。作为例子,我们思考如何实现一个排列算法。我们从全排列<sup>[72][94]</sup>出发,定义更一般的排列函数 *perm* *X* *r*,列举从长度为 *n* 的列表 *X* 中选出 *r* 个元素的全部排列。一共有  $P_n^r = \frac{n!}{(n-r)!}$  种不同的排列。

$$\text{perm } X \ r = \begin{cases} |X| < r \text{ 或 } r = 0: & [[]] \\ \text{否则:} & [x:ys \mid x \leftarrow X, ys \leftarrow \text{perm } (\text{delete } x \ X) \ (r - 1)] \end{cases} \quad (1.36)$$

如果选出 0 个元素排列,或列表中元素的个数小于  $r$ ,排列结果为空列表的列表 [[]]; 否则,我们逐一取出列表中每个元素  $x$ ,递归地从剩余的  $n - 1$  个元素中选择  $r - 1$  个元素排列,然后再将  $x$  置于每个排列的前面。

为了简化映射的迭代实现,下面的算法中使用了一个辅助节点。

```

1: function MAP( $f, X$ )
2:    $X' \leftarrow \text{CONS}(\perp, \text{NIL})$  ▷ 辅助节点
3:    $p \leftarrow X'$ 
4:   while  $X \neq \text{NIL}$  do
5:      $x \leftarrow \text{FIRST}(X)$ 
6:      $X \leftarrow \text{REST}(X)$ 
7:      $\text{REST}(p) \leftarrow \text{CONS}(f(x), \text{NIL})$ 
8:      $p \leftarrow \text{REST}(p)$ 
9:   return  $\text{REST}(X')$  ▷ 丢弃辅助节点

```

某些情况下我们只希望逐一处理表中的每个元素,而无需构造新的列表。例如打印一个列表中的每个元素:

```

1: function PRINT( $X$ )
2:   while  $X \neq \text{NIL}$  do
3:     print  $\text{FIRST}(X)$ 
4:      $X \leftarrow \text{REST}(X)$ 

```

通常,我们在遍历时传入一个过程  $P$ ,然后遍历列表,将  $P$  应用到每个元素上。

```

1: function FOR-EACH( $P, X$ )
2:   while  $X \neq \text{NIL}$  do
3:      $P(\text{FIRST}(X))$ 
4:      $X \leftarrow \text{REST}(X)$ 

```

作为例子,我们思考一下“ $n$  盏灯”趣题<sup>[96]</sup>:屋子里有  $n$  盏灯,都是熄灭的。我们执行下面的过程  $n$  次。

1. 将所有的灯都打开;
2. 扳动第 2、4、6……盏灯的开关。如果灯是亮的,则熄灭;如果是灭的,则点亮;
3. 每三个灯,扳动一次开关。第 3、6、9……位置上的灯的明暗状态切换;
4. ……

最后一轮的时候,只有最后一盏灯(第  $n$  盏)的开关被扳动。问最终有几盏灯是亮



的? 考虑穷举法。把  $n$  盏灯表示为一列 0、1 数字(0 灭、1 亮)。开始时都是灭的  $[0, 0, \dots, 0]$ 。将灯编号为 1 到  $n$ , 然后映射成一个关于  $(i, \text{亮/灭})$  的列表:

$$\text{lights} = \text{map } (i \mapsto (i, 0)) [1, 2, \dots, n]$$

这一映射将每个编号都对应到 0, 结果列表由  $n$  个数偶组成:  $L = [(1, 0), (2, 0), \dots, (n, 0)]$ 。操作  $L$  共  $n$  次。在第  $i$  次中, 逐一检查每对值  $(j, x)$ , 若  $i|j$  (即  $j \bmod i = 0$ ) 就翻转亮/灭。考虑  $1 - 0 = 1$  且  $1 - 1 = 0$ , 可将亮灭状态  $x$  切换为  $1 - x$ 。

$$\text{switch } i (j, x) = \begin{cases} j \bmod i = 0 : & (j, 1 - x) \\ \text{否则} : & (j, x) \end{cases} \quad (1.37)$$

第  $i$  轮操作实现为  $\text{map } (\text{switch } i) L$ 。这里使用了  $\text{switch}$  的柯里化形式。接下来定义函数  $op$ , 执行映射  $n$  次  $op [1, 2, \dots, n] L$ 。

$$\begin{aligned} op [] L &= L \\ op (i:is) L &= op \text{ is } (\text{map } (\text{switch } i) L) \end{aligned} \quad (1.38)$$

最后, 将数偶中第二个元素累加起来就得到答案。

$$\text{solve } n = \text{sum } (\text{map } \text{snd } (op [1, 2, \dots, n] L)) \quad (1.39)$$

下面的例子程序实现了穷举法。

```

solve = sum ◦ (map snd) ◦ proc where
  lights = map (λi → (i, 0)) [1..n]
  proc n = operate [1..n] lights
  operate [] xs = xs
  operate (i:is) xs = operate is (map (switch i) xs)
  switch i (j, x) = if j `mod` i == 0 then (j, 1 - x) else (j, x)
```

我们列出灯的数目为 1、2、……、100 盏时的答案(人为添加了换行):

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

这一结果很有规律:3 盏灯以内时, 最后有 1 盏灯亮;4 盏到 8 盏灯时, 最后有 2 盏灯亮;9 盏到 15 盏灯时, 最后有 3 盏灯亮;……看起来灯的数目为  $i^2$  到  $(i + 1)^2 - 1$  盏时, 最后有  $i$  盏灯亮。我们可以证明这一结论:

证明. 将  $n$  盏灯编号为 1 到  $n$ , 考虑最后仍然亮的那些灯。由于初始时, 所有灯都是灭的, 我们可以确定, 被扳动奇数次开关的灯最后是亮的。对于编号为  $i$  的灯, 若  $i$  可以被  $j$  整除 (表示为  $j|i$ ), 则第  $j$  轮, 它的开关被扳动一次。所以当灯的编号含有奇数个因子时, 最后的状态是亮的。为了找出最后亮的灯, 我们需要找出所有含有奇数个因子的数。对于任意自然数  $n$ , 记  $S$  为  $n$  的所有因子的集合。  $S$  初始化为  $\emptyset$ , 若  $p$  为  $n$  的一个因子, 则必然存在一个正整数  $q$ , 使得  $n = pq$ 。也就是说  $q$  也是  $n$  的因子。因此当且仅当  $p \neq q$  时, 我们向集合  $S$  中添加两个不同的因子, 这样  $|S|$  将总是偶数。除非  $p = q$ , 此时  $n$  必定是完全平方数。这时只能向集合  $S$  中增加一个因子, 从而导致奇数个因子。  $\square$

我们可以寻找  $n$  以内的完全平方数来解决这一趣题。

$$\text{solve}(n) = \lfloor \sqrt{n} \rfloor \quad (1.40)$$

下面的例子程序输出 1、2、……、100 盏灯时的结果:

```
map (floor ∘ sqrt) [1..100]
```

映射是一个抽象概念, 它不仅局限于列表, 也可以扩展到许多复杂的代数结构。下一章我们会介绍如何对树结构进行映射。只要我们能够遍历一个结构, 并且有空结构的定义, 就可以使用映射的概念。

### 1.3.2 反转

如何用最小的空间反转一个单向链表是一道经典题目, 需要仔细操作节点的引用。有一个简单的策略: (1) 先写出一个纯递归解; (2) 转换为尾递归; (3) 将尾递归转换为命令式操作。纯递归解很直观:

$$\begin{aligned} \text{reverse } [] &= [] \\ \text{reverse } (x:xs) &= \text{append } (\text{reverse } xs) x \end{aligned}$$

接下来进行尾递归优化。使用一个累积器来记录反转结果, 并传入空列表来启动:  $\text{reverse} = \text{reverse}' []$ 。

$$\begin{aligned} \text{reverse}' a [] &= a \\ \text{reverse}' a (x:xs) &= \text{reverse}' (x:a) xs \end{aligned} \quad (1.41)$$

不同于在尾部添加,  $\text{cons}(:)$  是常数时间操作。我们不断从列表的头部逐一取出元素, 将其置于累积结果的前面。这相当于将全部元素压入一个堆栈, 然后再依次弹出。整体上是线性时间的。由于尾递归无需记录上下文环境, 我们可以将其优化为循环迭代:

- 1: **function** REVERSE( $X$ )
- 2:      $A \leftarrow \text{NIL}$

```

3:   while X ≠ NIL do
4:     A ← CONS(FIRST(X), A)
5:     X ← REST(X)
6:   return A

```

但是,这一算法生成了一个新的反转列表,而不是在原列表上直接修改。我们接下来要通过重用  $X$  将其改为就地修改的形式。如下面的例子程序:

```

List<T> reverse(List<T> xs) {
  List<T> p, ys = null
  while xs ≠ null {
    p = xs
    xs = xs.next
    p.next = ys
    ys = p
  }
  return ys
}

```

## 练习 1.9

1.9.1. 使用尾递归求  $[(k, v)]$  列表中  $v$  值最大的元素。

答案: 351 页

## 1.4 子列表

数组可以快速地分割为连续的切片,而列表分割则需要遍历,因而这类操作都是线性时间的。 $take$  从列表中取出前  $n$  个元素,相当于:  $sublist\ 1\ n\ X$ 。 $drop$  从列表中丢弃前  $n$  个元素,相当于:  $sublist\ (n + 1)\ |X|\ X$ 。它和  $take$  对称<sup>6</sup>:

$$\begin{aligned}
 take\ 0\ xs &= [] & drop\ 0\ xs &= xs \\
 take\ n\ [] &= [] & drop\ n\ [] &= [] \\
 take\ n\ (x:xs) &= x : take\ (n - 1)\ xs & drop\ n\ (x:xs) &= drop\ (n - 1)\ xs
 \end{aligned}
 \tag{1.42}$$

这一算法是这样处理越界情况的:如果  $n > |X|$  或  $n$  为负数,最终转化为  $X$  为空的边界情况。我们把对应的迭代实现留作本节习题。使用取出和丢弃操作,可以在列表的任何位置获取指定长度的子列表。

$$sublist\ s\ n\ X = take\ n\ (drop\ (s - 1)\ X) \tag{1.43}$$

另外一种形式,是传入左侧和右侧的边界:

$$slice\ s\ e\ X = drop\ (s - 1)\ (take\ e\ X) \tag{1.44}$$

<sup>6</sup>有些语言内置了这两个操作,例如 Python 中:  $xs[:m]$  和  $xs[m:]$  分别相当于  $take/drop$ 。

闭区间  $[s, e]$  包括两端。我们也可以在指定位置把列表分割开：

$$\mathit{splitAt} \ i \ X = (\mathit{take} \ i \ X, \mathit{drop} \ i \ X) \quad (1.45)$$

$\mathit{take}/\mathit{drop}$  指定截取或丢弃的个数。我们可以对其扩展,只要某种条件成立,就不断取出或者丢弃元素,称为  $\mathit{takeWhile}/\mathit{dropWhile}$ 。它们逐一检查元素是否满足给定条件  $p$ ,如果不满足,则停止检查剩余的部分,这和后面介绍的过滤算法有所不同。

$$\begin{aligned} \mathit{takeWhile} \ p \ [] &= [] & \mathit{dropWhile} \ p \ [] &= [] \\ \mathit{takeWhile} \ p \ (x:xs) &= \begin{cases} p(x) : x : \mathit{takeWhile} \ p \ xs \\ \text{否则} : [] \end{cases} & \mathit{dropWhile} \ p \ (x:xs) &= \begin{cases} p(x) : \mathit{dropWhile} \ p \ xs \\ \text{否则} : x:xs \end{cases} \end{aligned} \quad (1.46)$$

### 1.4.1 切分和分组

切分和分组操作将列表中的元素重新安排成若干子列表。通常一边遍历一边进行,使得时间复杂度为线性。切分可以被认为是一种特殊的  $\mathit{split}$ ,我们不是在指定的位置将列表分开,而是检查每个元素是否满足某一条件,根据条件找出最长前缀。切分结果是一对子列表,一个是最长前缀,另一个包含剩余的部分。切分有两种类型:一种是满足条件的最长子列表;另一种是**不**满足条件的最长子列表。前者称为  $\mathit{span}$ ,后者称为  $\mathit{break}$ 。

$$\begin{aligned} \mathit{span} \ p \ [] &= ([], []) \\ \mathit{span} \ p \ (x:xs) &= \begin{cases} p(x) : (x:as, bs) \text{ 其中 } (as, bs) = \mathit{span} \ p \ xs \\ \text{否则} : ([], x:xs) \end{cases} \end{aligned} \quad (1.47)$$

只需要把条件取逻辑非,就可以实现  $\mathit{break} \ p = \mathit{span} \ (\neg p)$ 。 $\mathit{span}$  和  $\mathit{break}$  都寻找最长前缀。一旦条件打破就立即停下,忽略剩余部分。下面是 SPAN 的迭代实现：

```

1: function SPAN( $p, X$ )
2:    $A \leftarrow X$ 
3:    $tail \leftarrow \text{NIL}$ 
4:   while  $X \neq \text{NIL}$  且  $p(\text{FIRST}(X))$  do
5:      $tail \leftarrow X$ 
6:      $X \leftarrow \text{REST}(X)$ 
7:   if  $tail = \text{NIL}$  then
8:     return ( $\text{NIL}, X$ )
9:    $\text{REST}(tail) \leftarrow \text{NIL}$ 
10:  return ( $A, X$ )

```

$\mathit{span}$  和  $\mathit{break}$  将列表切分为两部分,分组将列表中的元素分成若干子列表。例如将字符串分割为若干子串,每个包含连续相同字符：

$$\mathit{group} \ \text{“Mississippi”} = [\text{“M”}, \text{“i”}, \text{“ss”}, \text{“i”}, \text{“ss”}, \text{“i”}, \text{“pp”}, \text{“i”}]$$

再例如,给出一列数字: $X = [15, 9, 0, 12, 11, 7, 10, 5, 6, 13, 1, 4, 8, 3, 14, 2]$ ,把它分成若干组,每组中的元素按降序排列:

$$\text{group } X = [[15, 9, 0], [12, 11, 7], [10, 5], [6], [13, 1], [4], [8, 3], [14, 2]]$$

这两个例子都具有实用价值。字符串分组后,可构造基数树用于快速文本搜索(第6章)。有序子列表可实现自然归并排序(第13章)。我们把分组条件抽象成某种关系 $\sim$ 。它用于判断两个相邻元素 $x, y$ 是否“等价”: $x \sim y$ 。我们遍历列表,每次比较两个元素。如果等价,就把它们置于一组;否则把 $x, y$ 分置于两组中。

$$\begin{aligned} \text{group } \sim [] &= [[]] \\ \text{group } \sim [x] &= [[x]] \\ \text{group } \sim (x:y:xs) &= \begin{cases} x \sim y : (x:ys):yss, \text{ 其中 } (ys:yss) = \text{group } \sim (y:xs) \\ \text{否则} : [x]:ys:yss \end{cases} \end{aligned} \tag{1.48}$$

这一算法的时间复杂度为 $O(n)$ ,其中 $n$ 是长度。也可以用迭代的方式实现这一算法。若 $X$ 不为空,我们将分组结果初始化为 $[[x_1]]$ 。然后从第二个元素开始遍历列表,若相邻的两个元素“等价”,我们就将遍历到的元素放入最后一组,否则就新建一个组。

```

1: function GROUP( $\sim, X$ )
2:   if  $X = \text{NIL}$  then
3:     return [[]]
4:    $x \leftarrow \text{FIRST}(X)$ 
5:    $X \leftarrow \text{REST}(X)$ 
6:    $g \leftarrow [x]$ 
7:    $G \leftarrow [g]$ 
8:   while  $X \neq \text{NIL}$  do
9:      $y \leftarrow \text{FIRST}(X)$ 
10:    if  $x \sim y$  then
11:       $g \leftarrow \text{APPEND}(g, y)$ 
12:    else
13:       $g \leftarrow [y]$ 
14:       $G \leftarrow \text{APPEND}(G, g)$ 
15:     $x \leftarrow y$ 
16:     $X \leftarrow \text{NEXT}(X)$ 
17:   return  $G$ 

```

如果添加 APPEND 没有尾部引用优化,这一实现的时间复杂度会退化为 $O(n^2)$ 。如果不关心顺序,可以将 APPEND 改为 CONS。使用分组函数,本节开头的两个例子可实现为: $\text{group}(=)$ “Mississippi”和 $\text{group}(\geq) X$ 。也可以使用 *span* 来实现分组。传

入一个条件, *span* 将列表分割成两部分: 其中之一是满足条件的最长子列表。我们对剩余部分不断执行 *span* 处理完所有元素。但是传入 *span* 的条件函数是一元函数, 而分组条件要求是二元函数。可以用柯里化来解决这一差异: 将第一个元素传入二元判断函数并固定。

$$\begin{aligned} \text{group} \sim [] &= [[]] \\ \text{group} \sim (x:xs) &= (x:as) : \text{group} \sim bs, \text{其中} : (as, bs) = \text{span} (x \sim) xs \end{aligned} \quad (1.49)$$

虽然这个新函数可以将文本中的相同字母分组, 但却不能正确地把数字按降序分组:  $\text{group} (\geq) X = [[15,9,0,12,11,7,10,5,6,13,1,4,8,3,14,2]]$ 。第一个元素是 15, 它被置于  $\geq$  的左侧进行比较。15 是列表中的最大元素, 因此 *span* 把所有元素都置于 *as* 中, 而 *bs* 为空。这并不是错误, 而是正确的行为。因为分组被设计为将“等价”的元素放到一起。严格说来, 等价关系( $\sim$ )条件必须满足三个性质: 自反性、对称性、传递性。

1. 自反性:  $x \sim x$ ;
2. 对称性:  $x \sim y \Leftrightarrow y \sim x$ 。
3. 传递性:  $x \sim y, y \sim z \Rightarrow x \sim z$ ;

对“Mississippi”分组时, 使用等号(=), 上述三个条件都满足, 产生了正确结果。但将柯里化的大于等于号( $\geq$ )作为等价关系时, 违反了自反性和对称性, 因而无法按预期分组。用 *span* 实现的第二个分组算法, 将含义限制为更严格的等价关系, 而第一个分组算法则无此限制。它仅检查任何两个相邻元素是否满足条件, 这比等价条件要弱。

### 练习 1.10

- 1.10.1. 修改 *take/drop*, 当  $n$  是负数时 *take* 返回 [], *drop* 返回全部列表。
- 1.10.2. 实现就地修改的 *take* 和 *drop* 算法。
- 1.10.3. 将 *sublist* 和 *slice* 改写为柯里化形式, 无需  $X$  作为参数。
- 1.10.4. 考虑下面 *span* 的实现:

$$\begin{aligned} \text{span } p [] &= ([], []) \\ \text{span } p (x:xs) &= \begin{cases} p(x) : (x : as, bs), \text{其中} : (as, bs) = \text{span}(p, xs) \\ \text{否则} : (as, x : bs) \end{cases} \end{aligned}$$

它和我们本节给出的实现有何不同?

## 1.5 叠加

几乎所有的列表算法都有着共同的结构。这不是一个巧合。这种共性本质上来自列表的递归性质。我们可以将列表算法抽象到更高层次的概念：叠加<sup>7</sup>。它本质上是所有列表计算的初始代数<sup>[99]</sup>。比较 *sum*、*product*、*sort*，它们都有共同的结构：列表为空时的结果，求和时为 0、求积时为 1、排序时为 `[]`；对表头元素和递归结果进行二元运算，求和时是相加、求积时是相乘、排序时是按序插入。我们将空列表时的结果抽象为**初始值**  $z$ （抽象的零），二元运算抽象为  $\oplus$ 。定义如下结构：

$$\begin{aligned} h \oplus z [] &= z \\ h \oplus z (x:xs) &= x \oplus (h \oplus z xs) \end{aligned} \tag{1.50}$$

输入列表  $X = [x_1, x_2, \dots, x_n]$ ，计算过程展开如下：

$$\begin{aligned} &h \oplus z [x_1, x_2, \dots, x_n] \\ &= x_1 \oplus (h \oplus z [x_2, x_3, \dots, x_n]) \\ &= x_1 \oplus (x_2 \oplus (h \oplus z [x_3, \dots, x_n])) \\ &\dots \\ &= x_1 \oplus (x_2 \oplus (\dots(x_n \oplus (h \oplus z []))\dots)) \\ &= x_1 \oplus (x_2 \oplus (\dots(x_n \oplus z)\dots)) \end{aligned}$$

这些括弧是必需的，它限制计算顺序从最右侧开始( $x_n \oplus z$ )，不断向左侧进行直到  $x_1$ 。这和图 1.3 描述的折扇相似。折扇由竹子和纸制成。多根扇骨在末端被轴穿在一起。把展开的扇形逐渐折叠，最终收起成一根。



图 1.3: 折扇

这些扇骨组成了一个列表。收起扇子的二元操作是将一根扇骨叠在已收起的部分之上。最初的收起部分为空。收起的过程从一端开始，不断应用二元操作，直到所有的扇骨都叠在一起。求和与求积算法和收起折扇的过程是相同的。

<sup>7</sup>也叫作 reduce

$$\begin{array}{ll}
 \text{sum } [1, 2, 3, 4, 5] = 1 + (2 + (3 + (4 + 5))) & \text{product } [1, 2, 3, 4, 5] = 1 \times (2 \times (3 \times (4 \times 5))) \\
 = 1 + (2 + (3 + 9)) & = 1 \times (2 \times (3 \times 20)) \\
 = 1 + (2 + 12) & = 1 \times (2 \times 60) \\
 = 1 + 14 & = 1 \times 120 \\
 = 15 & = 120
 \end{array}$$

我们称这一过程为**叠加**。特别地,由于计算从右侧一端开始,我们将其记为 *foldr* :

$$\begin{aligned}
 \text{foldr } f \ z \ [] &= z \\
 \text{foldr } f \ z \ (x:xs) &= f \ x \ (\text{foldr } f \ z \ xs)
 \end{aligned} \tag{1.51}$$

使用 *foldr*, 求和与求积可以定义如下:

$$\begin{aligned}
 \sum_{i=1}^n x_i &= x_1 + (x_2 + (x_3 + \dots + (x_{n-1} + x_n))) \dots \\
 &= \text{foldr } (+) \ 0 \ [x_1, x_2, \dots, x_n]
 \end{aligned} \tag{1.52}$$

$$\begin{aligned}
 \prod_{i=1}^n x_i &= x_1 \times (x_2 \times (x_3 \times \dots + (x_{n-1} \times x_n))) \dots \\
 &= \text{foldr } (\times) \ 1 \ [x_1, x_2, \dots, x_n]
 \end{aligned} \tag{1.53}$$

或者写成柯里化形式:  $\text{sum} = \text{foldr } (+) \ 0$  和  $\text{product} = \text{foldr } (\times) \ 1$ 。插入排序算法可定义为:  $\text{sort} = \text{foldr } \text{insert} \ []$ 。我们可以把 *foldr* 转换为尾递归,计算是从左向右进行的。我们将其记为 *foldl*:

$$\begin{aligned}
 \text{foldl } f \ z \ [] &= z \\
 \text{foldl } f \ z \ (x:xs) &= \text{foldl } f \ (f \ z \ x) \ xs
 \end{aligned} \tag{1.54}$$

以 *sum* 为例,可以看到计算是自左向右展开的:

$$\begin{aligned}
 &\text{foldl } (+) \ 0 \ [1, 2, 3, 4, 5] \\
 &= \text{foldl } (+) \ (0 + 1) \ [2, 3, 4, 5] \\
 &= \text{foldl } (+) \ (0 + 1 + 2) \ [3, 4, 5] \\
 &= \text{foldl } (+) \ (0 + 1 + 2 + 3) \ [4, 5] \\
 &= \text{foldl } (+) \ (0 + 1 + 2 + 3 + 4) \ [5] \\
 &= \text{foldl } (+) \ (0 + 1 + 2 + 3 + 4 + 5) \ [] \\
 &= 0 + 1 + 2 + 3 + 4 + 5
 \end{aligned}$$

每一步都推迟了  $f \ z \ x$  的计算,这是惰性求值时的行为。否则每次调用的求值序列为  $[1, 3, 6, 10, 15]$ 。一般来说, *foldl* 可以展开为下面的形式(中缀记法):

$$\text{foldl } (\oplus) \ z \ [x_1, x_2, \dots, x_n] = z \oplus x_1 \oplus x_2 \oplus \dots \oplus x_n \tag{1.55}$$

*foldl* 是尾递归的,可以将其实现为循环。通常称作 REDUCE。

- 1: **function** REDUCE( $f, z, X$ )
- 2:     **while**  $X \neq \text{NIL}$  **do**
- 3:          $z \leftarrow f(z, \text{FIRST}(X))$



```

4:      X ← REST(X)
5:      return z

```

*foldr* 和 *foldl* 各自有适合的应用场景,它们并不总能互换。例如,某些容器只支持从一端添加元素(如栈)。我们要定义一个 *fromList* 函数,从一个列表构建出容器(柯里化):

$$\text{fromList} = \text{foldr add } \emptyset$$

其中  $\emptyset$  是空容器。单向链表本身就是这样一种容器。在头部添加元素的性能要远高于在尾部添加。如果希望复制列表并保持顺序, *foldr* 就是一个自然的选择,而 *foldl* 则产生逆序的列表。在迭代实现时为了解决逆序问题,我们可以先反转列表,再执行 reduce 操作:

```

1: function REDUCE-RIGHT(f, z, X)
2:   return REDUCE(f, z, REVERSE(X))

```

有人认为应优先使用 *foldl*, 因为它是尾递归的,同时满足函数式和命令式场景,并且还是在线算法。但在处理无穷列表(用流和惰性求值实现)时,就只能用 *foldr*。下面的例子程序将一个无穷列表中的每个元素都置于一个单独的列表中,并返回前 10 个:

$$\begin{aligned} & \text{take } 10 \ (\text{foldr } (x \ xs \mapsto [x]:xs) \ [] \ [1, 2, \dots]) \\ & \Rightarrow [[1], [2], [3], [4], [5], [6], [7], [8], [9], [10]] \end{aligned}$$

这里不能使用 *foldl*, 否则计算永远不会完成。当左右没有区别时,我们用统一的符号 *fold*。本书也使用符号 *fold<sub>l</sub>* 和 *fold<sub>r</sub>* 来强调叠加本身而非方向。尽管本章内容是关于列表的,但是叠加概念是抽象的。它可以应用到其它代数结构。我们可以对树([99]2.6 节)、队列和更复杂的结构进行叠加。只要它满足下面这两个条件:(1)定义了空(例如空树);(2)可分解的递归结构(例如一棵树可分解为子树和元素)。人们进一步将这些概念抽象为可叠加、么半群、可遍历等等。

作为例子,我们用 *fold* 和 *map* 来解决 *n* 盏灯趣题。在穷举法中,我们创建了一个列表,每个元素是一对值  $(i, s)$ , 包含灯的序号 *i* 和明暗状态 *s*。每轮操作中,如果灯的序号 *i* 能被轮数 *j* 整除,就翻转灯的状态。这一过程可以用 *fold* 定义:

$$\text{foldr step } [(1, 0), (2, 0), \dots, (n, 0)] \ [1, 2, \dots, n]$$

初始时所有灯都是灭的。要折叠的列表是从 1 到 *n* 的轮数。函数 *step* 接受两个参数:轮数 *i* 和“灯序号/状态”列表  $\text{step } i \ L = \text{map } (\text{switch } i) \ L$ 。 *foldr* 的结果是最终的序号/明暗状态列表。接下来用 *map* 从每对值中提取出状态,再用 *sum* 求出有几盏灯是点亮的:

$$\text{sum } (\text{map snd } (\text{foldr step } [(1, 0), (2, 0), \dots, (n, 0)] \ [1, 2, \dots, n])) \quad (1.56)$$

如果用“ $\#$ ”(式 (1.15))在一组列表上折叠,相当于把它们串联成一个列表。这和数字累加是类似的:

$$\text{concat} = \text{fold}_r (\#) [] \quad (1.57)$$

例如:  $\text{concat} [[1], [2, 3, 4], [5, 6, 7, 8, 9]] \Rightarrow [1, 2, 3, 4, 5, 6, 7, 8, 9]$ 。

### 练习 1.11

1.11.1. 为了用  $\text{foldr}$  定义插入排序,我们将插入函数设计成  $\text{insert } x \ X$ , 这样排序可以表示为:  $\text{sort} = \text{foldr } \text{insert } []$ 。  $\text{foldr}$  的类型为:

$$\text{foldr} :: (A \rightarrow B \rightarrow B) \rightarrow B \rightarrow [A] \rightarrow B$$

其中第一个参数  $f$  的类型是:  $A \rightarrow B \rightarrow B$ , 初始值  $z$  的类型为  $B$ 。它对元素类型为  $A$  的列表进行叠加, 最终结果的类型为  $B$ 。如何用  $\text{foldl}$  定义插入排序?  $\text{foldl}$  的类型是什么?

1.11.2.  $\text{concat}$  的时间复杂度是什么? 设计一个线性时间的  $\text{concat}$  算法。

1.11.3. 使用  $\text{foldr}$  来定义  $\text{map}$ 。

答案: 352 页

## 1.6 查找和过滤

查找和过滤都是抽象概念, 不仅限于列表, 也可用于更广泛的对象。对于列表, 通常需要遍历以找到结果。最基本的操作是判断元素  $x$  是否属于列表  $X$ ? 我们可以遍历列表将每个元素和  $x$  进行对比。

$$\begin{aligned} a \in [] &= \text{False} \\ a \in (b:bs) &= \begin{cases} b = a : \text{True} \\ b \neq a : a \in bs \end{cases} \end{aligned} \quad (1.60)$$

这一操作也称作  $\text{elem}$ , 复杂度为  $O(n)$ 。即使列表有序, 仍然无法直接优化到  $O(\lg n)$ 。这是因为列表不支持常数时间的随机访问, 无法使用二分查找(见第 3 章)。接下来扩展  $\text{elem}$  操作。在  $n$  盏灯趣题中, 我们使用了“键/值”对列表  $[(k, v)]$ 。每对元素包含键、值, 称作“关联列表”。我们希望通过某个键查询到其对应的值。

$$\begin{aligned} \text{lookup } x [] &= \text{Nothing} \\ \text{lookup } x ((k, v):kvs) &= \begin{cases} k = x : \text{Just } (k, v) \\ k \neq x : \text{lookup } x \text{ } cvs \end{cases} \end{aligned} \quad (1.61)$$

和  $\text{elem}$  不同, 我们不仅想知道  $x$  存在与否, 还希望找到对应的值。由于键  $x$  并不一定存在, 我们引入“可能”代数类型 **Maybe**  $A$ , 它有两种不同的值: 类型  $A$  的某个值

$a$  或是空。分别记为 *Just a* 或 *Nothing*。这是一种解决空引用的方法<sup>8</sup>([\[99\]4.2.2 节](#))。进一步扩展 *lookup* 到一般情况。不再比较是否等于某个值,而是查找满足某一条件的元素:

$$\begin{aligned} \text{find } p [] &= \text{Nothing} \\ \text{find } p (x:xs) &= \begin{cases} p(x) : \text{Just } x \\ \text{否则} : \text{find } p xs \end{cases} \end{aligned} \quad (1.62)$$

尽管可能有多个元素满足条件, *find* 只返回第一个。我们可以把它扩展为查找满足条件的全部元素,这一过程通常称作过滤,如图 1.4 所示。其定义为 (ZF 表达式):  $\text{filter } p X = [x | x \leftarrow X, p x]$ 。可以用逻辑与将过滤条件串联起来:  $\text{filter } (p_1 \wedge p_2 \wedge \dots) X = [x | x \leftarrow X, p_1(x), p_2(x), \dots]$ <sup>9</sup>。



图 1.4: 输入:  $[x_1, x_2, \dots, x_n]$ , 输出:  $[x'_1, x'_2, \dots, x'_m]$ 。满足:  $\forall x'_i \Rightarrow p(x'_i)$ 。

和 *find* 不同,如果没有任何元素满足条件, *filter* 返回空列表。它逐一扫描列表,检查每个元素:

$$\begin{aligned} \text{filter } p [] &= [] \\ \text{filter } p (x:xs) &= \begin{cases} p(x) : x : \text{filter } p xs \\ \text{否则} : \text{filter } p xs \end{cases} \end{aligned} \quad (1.63)$$

这一算法从右向左构造结果。在迭代实现中,如果用 APPEND 来构造结果,性能会下降到  $O(n^2)$ 。如果用 CONS 替代,则结果是逆序的。我们可以再执行一次线性时间的反转(见练习)。

```

1: function FILTER( $p, X$ )
2:    $X' \leftarrow \text{NIL}$ 
3:   while  $X \neq \text{NIL}$  do
4:     if  $p(\text{FIRST}(X))$  then
5:        $X' \leftarrow \text{APPEND}(X', \text{FIRST}(X))$  ▷ 线性时间
6:      $L \leftarrow \text{REST}(X)$ 

```

从右向左进行计算的过程可以用 *foldr* 来定义。设计函数  $f$  检查每个元素,如果符合条件就添加到结果中:  $f p x as = \text{if } p x \text{ then } x:as \text{ else } as$ 。定义过滤为 (柯里化):

$$\text{filter } p = \text{foldr } (x as \mapsto f p x as) [] \quad (1.64)$$

<sup>8</sup>类似于某些编程环境的 `Optional<A>`。

<sup>9</sup>在上下文清楚的情况下可以简记为:  $[x \leftarrow X, p_1(x), p_2(x), \dots]$  或  $[p_1(x), p_2(x), \dots]$ 。

可以进一步简化为(称作  $\eta$  变换<sup>[73]</sup>):

$$\text{filter } p = \text{foldr } (f \ p) \ [] \quad (1.65)$$

过滤是一个通用的概念。不仅限于列表,可以对任何可遍历的结构应用某个判定条件,过滤出感兴趣的部分。

匹配一般是指在一个结构中寻找某一模式。即使限定为列表和字符串,匹配仍是一个广泛、深入的内容(第 14 章)。我们仅考虑检查列表  $as$  是否出现在列表  $bs$  中的问题。这里有两个特例:判断  $as$  是否是  $bs$  的前缀、后缀。 $\text{span}$  寻找符合某个条件的最长前缀。使用类似的方法逐一比较  $as$ 、 $bs$  中每个元素。若  $as$  是  $bs$  的前缀,记为:  $as \subseteq bs$ 。

$$\begin{aligned} [] \subseteq bs &= \text{True} \\ (a:as) \subseteq [] &= \text{False} \\ (a:as) \subseteq (b:bs) &= \begin{cases} a \neq b: \text{False} \\ a = b: as \subseteq bs \end{cases} \end{aligned} \quad (1.66)$$

由于扫描列表,前缀检查是线性时间的。但我们不能用同样的方法来检查后缀:对齐尾部,然后从右向左比较。这一点与数组不同。为了实现线性时间的后缀检查,可以反转两个列表,转换为前缀检查:

$$as \supseteq bs = \text{reverse}(as) \subseteq \text{reverse}(bs) \quad (1.67)$$

使用  $\subseteq$ ,可以解决子列表问题(中缀检查)。定义空列表是任何列表的中缀,遍历列表  $bs$  不断进行前缀检查:

$$\begin{aligned} \text{infix? } (a:as) \ [] &= \text{False} \\ \text{infix? } A \ B &= \begin{cases} as \subseteq (b:bs): \text{True} \\ \text{否则}: \text{infix? } A \ bs \end{cases} \end{aligned} \quad (1.68)$$

下面是对应的迭代实现:

```

1: function IS-INFIX(A, B)
2:   if A = NIL then
3:     return TRUE
4:   n ← |A|
5:   while B ≠ NIL 且 n ≤ |B| do
6:     if A ⊆ B then
7:       return TRUE
8:     B ← REST(B)
9:   return FALSE

```

由于前缀检测需要线性时间，并且在遍历时被不断调用，这一算法的复杂度为  $O(mn)$ ，其中  $m$  和  $n$  分别是两个列表的长度。对称地，我们可以枚举出  $B$  的所有后缀，然后检查  $A$  是否是某个后缀的前缀：

$$\text{infix? } A B = \exists S \in \text{suffixes } B, A \subseteq S \quad (1.69)$$

下面的例子程序使用列表解析实现了这一方法：

```
isInfixOf a b = (not ◦ null) [s | s ← tails b, a `isPrefixOf` s]
```

其中 `isPrefixOf` 进行前缀检查。`tails` 枚举一个列表的所有后缀(本节练习)。

## 练习 1.12

1.12.1. 使用 `reverse` 实现线性时间的过滤算法。

1.12.2. 给定一个列表，枚举出它的所有后缀。

答案: 353 页

## 1.7 关联和分解

关联列表是一种简易映射(字典)，用以处理少量数据。它比树、堆简单，但性能是线性的。在  $n$  盏灯趣题中，我们用逐一映射创建关联列表：`map (i ↦ (i, 0)) [1, 2, ..., n]`。为了“关联”两个列表，定义 `zip` 函数：

$$\begin{aligned} \text{zip } as [] &= [] \\ \text{zip } [] bs &= [] \\ \text{zip } (a:as) (b:bs) &= (a,b) : \text{zip } as bs \end{aligned} \quad (1.70)$$

这一定义可处理长度不同的列表。关联结果的长度将和较短的一个相同。我们甚至可以关联无穷列表(利用惰性求值)，例如<sup>10</sup>：`zip [0, 0, ...] [1, 2, ..., n]`。给定单词列表，我们可以给每个单词顺序编号为：`zip [1, 2, ...] [a, an, another, ...]`。`zip` 从右向左构建结果。我们可以用 `foldr` 定义它。复杂度为  $O(m)$ ，其中  $m$  是较短列表的长度。在迭代实现时，如果使用 `APPEND`，其性能会下降为平方时间。可以使用 `CONS` 再反转结果。但无法处理无穷列表。在命令式环境中，可以复用  $A$  来存储结果(视为一种映射：将一系列元素映射为一系列元素对)。

```
1: function ZIP(A, B)
2:   C ← NIL
3:   while A ≠ NIL 且 B ≠ NIL do
4:     C ← APPEND(C, (FIRST(A), FIRST(B)))
5:     A ← REST(A)
6:     B ← REST(B)
```

▷ 线性时间

<sup>10</sup>或 `zip (repeat 0) [1..n]`，其中 `repeat x = x : repeat x`

7: **return C**

进一步扩展 *zip* 关联多个列表,有些编程环境提供了 *zip*、*zip3*、*zip4*……。有时我们要应用某种组合函数,而不仅是构建元素对。例如,给定水果单价的列表:苹果、橙子、香蕉……单价为 [1.00, 0.80, 10.05, …](元);顾客购买数量为: [3, 1, 0, …],表示购买了 3 个苹果,1 个橙子、0 个香蕉。下面的程序计算应付金额:

$$\begin{aligned} \text{pays } us [] &= [] \\ \text{pays } [] qs &= [] \\ \text{pays } (u:us) (q:qs) &= uq : \text{pays } us qs \end{aligned}$$

除了用乘法代替 *cons*,其计算结构和 *zip* 相同。我们将组合函数抽象为 *f* 并传给 *zip* 定义出一般关联操作:

$$\begin{aligned} \text{zipWith } f as [] &= [] \\ \text{zipWith } f [] bs &= [] \\ \text{zipWith } f (a:as) (b:bs) &= (f a b) : \text{zipWith } f as bs \end{aligned} \tag{1.71}$$

利用 *zipWith* 可以定义内积(点积)<sup>[98]</sup>:  $A \cdot B = \text{sum } (\text{zipWith } (\cdot) A B)$ 。*zipWith* 结合惰性求值还可以定义无穷斐波那契数列:

$$F = 0 : 1 : \text{zipWith } (+) F F' \tag{1.72}$$

它的含义是 *F* 是无穷的斐波那契数列,第一个元素是 0,第二个元素是 1。*F'* 是去掉头部元素的无穷斐波那契数列。从第三个元素起,每个斐波那契数,都是 *F* 和 *F'* 中对应元素的和。下面的例子程序列出了前 15 个斐波那契数:

```
fib = 0 : 1 : zipWith (+) fib (tail fib)

take 15 fib
[0,1,1,2,3,5,8,13,21,34,55,89,144,233,377]
```

*zip* 的逆运算是 *unzip*, 将关联列表分解成两个列表。可以使用 *foldr* 定义分解(柯里化):

$$\text{unzip} = \text{foldr } ((a, b) (as, bs) \mapsto (a:as, b:bs)) ([], []) \tag{1.73}$$

对于买水果的例子,若单价信息以关联列表的形式给出:  $U = [(\text{apple}, 1.00), (\text{orange}, 0.80), (\text{banana}, 10.05), \dots]$ , 购买数量也是关联列表:  $Q = [(\text{apple}, 3), (\text{orange}, 1), (\text{banana}, 0), \dots]$ 。计算总金额时,从两个关联列表中分解出单价和数量,然后计算其内积:

$$\text{pay} = \text{sum } (\text{zipWith } (\cdot) \text{snd}(\text{unzip } U) \text{snd}(\text{unzip } Q)) \tag{1.74}$$

*zip* 和 *unzip* 的概念是抽象的。我们可以扩展 *zip* 以关联两棵树,节点中的数据是元素对,分别来自两棵树中。抽象的 *zip* 和 *unzip* 还可以用于跟踪复杂结构的遍历路径,从而模拟命令式环境中的父节点引用(<sup>[10]</sup>的最后一章)。

列表是构建复杂数据结构和算法的基础,对于函数式编程尤为重要。我们介绍了构建、分解、更改、变换列表的基本算法;介绍了如何在列表中搜索、过滤、进行计算。尽管大多数编程环境都提供了工具来支持列表,我们不应仅把它们看作黑盒子。Rabhi 和 Lapalme<sup>[72]</sup>介绍了关于列表的许多函数式算法。Haskell 标准库提供了关于基础算法的详细文档。伯德<sup>[1]</sup>给出了很多与叠加相关的例子,并介绍了“叠加融合定律”。

### 练习 1.13

1.13.1. 设计 *iota* 算法(希腊字母 *I*)其用法如下:

- $iota(\dots, n) = [1, 2, 3, \dots, n]$ ;
- $iota(m, n) = [m, m + 1, m + 2, \dots, n]$ , 其中  $m \leq n$ ;
- $iota(m, m + a, \dots, n) = [m, m + a, m + 2a, \dots, m + ka]$ ,  $k$  是使得  $m + ka \leq n$  的最大整数;
- $iota(m, m, \dots) = repeat(m) = [m, m, m, \dots]$ ;
- $iota(m, \dots) = [m, m + 1, m + 2, \dots]$ 。

1.13.2. 实现线性时间的命令式 *zip* 算法。

1.13.3. 用叠加定义 *zip*(提示:定义两个列表的叠加  $foldr2 f z xs ys$ )。

1.13.4. 使用 *zip* 实现 *lastAt*。

1.13.5. 编写一个程序从列表中去除重复的元素。在命令式环境中,请用就地修改的方式删除这些重复元素。在纯函数环境中,构建一个只含有不同元素的新列表。结果列表中的元素顺序应保持和原列表中的一致。这一算法的复杂度是怎样的?如果允许使用额外的数据结构,可以如何简化实现?

1.13.6. 可以用列表来表示十进制的非负整数。例如 1024 可以表示为:“ $4 \rightarrow 2 \rightarrow 0 \rightarrow 1$ ”。一般来说,  $n = d_m \dots d_2 d_1$  可以表示为:“ $d_1 \rightarrow d_2 \rightarrow \dots \rightarrow d_m$ ”。任给两个用列表表示的数  $a$  和  $b$ 。实现它们的算数运算(加减乘除)。

1.13.7. 在命令式环境中,循环列表是一种有缺陷的列表:某个节点指向了以前的位置,如图 1.6 所示。当遍历的时候,会陷入无限循环。设计一个算法检查某个列表是否含有循环。在此基础上,设计算法找到循环开始的节点(被两个不同祖先指向的节点)。

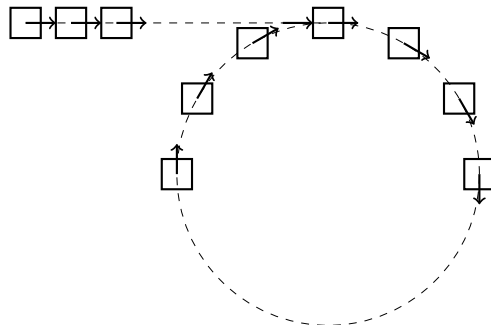


图 1.5: 带有循环的列表

答案: [353](#) 页



## 第二章 二叉搜索树

数组和链表通常被认为是最基础的数据结构,其实它们并不简单(第 12 章)。在某些系统中,数组是最基本的组件,甚至链表也可以由数组来实现(3.4)。另一方面,在函数式环境中,链表被作为最基本的组件来实现数组和其它更复杂的数据结构。二叉搜索树是另一种基础数据结构。乔·本特利在《编程珠玑》<sup>[2]</sup>中,讨论了如何统计一段文字中各单词的个数。下面的例子程序给出了一个解法。

```
void wordCount(Input in) {
    Map<String, Int> map
    while String w = read(in) {
        map[w] = if map[w] == null then 1 else map[w] + 1
    }
    for var (w, c) in map {
        print(w, ":", c)
    }
}
```

### 2.1 定义

这里的 `map` 是用二叉搜索树实现的字典数据结构。我们用单词作为键,用单词出现的次数作为值。这段程序展示了二叉搜索树的典型应用。我们先了解一下二叉树。一棵二叉树或者为空( $\emptyset$ )<sup>1</sup>,或者包含三个部分:一个元素  $k$  和左右两个分支  $l, r$ , 这两个分支也都是二叉树,记为  $(l, k, r)$ 。左右分支也被称为左子树和右子树,或统称为孩子。也可以说一棵树由若干节点构成。节点或为空或包含类型为  $K$  的值。二叉树的类型为  $Tree\ K$ 。如果一个节点的左右子树都为空,我们称之为叶子节点,否则称为分支节点。

二叉搜索树是一种特殊的二叉树,它的值可以进行比较<sup>2</sup>,并且满足:对于任何非空节点  $(l, k, r)$ ,左侧分支中所有的值都小于  $k$ ;  $k$  小于右侧分支中所有的值。图 2.2 展示了一棵二叉搜索树。和图 2.1 比较,可以看到节点组织方式的不同。一棵二叉树的值可以是任意类型,而二叉搜索树要求它的值必须能进行比较<sup>3</sup>。为了强调这种区别,我们特别称二叉搜索树的值为键 (key),称节点中存储的其它数据为值 (value),其类型

<sup>1</sup> $\emptyset$  是数学家安德烈·韦伊引入的挪威语字母,用来表示空集。

<sup>2</sup>广义比较,例如大小,先后、包含等序关系。本章中的“小于”及其符号  $<$  是抽象的比较。

<sup>3</sup>只要能进行抽象的“小于”和“等于”比较就足够了。

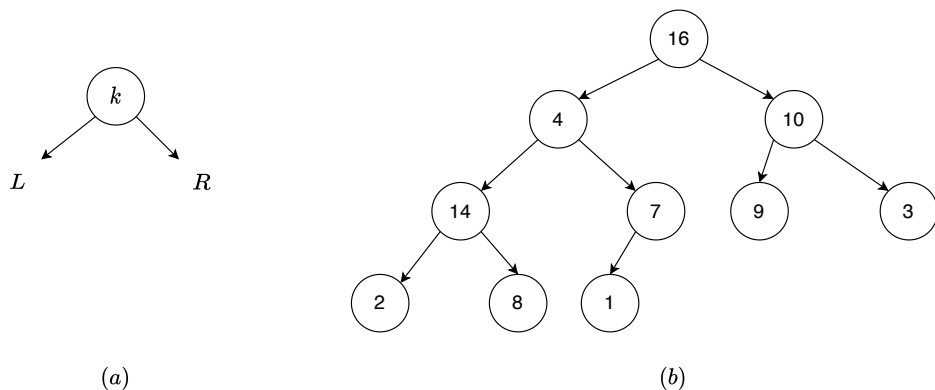


图 2.1: 二叉树的结构和例子

为  $Tree(K, V)$ 。为了方便上溯到祖先,可以在节点中存储父节点引用。我们在上下文清楚的情况下忽略值。本章附录给出了例子定义。在函数式环境中,一般不使用引用或指针回溯,而用自顶向下的递归进行计算。下面是函数式定义的例子(这类定义称为代数数据类型,简称为 ADT):

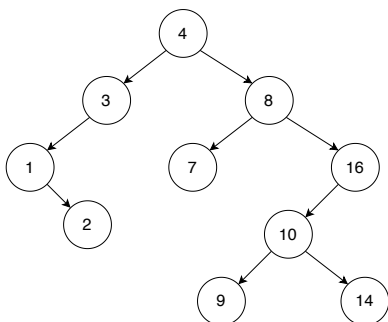


图 2.2: 二叉搜索树的例子

```
data Tree a = Empty | Node (Tree a) a (Tree a)
```

## 2.2 插入

向二叉搜索树插入  $k$  (和相关的数)时,需要确保树中元素有序。插入操作定义如下:(1) 如果树为空,创建一个元素为  $k$  的叶子节点;(2) 如果  $k$  小于根节点中的元素,将它插入到左子树中;否则,将  $k$  插入到右子树中。当  $k$  等于根节点中的元素时,说明它已经存在了。我们可以覆盖掉以前的数据(更新),或把新数据添加在后面,也可以跳过不做处理。简单起见,我们忽略这一情况。

$$\begin{aligned}
 \text{insert } k \ \emptyset &= (\emptyset, k, \emptyset) \\
 \text{insert } k \ (l, x, r) &= \begin{cases} k < x : (\text{insert } k \ l, x, r) \\ \text{否则} : (l, x, \text{insert } k \ r) \end{cases} \quad (2.1)
 \end{aligned}$$

下面是相应的例子程序：

```

insert k Empty = Node Empty k Empty
insert k (Node l x r) | k < x = Node (insert k l) x r
                       | otherwise = Node l x (insert k r)

```

这一例子程序使用了模式匹配。本章附录给出了不使用此特性的例子。也可以消除递归,用迭代实现插入：

```

1: function INSERT(T, k)
2:   root ← T
3:   x ← CREATE-LEAF(k)
4:   parent ← NIL
5:   while T ≠ NIL do
6:     parent ← T
7:     if k < KEY(T) then
8:       T ← LEFT(T)
9:     else
10:      T ← RIGHT(T)
11:    PARENT(x) ← parent
12:    if parent = NIL then
13:      return x
14:    else if k < KEY(parent) then
15:      LEFT(parent) ← x
16:    else
17:      RIGHT(parent) ← x
18:    return root

19: function CREATE-LEAF(k)
20:   x ← EMPTY-NODE
21:   KEY(x) ← k
22:   LEFT(x) ← NIL
23:   RIGHT(x) ← NIL
24:   PARENT(x) ← NIL
25:   return x

```

▷ *T* 为空

其中 KEY(*T*) 获取二叉树节点中的值,它可以定如下：

$$\begin{aligned} \text{key } \emptyset &= \text{Nothing} \\ \text{key } (l, k, r) &= \text{Just } k \end{aligned} \tag{2.2}$$

我们可以连续插入,将一组元素转变成二叉搜索树:

$$\begin{aligned} \text{fromList } [] &= \emptyset \\ \text{fromList } (x:xs) &= \text{insert } x (\text{fromList } xs) \end{aligned}$$

或通过叠加(第1章)将  $\text{fromList}$  定义为(柯里化):  $\text{fromList} = \text{foldr } \text{insert } \emptyset$ 。我们特意将参数顺序设计为对称的  $\text{insert } k t$  和  $\text{INSERT}(T, k)$ 。前者用  $\text{foldr}$ , 后者用  $\text{foldl}$ (或  $\text{for}$  循环):

```
1: function FROM-LIST( $X$ )
2:    $T \leftarrow \text{NIL}$ 
3:   for each  $x$  in  $X$  do
4:      $T \leftarrow \text{INSERT}(T, x)$ 
5:   return  $T$ 
```

## 2.3 遍历

遍历是指依次访问二叉树中的每个元素。有三种遍历方法:前序遍历、中序遍历、后序遍历。它们是按照访问根节点和子节点的先后顺序命名的。

- 前序遍历:先访问**根节点**,然后访问左子树,最后访问右子树;
- 中序遍历:先访问左子树,然后访问**根节点**,最后访问右子树;
- 后序遍历:先访问左子树,然后访问右子树,最后访问**根节点**。

所有的“访问”操作都是递归的。**先**访问根后访问子分支称为**先序**,在访问左右分支的**中间**访问根称为**中序**,先访问子分支**后**访问根称为**后序**。对于图 2.2 中的二叉树,三种遍历的结果分别如下:

- 前序遍历:4, 3, 1, 2, 8, 7, 16, 10, 9, 14
- 中序遍历:1, 2, 3, 4, 7, 8, 9, 10, 14, 16
- 后序遍历:2, 1, 3, 7, 9, 14, 10, 16, 8, 4

特别地,中序遍历会按照从小到大的顺序输出元素。二叉搜索树的定义保证了这一性质(练习 2.1.3)。定义  $\text{map}$  按中序遍历顺序将函数  $f$  应用到每个元素上,从而映射成另一棵同构的树。

$$\begin{aligned} \text{map } f \emptyset &= \emptyset \\ \text{map } f (l, k, r) &= (\text{map } f l, f k, \text{map } f r) \end{aligned} \tag{2.3}$$

如果只访问、操作各节点的值, 无需创建另外一棵树, 我们可以实现如下的中序遍历:

```

1: function TRAVERSE(T, f)
2:   if T ≠ NIL then
3:     TRAVERSE(LEFT(T), f)
4:     f(KEY(T))
5:     TRAVERSE(RIGHT(T), f)

```

我们可以修改 *map*, 通过中序遍历将一棵二叉搜索树转化为有序序列:

$$\begin{aligned} toList \ \emptyset &= [] \\ toList \ (l, k, r) &= toList \ l \ ++ \ [k] \ ++ \ toList \ r \end{aligned} \quad (2.4)$$

据此可以得到一个排序方法: 先把无序列表转化为二叉搜索树, 再用中序遍历把树转换回有序列表。称为“树排序”:  $sort \ X = toList \ (fromList \ X)$ , 或写成函数组合<sup>[8]</sup>的形式:

$$sort = toList \ \circ \ fromList \quad (2.5)$$

我们可以对二叉树进行中序叠加(叠加的概念见第一章)。

$$\begin{aligned} foldt \ f \ g \ z \ \emptyset &= z \\ foldt \ f \ g \ z \ (l, k, r) &= g \ (foldt \ f \ g \ z \ l) \ (f \ k) \ (foldt \ f \ g \ z \ r) \end{aligned} \quad (2.6)$$

这里  $f : A \rightarrow B$ , 将树中类型为  $A$  的元素映射为  $B$ , 这样  $f(k) = m$ 。它递归地对左右子树叠加(从  $z$  开始)得到  $x$  和  $y$ , 然后用  $g$  将三个值组合起来  $g \ x \ m \ y$ 。利用 *foldt*, 我们可以定义二叉树的映射 *map* 为:

$$map \ f = foldt \ f \ (x \ m \ y \mapsto (x, m, y)) \ \emptyset \quad (2.7)$$

*foldt* 在叠加时利用三元函数  $g$  保存了二叉树的结构, 如果不关心结构, 可以利用二元函数  $f : A \times B \rightarrow B$  实现叠加, 将类型为  $Tree \ A$  的树叠加为类型为  $B$  的值:

$$\begin{aligned} fold \ f \ z \ \emptyset &= z \\ fold \ f \ z \ (l, k, r) &= fold \ f \ (f \ k \ (fold \ f \ z \ r)) \ l \end{aligned} \quad (2.8)$$

例如:  $sum = fold \ (+) \ 0$  将二叉树中所有的元素累加起来。  $length = fold \ (x \ n \mapsto n + 1) \ 0$  可以统计树中元素个数。但是 *fold* 无法定义 *map*, 因为二元函数  $f$  丢失了树结构。

## 练习 2.1

2.1.1. 给定如下前序遍历和中序遍历的结果, 请重建出二叉树, 并给出后序遍历的结果。

- 前序遍历结果:1, 2, 4, 3, 5, 6
- 中序遍历结果:4, 2, 1, 5, 3, 6
- 后序遍历结果:?

2.1.2. 编程实现从前序遍历和中序遍历的结果重建二叉树。

2.1.3. 证明对二叉搜索树进行中序遍历可以将全部元素按照从小到大的顺序输出。

2.1.4. 树排序的复杂度是什么?

2.1.5. 使用叠加定义 *toList*

2.1.6. 使用叠加定义 *depth t*, 获得一棵二叉树的高度。

答案: 360 页

## 2.4 查找

由于二叉搜索树中的元素是按序递归存储的, 它可以方便地支持各种搜索。这也是人们将其命名为“搜索树”的原因。有三种不同类型的搜索: 1) 在树中查找一个键; 2) 寻找最大或最小元素; 3) 查找某一元素的前驱(上一个)或后继(下一个)元素。考虑在类型为  $Tree(K, V)$  的二叉搜索树中查找某个键  $x$  对应的值:

- 如果树为空,  $x$  不存在;
- 如果根节点元素  $(k, v)$  中  $k = x$ , 结果为  $v$ ;
- 如果  $x < k$ , 在左子树中递归查找; 否则在右子树中递归查找。

$$\begin{aligned}
 \text{lookup } x \ \emptyset &= \text{Nothing} \\
 \text{lookup } x \ (l, (k, v), r), x &= \begin{cases} k = x : \text{Just } v \\ x < k : \text{lookup } x \ l \\ \text{否则} : \text{lookup } x \ r \end{cases} \quad (2.9)
 \end{aligned}$$

这里使用 *Maybe* 类型<sup>4</sup> 处理未找到的情况。记树的高度为  $h$ , 查找算法的复杂度为  $O(h)$ 。如果二叉树平衡(见第四章), 对于  $n$  个元素的二叉树, 查找的性能为  $O(\lg n)$ 。如果二叉树很不平衡, 查找的时间在最坏情况下退化到  $O(n)$ 。下面是消除递归的迭代式实现:

```

1: function LOOKUP(T, x)
2:   while T ≠ NIL 且 KEY(T) ≠ x do
3:     if x < KEY(T) then
4:       T ← LEFT(T)
5:     else

```

<sup>4</sup>也叫作 `Optional<T>` 类型, 见第一章。

6:  $T \leftarrow \text{RIGHT}(T)$

7: **return** VALUE( $T$ ) ▷  $T = \text{NIL}$  返回  $\emptyset$

在二叉搜索树中,较小的元素总是位于左侧分支,而较大的元素总是位于右侧分支。可以利用这一特性来定位最大、最小元素。为了找到最小元素,我们可以不断向左侧前进,直到左侧分支为空。对称地,我们可以通过不断向右侧前进找到最大元素。这两个函数的性能都是  $O(h)$ ,其中  $h$  是树的高度。

$$\begin{aligned} \min(\emptyset, k, r) &= k & \max(l, k, \emptyset) &= k \\ \min(l, k, r) &= \min l & \max(l, k, r) &= \max r \end{aligned} \quad (2.10)$$

有时需要把二叉搜索树当作容器遍历。例如从最小的元素开始,逐一向前移动到最大元素,或者按需先后移动。下面的例子程序升序输出树中的元素:

```
void printTree (Node<T> t) {
    for var it = Iterator(t), it.hasNext(), it = it.next() {
        print(it.get(), ", ")
    }
}
```

这需要查找一个节点的前驱或后继元素。 $x$  的后继定义为满足  $x < y$  的最小  $y$ 。如果  $x$  的右子树不为空,则右子树中的最小值就是后继。图 2.3 中 8 的后继元素为 9,它是 8 的右子树中的最小值。如果  $x$  的右子树为空,需要向上回溯,找到最近的一个祖先  $p$ ,使得  $p$  的左子树也是  $x$  的祖先。在图 2.3 中,元素 2 所在的节点没有右侧分支,向上回溯一步到元素 1,但是 1 没有左侧分支,继续向上查找到元素 3。3 的左子树也是 2 的祖先。这样 2 的后继为 3。

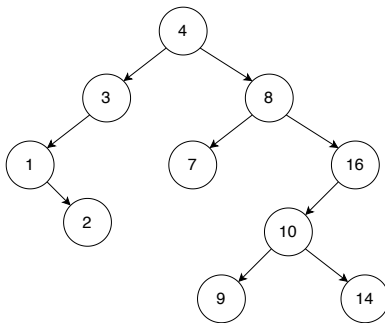


图 2.3: 8 的后继为其右侧分支的最小值 9;为了获得 2 的后继,首先向上找到 1,它没有左子树,继续向上找到 3,3 的左子树也是 2 的祖先,故而后继为 3。

如果沿着父节点引用一直回溯到了根节点,仍然没有找到位于右侧的祖先,这说明  $x$  没有后继(树中最后一个元素)。

```
1: function SUCC( $x$ )
2:   if RIGHT( $x$ )  $\neq$  NIL then
3:     return MIN(RIGHT( $x$ ))
```

```

4:  else
5:      p ← PARENT(x)
6:      while p ≠ NIL 且 x = RIGHT(p) do
7:          x ← p
8:          p ← PARENT(p)
9:      return p

```

当  $x$  没有后继时返回 NIL。寻找前驱元素与此对称：

```

1: function PRED(x)
2:     if LEFT(x) ≠ NIL then
3:         return MAX(LEFT(x))
4:     else
5:         p ← PARENT(x)
6:         while p ≠ NIL 且 x = LEFT(p) do
7:             x ← p
8:             p ← PARENT(p)
9:     return p

```

纯函数式环境中一般不使用父节点引用<sup>5</sup>。有的实现在遍历时留下标记，用以将来回溯甚至重建整棵树，称为 zipper<sup>[10]</sup> 最后一章)。查找前驱和后继的初衷是“作为一个通用容器，遍历树中的全部元素”。在纯函数式环境中，我们通常用 *map* 中序遍历所有元素。前驱和后续的查找，仅在命令式环境中才有意义。

## 练习 2.2

- 2.2.1. 判断某个值  $k$  是否存在于二叉搜索树  $t$  中。
- 2.2.2. 使用 PRED 和 SUCC 实现一个二叉搜索树的迭代器。用它遍历一棵含有  $n$  个元素的树的复杂度是什么？
- 2.2.3. 下面程序可以遍历一个区间  $[a, b]$  内的元素：

```

for_each (m.lower_bound(12), m.upper_bound(26), f)

```

试用纯函数式的方法解决这一问题

答案: 361 页

## 2.5 删除

我们必须保证删除后二叉搜索树仍保持有序：节点的左侧分支元素仍小于节点中元素，右侧分支元素仍大于节点中元素。直接删除节点会破坏这一性质。删除节点  $x$  的方法如下<sup>[6]</sup>：(1) 如果  $x$  是叶子节点或只有一棵非空子树，直接将  $x$  “切掉”；(2) 如果

<sup>5</sup>ML 或 OCaml 中有 `ref` 引用概念，这里我们限于纯函数式环境。



$x$  有两棵非空子树,用右子树中的最小值  $y$  替换掉  $x$ ,然后将原先的  $y$ “切掉”。我们利用了这样的特性:右子树中的最小值不可能有两棵非空子树。这样就把第二种情况转化为第一种,从而直接将原最小节点“切掉”。如同图 2.4、图 2.5、图 2.6 所示。

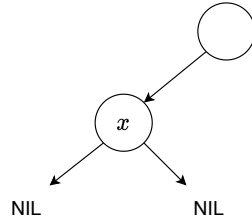
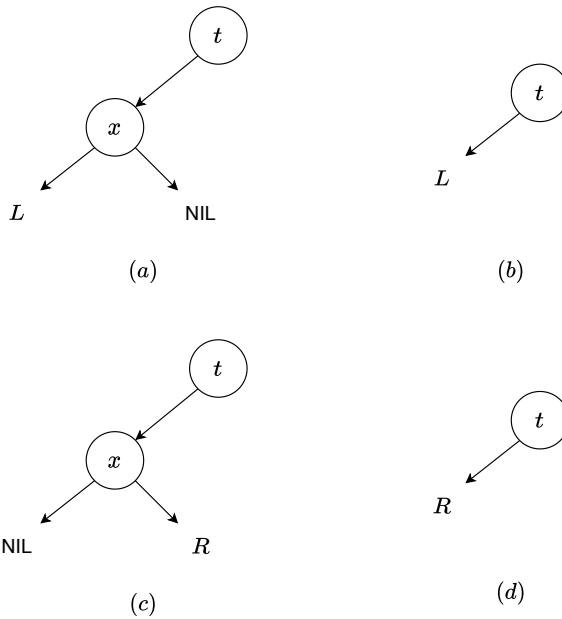
图 2.4: 直接“切掉”叶子节点  $x$ 。

图 2.5: 删除只有一个非空分支的节点

$$\begin{aligned}
 delete\ x\ \emptyset &= \emptyset \\
 delete\ x\ (l, k, r) &= \begin{cases} x < k : (delete\ x\ l, k, r) \\ x > k : (l, k, delete\ x\ r) \\ x = k : del\ l\ r \end{cases} \quad (2.11)
 \end{aligned}$$

其中:

$$\begin{aligned}
 del\ \emptyset\ r &= r \\
 del\ l\ \emptyset &= l \\
 del\ l\ r &= (l, y, delete\ y\ r), y = \min\ r \quad (2.12)
 \end{aligned}$$

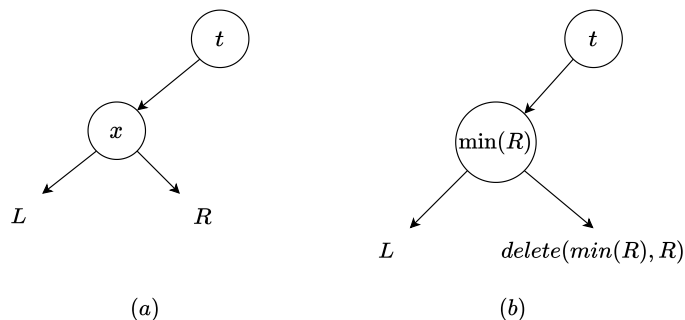


图 2.6: 删除有两个非空分支的节点

删除算法的复杂度为  $O(h)$ , 其中  $h$  是树的高度。命令式算法需要在删除后设置父节点。

```

1: function DELETE( $T, x$ )
2:    $r \leftarrow T$ 
3:    $x' \leftarrow x$                                      ▷ 保存  $x$ 
4:    $p \leftarrow \text{PARENT}(x)$ 
5:   if LEFT( $x$ ) = NIL then
6:      $x \leftarrow \text{RIGHT}(x)$ 
7:   else if RIGHT( $x$ ) = NIL then
8:      $x \leftarrow \text{LEFT}(x)$ 
9:   else                                               ▷ 两个分支都不为空
10:     $y \leftarrow \text{MIN}(\text{RIGHT}(x))$ 
11:    KEY( $x$ )  $\leftarrow$  KEY( $y$ )
12:    VALUE( $x$ )  $\leftarrow$  VALUE( $y$ )
13:    if PARENT( $y$ )  $\neq$   $x$  then                       ▷  $y$  的左子树为空
14:      LEFT(PARENT( $y$ ))  $\leftarrow$  RIGHT( $y$ )
15:    else                                             ▷  $y$  是右子树的根
16:      RIGHT( $x$ )  $\leftarrow$  RIGHT( $y$ )
17:    if RIGHT( $y$ )  $\neq$  NIL then
18:      PARENT(RIGHT( $y$ ))  $\leftarrow$  PARENT( $y$ )
19:    Remove  $y$ 
20:    return  $r$ 
21: if  $x \neq$  NIL then
22:   PARENT( $x$ )  $\leftarrow$   $p$ 
23: if  $p =$  NIL then                                   ▷ 删除根
24:    $r \leftarrow x$ 
25: else

```

```

26:         if LEFT( $p$ ) =  $x'$  then
27:             LEFT( $p$ )  $\leftarrow x$ 
28:         else
29:             RIGHT( $p$ )  $\leftarrow x$ 
30:     Remove  $x'$ 
31:     return  $r$ 

```

假定  $x$  不为空, 首先记录下根节点、待删除节点及其父节点。删除时还需要处理  $y$  是  $x$  右子树的根节点这一特殊情况。此后还要把之前保存的父节点重新设好。如果父节点为空, 则要删除的是根节点。这种情况下, 需要返回新的根。父节点设置好后, 就可以安全删除  $x$  了。

二叉搜索树操作的复杂度依赖于树的高度  $h$ 。如果树不平衡,  $O(h)$  接近  $O(n)$ 。反之如果平衡,  $O(h)$  接近  $O(\lg n)$ 。第四、五章介绍两种自动维持平衡的方法。还有一种简单的平衡策略 (<sup>[4]</sup> 第 265-268 页): 通过随机构建来减小不平衡的概率。先用随机函数打乱元素的次序, 然后再依次插入。

可以用二叉搜索树来实现键值映射 (Map, 也称为关联数据或字典), 存储若干“键、值”对。键是唯一的, 每个键映射为一个值。令键的类型是  $K$ , 值的类型是  $V$ , 映射的类型为  $Map\ K\ V$  或  $Map<K, V>$ 。非空映射包含  $n$  个关联 (映射) 关系:  $\{k_1 \mapsto v_1, k_2 \mapsto v_2, \dots, k_n \mapsto v_n\}$ 。使用二叉搜索树实现映射时, 我们限制  $K$  为有序集合。每个二叉树节点存储一对键、值, 类型为  $Tree(K, V)$ 。我们用二叉搜索树的插入或更新算法将一对键、值关联起来。给定键  $k$ , 我们使用 *lookup* 操作获取映射值  $v$ 。如果  $k$  不存在, 则返回  $\emptyset$ 。第四、五章中的红黑树、AVL 树也可以用来实现键值映射。

### 练习 2.3

- 2.3.1. 当节点的两个分支都不为空时, 存在一种对称的删除算法: 用左子树的最大值替换待删除的节点, 然后将此最大值的节点“切下”。编程实现这一算法。
- 2.3.2. 编程实现随机构建。
- 2.3.3. 如何在一棵二叉树中找到“距离最远”的两个节点?

答案: 362 页

## 2.6 附录:例子代码

包含父节点引用的二叉搜索树的例子定义:

```

data Node<T> {
    T key
    Node<T> left
    Node<T> right
    Node<T> parent

    Node(T k) = Node(null, k, null)

```

```

Node(Node<T> l, T k, Node<T> r) {
    left = l, key = k, right = r
    if (left ≠ null) then left.parent = this
    if (right ≠ null) then right.parent = this
}
}

```

不使用模式匹配的递归插入算法:

```

Node<T> insert (Node<T> t, T x) {
    if (t == null) {
        return Node(null, x, null)
    } else if (t.key < x) {
        return Node(insert(t.left, x), t.key, t.right)
    } else {
        return Node(t.left, t.key, insert(t.right, x))
    }
}
}

```

映射和叠加:

```

mapt _ Empty = Empty
mapt f (Node l x r) = Node (mapt f l) (f x) (mapt f r)

foldt _ _ z Empty = z
foldt f g z (Node l k r) = g (foldt f g z l) (f k) (foldt f g z r)

maptr :: (a → b) → Tree a → Tree b
maptr f = foldt f Node Empty

fold _ z Empty = z
fold f z (Node l k r) = fold f (k `f` (fold f z r)) l

```

消除递归的查找算法:

```

Optional<Node<T>> lookup (Node<T> t, T x) {
    while (t ≠ null and t.key ≠ x) {
        if (x < t.key) {
            t = t.left
        } else {
            t = t.right
        }
    }
    return Optional.of(t);
}
}

```

迭代寻找最小元素:

```

Optional<Node<T>> min (Node<T> t) {
    while (t ≠ null and t.left ≠ null) {
        t = t.left
    }
    return Optional.of(t);
}
}

```

寻找给定节点的后继:

```
Optional<Node<T>> succ (Node<T> x) {
  if (x == null) {
    return Optional.Nothing
  } else if (x.right != null) {
    return min(x.right)
  } else {
    p = x.parent
    while (p != null and x == p.right) {
      x = p
      p = p.parent
    }
    return Optional.of(p);
  }
}
```

删除:

```
delete _ Empty = Empty
delete x (Node l k r) | x < k = Node (delete x l) k r
                    | x > k = Node l k (delete x r)
                    | otherwise = del l r

where
  del Empty r = r
  del l Empty = l
  del l r = let k' = min r in Node l k' (delete k' r)
```



## 第三章 插入排序

插入排序是一种简单直观的排序算法<sup>1</sup>。在第一章中,我们给出了它的简明定义:对于一组可比较的元素,我们不断从中取出元素,按序将其插入到一个列表中。由于每次插入都需要线性时间,排序的复杂度为  $O(n^2)$ ,其中  $n$  是元素的个数,不如其它分治排序算法,如快速排序和归并排序。尽管如此,它仍然应用于现代软件中。如快速排序的实现通常在数据集较小的时候,回退到插入排序。

### 3.1 简介

扑克的抓牌过程形像地描述了插入排序的思想([4] 第 15 - 19 页)。任何时候,玩家手中的牌都是有序的。每当抓到一张新牌,就按照牌的点数,插入到合适的位置。如图 3.1。模拟这一过程就可以实现插入排序:



图 3.1: 将草花 8 插入到一手牌中

```
1: function SORT( $A$ )
2:    $S \leftarrow []$ 
3:   for each  $a \in A$  do
4:     INSERT( $a, S$ )
5:   return  $S$ 
```

这一实现将排序结果存储在新数组  $S$  中,也可以复用原数组进行就地排序:

---

<sup>1</sup>忽略冒泡排序算法

```

1: function SORT( $A$ )
2:   for  $i \leftarrow 2$  to  $|A|$  do
3:     ordered insert  $A[i]$  to  $A[1 \dots (i - 1)]$ 

```

其中索引  $i$  的范围从 1 到  $n = |A|$ 。只含有一个元素的子数组  $A[1]$  是已序的, 我们从第二个元素开始插入。当处理第  $i$  个元素时, 所有  $i$  之前的元素是已序的。我们不断将未排序的元素插入, 如图 3.2 所示。

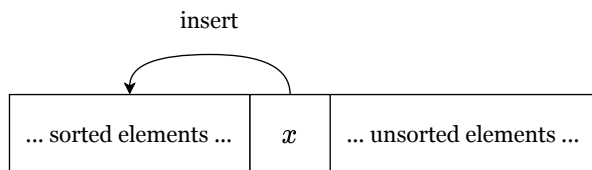


图 3.2: 不断将元素插入已序部分

## 3.2 插入

第一章给出了列表的插入算法。对于数组, 也可以通过逐一检查找到插入位置。检查可以从左向右或者从右向左进行。下面的实现是从右向左进行检查的:

```

1: function SORT( $A$ )
2:   for  $i \leftarrow 2$  to  $|A|$  do                                ▷ Insert  $A[i]$  to  $A[1 \dots (i - 1)]$ 
3:      $x \leftarrow A[i]$                                           ▷ 将  $A[i]$  保存到  $x$ 
4:      $j \leftarrow i - 1$ 
5:     while  $j > 0$  且  $x < A[j]$  do
6:        $A[j + 1] \leftarrow A[j]$ 
7:        $j \leftarrow j - 1$ 
8:      $A[j + 1] \leftarrow x$ 

```

由于数组是连续存储的, 在某一位置插入元素是一个代价较高的操作。若在第  $i$  个位置插入元素  $x$ , 需要把  $i$  后面的所有元素(包括  $A[i + 1]$ 、 $A[i + 2]$ ……)都向右移动一个位置。将第  $i$  个位置空出以放入  $x$ 。如图 3.3 所示。

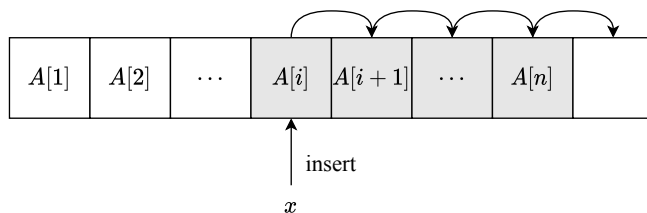


图 3.3: 将元素  $x$  插入数组  $A$  中的第  $i$  个位置



数组的长度为  $n$ , 若比较  $x$  和前  $i$  个元素后, 定位到了插入位置。接下来需要将剩余的  $n - i + 1$  的元素向后移动, 再将  $x$  放入第  $i$  个位置。整体上看, 我们相当于从左向右遍历了整个数组。另一方面, 如果从右向左处理, 则需要检查  $n - i + 1$  个元素, 并执行相同数量的移动操作。无论是从左向右或从右向左处理, 插入操作都需要线性时间, 因此插入排序的总体复杂度为  $O(n^2)$ 。我们也可以定义一个单独的 `INSERT()` 函数, 并在循环中调用。

### 练习 3.1

3.1.1. 实现从左向右处理的插入操作。

3.1.2. 定义单独的插入函数以实现插入排序。

答案: 363 页

## 3.3 二分查找

在玩扑克牌的时候, 人们并不是逐一比较找到插入位置的。我们之所以能快速定位, 是因为手中的牌在任何时刻都是已序的。二分查找是一种在已序序列中快速定位的方法。

```

1: function SORT( $A$ )
2:   for  $i \leftarrow 2$  to  $|A|$  do
3:      $x \leftarrow A[i]$ 
4:      $p \leftarrow$  BINARY-SEARCH( $x, A[1 \dots (i - 1)]$ )
5:     for  $j \leftarrow i$  down to  $p$  do
6:        $A[j] \leftarrow A[j - 1]$ 
7:      $A[p] \leftarrow x$ 

```

二分查找时, 数组中的片断  $A[1 \dots (i - 1)]$  是有序的。我们需要找到一个位置  $j$  使得  $A[j - 1] \leq x \leq A[j]$ 。比较  $x$  和中间位置元素  $A[m]$ , 其中  $m = \lfloor \frac{i}{2} \rfloor$ 。如果  $x < A[m]$ , 则递归地在前半序列中二分查找; 否则查找后半。由于每次都排除掉一半元素, 二分查找需要  $O(\lg i)$  的时间定位到插入点。

```

1: function BINARY-SEARCH( $x, A$ )
2:    $l \leftarrow 1, u \leftarrow 1 + |A|$ 
3:   while  $l < u$  do
4:      $m \leftarrow \lfloor \frac{l + u}{2} \rfloor$ 
5:     if  $A[m] = x$  then
6:       return  $m$ 
7:     else if  $A[m] < x$  then
8:        $l \leftarrow m + 1$ 
9:     else

```

▷ 重复元素

```

10:         u ← m
11:     return l

```

这一改进并不能提高插入排序的整体复杂度,结果仍然是  $O(n^2)$ 。逐一比较的插入排序需要  $O(n^2)$  次比较和  $O(n^2)$  次移动;使用二分查找后,比较次数减少到了  $O(n \lg n)$ ,但移动次数还是  $O(n^2)$ 。

### 3.4 列表

二分查找将总体比较次数降低到  $O(n \lg n)$ ,但由于要依次移动数组中的元素,整体复杂度仍然是  $O(n^2)$ 。另一方面,使用列表存储元素时,一旦移动到指定位置,插入操作本身是常数时间的。在第一章中,我们定义了如下的列表插入排序算法:

$$\begin{aligned}
 \text{sort } [] &= [] \\
 \text{sort } (x:xs) &= \text{insert } x (\text{sort } xs)
 \end{aligned} \tag{3.1}$$

或使用 *foldr* 的柯里化形式:  $\text{sort} = \text{foldr } \text{insert } []$ 。由于需要遍历,列表的 *insert* 算法仍是线性时间的:

$$\begin{aligned}
 \text{insert } x [] &= [x] \\
 \text{insert } x (y:ys) &= \begin{cases} x \leq y : x : y : ys \\ \text{否则} : y : \text{insert } x \ ys \end{cases}
 \end{aligned} \tag{3.2}$$

也可以不使用节点引用,而通过另一个索引数组来实现列表。对任何元素  $A[i]$ ,  $\text{Next}[i]$  保存了  $A[i]$  之后下一个元素的索引。也就是说  $A[\text{Next}[i]]$  是  $A[i]$  的下一个元素。其中有两个特殊索引:对于列表的末尾元素  $A[m]$ ,定义  $\text{Next}[m] = -1$ ,表示其指向 NIL;此外定义  $\text{Next}[0]$  指向列表的头部。利用索引数组,我们定义插入算法如下:

```

1: function INSERT(A, Next, i)
2:     j ← 0                                ▷ Next[0] 指向表头
3:     while Next[j] ≠ -1 and A[Next[j]] < A[i] do
4:         j ← Next[j]
5:     Next[i] ← Next[j]
6:     Next[j] ← i

7: function SORT(A)
8:     n ← |A|
9:     Next = [1, 2, ..., n, -1]           ▷ n + 1 个索引
10:    for i ← 1 to n do
11:        INSERT(A, Next, i)
12:    return Next

```

使用列表,尽管在引用位置进行插入只需要常数时间,但必须遍历才能找到插入位置。整体仍需要  $O(n^2)$  次比较。与数组不同,列表不支持随机访问,不能利用二分查找提升定位速度。

### 练习 3.2

3.2.1. 使用索引数组,排序结果是一个重新排列的索引。给出一个方法,根据新的索引  $Next$ ,重新排列数组  $A$ 。

答案: 364 页

## 3.5 二叉搜索树

我们遇到了一个困难境地:必须同时提高查找和插入的速度,仅提高其中之一仍然是  $O(n^2)$  的复杂度。一方面,我们希望用二分查找把比较次数降低到  $O(\lg n)$ ;另一方面,需要改变数据结构,因为数组不支持在指定位置以常数时间插入元素。在第二章中,我们介绍了二叉搜索树。它天然就支持二分查找。一旦定位到插入位置,我们可以用常数时间插入新节点。

```

1: function SORT( $A$ )
2:    $T \leftarrow \emptyset$ 
3:   for each  $x \in A$  do
4:      $T \leftarrow \text{INSERT-TREE}(T, x)$ 
5:   return TO-LIST( $T$ )

```

对于列表定义为  $sort = toList \circ fromList$ 。第二章给出了 INSERT-TREE()、TO-LIST()、fromList 的定义。平均情况下,树排序的复杂度为  $O(n \lg n)$ ,其中  $n$  是元素的个数。这达到了基于比较的排序算法时间下限 (<sup>[12]</sup> 第 180-193 页、<sup>[4]</sup> 第 167 页)。但在最坏情况下,当树极度不平衡时,其性能会下降到  $O(n^2)$ 。

插入排序通常作为第一个排序算法被介绍。它简单直观,但性能是平方级别的。插入排序不仅出现在教科书中,也出现在快速排序的工程实现中:在小数据集时回退到插入排序以抵消递归的代价。



## 第四章 红黑树

第二章的例子使用二叉搜索树来统计文章中单词的频次。能否用它来处理通讯录、查询联系人呢？如下面的例子所示：

```
void addrBook(Input in) {
    Map<String, String> dict
    while (String name, String addr) = read(in) {
        dict[name] = addr
    }
    loop {
        string name = read(Console)
        var addr = dict[name]
        if (addr == null) {
            print("not found")
        } else {
            print("address: ", addr)
        }
    }
}
```

但这个方法性能不佳，尤其是搜索 Zara、Zed、Zulu 等姓名时更加明显。通讯录是按照字典顺序排列的。如果依次把 1, 2, 3, ...,  $n$  插入二叉搜索树，就会得到图 4.1 中的结果。查找的复杂度为  $O(h)$ ,  $h$  为高。平衡时  $h = O(\lg n)$ 。这一极端情况下，查找的性能退化为  $O(n)$ ，等同于列表扫描。

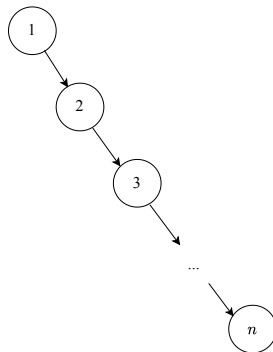


图 4.1: 不平衡的树

### 练习 4.1

4.1.1. 对于较大的通讯录,为了加快构建速度,可以使用两个并发的任务:一个从头部向后,另外一个从后向前读取。当两个任务相遇时结束。这样构建出的二叉搜索树是什么样子的?如果把通讯录分成更多片断,使用多任务会得到什么结果?参考图 4.2,找出更多的不平衡情况。

答案: 364 页

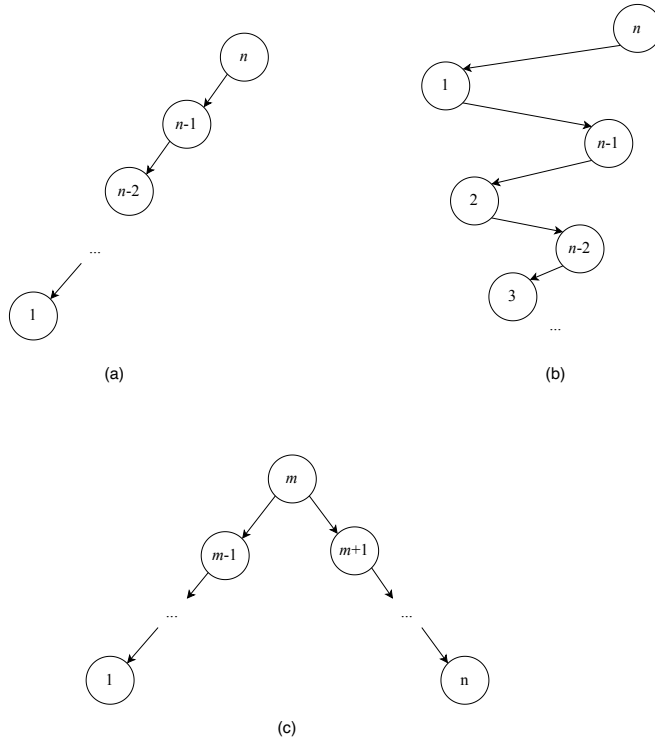


图 4.2: 一些不平衡的二叉树

## 4.1 平衡

为避免极不平衡的情况,可将输入序列打乱(2.5)。但这种方法有局限性,无法处理交互输入的序列。人们找到了一些保持平衡性的方法,它们大多依赖二叉树的旋转操作,可以在改变树结构的同时保持元素有序。本章介绍红黑树,下一章介绍 AVL 树,它们都是自平衡树。第 8 章还会介绍伸展树,它能够随着操作,逐步恢复平衡。不同的二叉树可以产生相同的中序遍历结果。树旋转如图 4.3 所示,可以通过模式匹配来定义:

$$\begin{aligned}
 rotate_l(a, x, (b, y, c)) &= ((a, x, b), y, c) & rotate_r((a, x, b), y, c) &= (a, x, (b, y, c)) \\
 rotate_l T &= T & rotate_r T &= T
 \end{aligned}
 \tag{4.1}$$

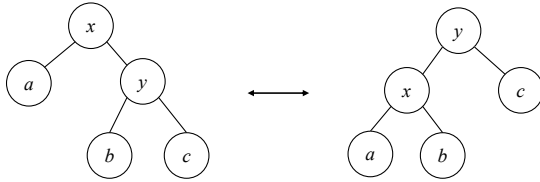


图 4.3: 树的左右旋转

如果模式没有匹配(例如两棵子树都为空),每个式子的第二行保持树不变。旋转也可以通过一系列步骤实现。我们需要将子树和父引用设置正确。在旋转时,我们传入根节点  $T$  和要旋转的子树  $x$ :

```

1: function LEFT-ROTATE( $T, x$ )
2:    $p \leftarrow$  PARENT( $x$ )
3:    $y \leftarrow$  RIGHT( $x$ )           ▷ 设  $y \neq$  NIL
4:    $a \leftarrow$  LEFT( $x$ )
5:    $b \leftarrow$  LEFT( $y$ )
6:    $c \leftarrow$  RIGHT( $y$ )
7:   REPLACE( $x, y$ )                 ▷ 用  $y$  替换  $x$ 
8:   SET-SUBTREES( $x, a, b$ )          ▷ 令  $a, b$  为  $x$  的子树
9:   SET-SUBTREES( $y, x, c$ )         ▷ 令  $x, c$  为  $y$  的子树
10:  if  $p =$  NIL then              ▷ 此前  $x$  是根节点
11:     $T \leftarrow y$ 
12:  return  $T$ 

```

右旋 RIGHT-ROTATE 的实现是对称的(练习 4.2)。REPLACE( $x, y$ ) 使用  $y$  替换  $x$ :

```

1: function REPLACE( $x, y$ )
2:    $p \leftarrow$  PARENT( $x$ )
3:   if  $p =$  NIL then             ▷  $x$  是根节点
4:     if  $y \neq$  NIL then PARENT( $y$ )  $\leftarrow$  NIL
5:   else if LEFT( $p$ ) =  $x$  then
6:     SET-LEFT( $p, y$ )
7:   else
8:     SET-RIGHT( $p, y$ )
9:   PARENT( $x$ )  $\leftarrow$  NIL

```

SET-SUBTREES( $x, L, R$ ) 将  $L$  设为  $x$  的左子树,  $R$  设为右子树:

```

1: function SET-SUBTREES( $x, L, R$ )
2:   SET-LEFT( $x, L$ )
3:   SET-RIGHT( $x, R$ )

```

它进一步调用 SET-LEFT 和 SET-RIGHT 完成子树的设置:

```

1: function SET-LEFT( $x, y$ )
2:   LEFT( $x$ )  $\leftarrow y$ 
3:   if  $y \neq \text{NIL}$  then PARENT( $y$ )  $\leftarrow x$ 

4: function SET-RIGHT( $x, y$ )
5:   RIGHT( $x$ )  $\leftarrow y$ 
6:   if  $y \neq \text{NIL}$  then PARENT( $y$ )  $\leftarrow x$ 

```

通过对比,可以看到模式匹配如何简化树旋转的实现。从这一点出发冈崎在 1995 年实现了红黑树的纯函数式算法<sup>[13]</sup>。

## 练习 4.2

4.2.1. 实现右旋 RIGHT-ROTATE 操作。

答案: 365 页

## 4.2 定义

红黑树是一种自平衡二叉搜索树<sup>[14]</sup>。它是 2-3-4 树的等价形式<sup>1</sup>。通过对节点进行着色和旋转,红黑树可以高效地保持平衡。我们在二叉搜索树的定义上给节点赋予红、黑颜色。我们称一棵树为红黑树,如果它满足下面 5 条性质<sup>[4]</sup>:

1. 节点的颜色为红色或黑色。
2. 根节点为黑色。
3. NIL 节点为黑色。
4. 如果一个节点为红色,则它的两个子节点都是黑色。
5. 从任一节点出发到所有叶子节点的路径上包含相同数量的黑色节点。

为什么这 5 条性质能保证红黑树的平衡性呢? 关键在于: 从根节点出发到达叶节点的所有路径中,最长路径不会超过最短路径两倍。性质 4 保证了不存在两个连续的红色节点。因此,最短的路径只含有黑色的节点。任何更长的路径一定含有红色节点。根据性质 5,从任何节点出发的所有的路径都含有相同数量的黑色节点,自然这条对于根节点也成立。这就最终保证了没有任何路径超过最短路径长度的两倍<sup>[14]</sup>。图 4.4 的例子展示了一棵红黑树。

由于 NIL 节点都是黑色的,通常隐藏不画出,如图 4.5 所示。所有不改变树结构的操作都和二叉搜索树相同,包括查找、最大、最小值等。只有插入和删除操作是特殊

<sup>1</sup>第 7 章, B 树。对于任一 2-3-4 树,都存在至少一棵红黑树,其元素顺序相同。



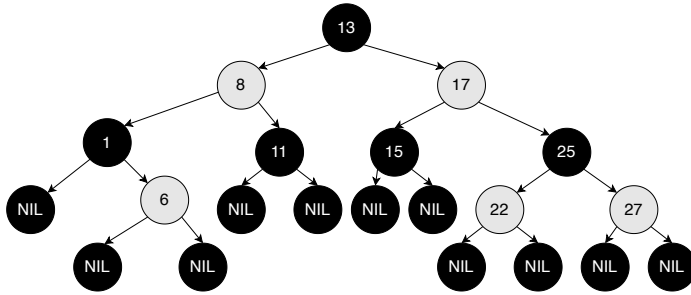


图 4.4: 红黑树

的。下面的例子程序在二叉搜索树的基础上增加了颜色定义。我们记空树为  $\emptyset$ , 否则记为  $(c, l, k, r)$ , 分别表示颜色  $c$  (红/黑)、元素  $k$ , 左右子树  $l, r$ 。

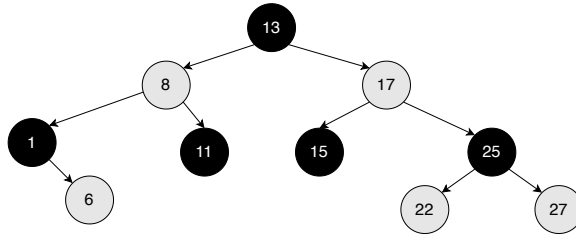


图 4.5: 隐藏 NIL 节点

```

data Color = R | B
data RBTREE a = Empty | Node Color (RBTREE a) a (RBTREE a)
  
```

### 练习 4.3

4.3.1. 证明含有  $n$  个节点的红黑树, 其高度  $h$  不会超过  $2\lg(n+1)$ 。

答案: 365 页

## 4.3 插入

插入算法包含两个步骤: 第一步和二叉搜索树相同, 树可能会变得不再平衡; 第二步修复红黑树的颜色性质。插入时, 我们令新节点为红色。只要它不是根节点, 除了第四条外的所有性质都可以满足。唯一的问题是可能引入两个相邻的红色节点, 共有 4 种情况需要修复。它们具有统一的形式<sup>[13]</sup>, 如图 4.6 所示。

四种情况都把红色向上移动一层。如果进行自底向上的递归修复, 可能会把根节点染成红色。根据性质 2, 最后需要把根节点变回黑色。利用模式匹配, 我们定义 *balance* 函数修复平衡。令节点的颜色变量为  $C$ , 取值为黑色  $B$  或红色  $R$ 。

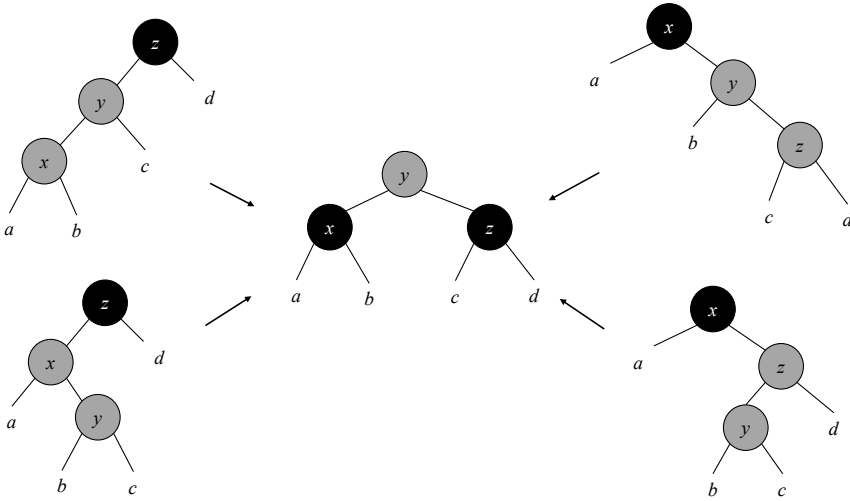


图 4.6: 插入后需要修复的四种情况

$$\begin{aligned}
 \text{balance } \mathcal{B} (\mathcal{R}, (\mathcal{R}, a, x, b), y, c) z d &= (\mathcal{R}, (\mathcal{B}, a, x, b), y, (\mathcal{B}, c, z, d)) \\
 \text{balance } \mathcal{B} (\mathcal{R}, a, x, (\mathcal{R}, b, y, c)) z d &= (\mathcal{R}, (\mathcal{B}, a, x, b), y, (\mathcal{B}, c, z, d)) \\
 \text{balance } \mathcal{B} a x (\mathcal{R}, b, y, (\mathcal{R}, c, z, d)) &= (\mathcal{R}, (\mathcal{B}, a, x, b), y, (\mathcal{B}, c, z, d)) \\
 \text{balance } \mathcal{B} a x (\mathcal{R}, (\mathcal{R}, b, y, c), z, d) &= (\mathcal{R}, (\mathcal{B}, a, x, b), y, (\mathcal{B}, c, z, d)) \\
 \text{balance } T &= T
 \end{aligned} \tag{4.2}$$

如果四种模式都不满足,最后一行保证此时不会改变树的形状。红黑树的插入算法定义为:  $\text{insert } x T = \text{makeBlack} (\text{ins } x T)$ , 或写成柯里化的形式:

$$\text{insert } x = \text{makeBlack} \circ \text{ins } x \tag{4.3}$$

其中:

$$\begin{aligned}
 \text{ins } x \emptyset &= (\mathcal{R}, \emptyset, x, \emptyset) \\
 \text{ins } x (\mathcal{C}, l, k, r) &= \begin{cases} x < k: \text{balance } \mathcal{C} (\text{ins } x l) k r \\ x > k: \text{balance } \mathcal{C} l k (\text{ins } x r) \end{cases}
 \end{aligned} \tag{4.4}$$

如果树为空,我们创建一个值为  $x$  的红色叶子节点;否则比较  $x$  和  $k$  的大小,递归地将  $x$  插入到子树中。然后用  $\text{balance}$  修复平衡。最后把根节点染成黑色。

$$\text{makeBlack} (\mathcal{C}, l, k, r) = (\mathcal{B}, l, k, r) \tag{4.5}$$

下面是对应的例子程序:

```

insert x = makeBlack ◦ (ins x) where
  ins x Empty = Node R Empty x Empty
  ins x (Node color l k r)
    | x < k    = balance color (ins x l) k r
    | otherwise = balance color l k (ins x r)

```

```

makeBlack (Node _ l k r) = Node B l k r

balance B (Node R (Node R a x b) y c) z d = Node R (Node B a x b) y (Node B c z d)
balance B (Node R a x (Node R b y c)) z d = Node R (Node B a x b) y (Node B c z d)
balance B a x (Node R b y (Node R c z d)) = Node R (Node B a x b) y (Node B c z d)
balance B a x (Node R (Node R b y c) z d) = Node R (Node B a x b) y (Node B c z d)
balance color l k r = Node color l k r

```

我们略去了重复值的处理。如果要插入的值已经存在,我们可以覆盖或丢弃,还可以在节点中用一个列表存储相应的数据<sup>[4]</sup>,269页)。图 4.7 给出了两棵红黑树。它们分别由序列 11, 2, 14, 1, 7, 15, 5, 8, 4 和 1, 2, ..., 8 构建而成。第二个例子说明,即使序列已序,红黑树仍然保持平衡。

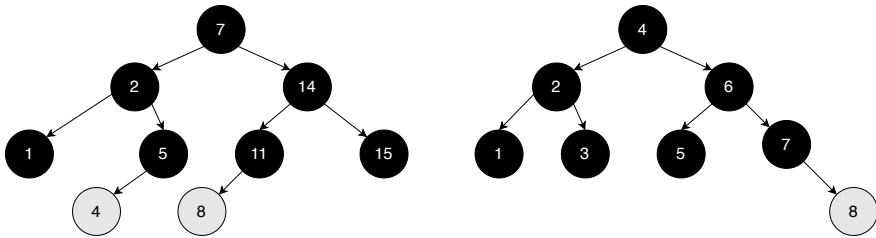


图 4.7: 插入产生的红黑树

算法自顶向下递归地插入、修复,对于高度为  $h$  的树,其复杂度为  $O(h)$ 。由于我们始终维护红黑树的颜色性质, $h$  和节点个数  $n$  呈对数关系。插入算法的复杂度为  $O(\lg n)$ 。

## 练习 4.4

4.4.1. 不使用模式匹配,分别检查四种情况实现 *insert* 算法。

答案: 366 页

## 4.4 删除

红黑树的删除比插入复杂。也可以通过模式匹配和递归简化删除<sup>2</sup>的实现。我们还可以利用其它方式来达到删除的效果。例如,一次性构建一棵树,用于后继的多次查找<sup>[5]</sup>。删除时在节点上加一个标记,当带有标记的节点超过 50%时,用未标记的节点重建一棵树。删除也会破坏红黑树的性质,需要进行修复。问题只发生在删除黑色节点时违反性质 5,使得某一路径上的黑色节点数目少于其它的路径。我们引入“双重黑色”<sup>[4]</sup>,290 页)节点来恢复第五条性质。一个这样的节点算作两个黑色节点。删除黑色节点  $x$  时,我们将黑色向上移动到父节点,或者向下移动到子树上。令接受黑色的节点为  $y$ 。如果  $y$  原来是红色,将其变为黑色;如果  $y$  原来是黑色,则变为“双重黑色”,记作  $B^2$ 。下面的例子程序增加了双重黑色的定义:

<sup>2</sup>实际上通过重用不变的部分重新构建了树。这一特性称作 *persist*

```

data Color = R | B | BB
data RBTREE a = Empty | BBEEmpty | Node Color (RBTREE a) a (RBTREE a)

```

由于空节点(NIL)是黑色,当将黑色移动到空节点时,其变为“双重黑色”空节点(BBEEmpty 或加粗的  $\emptyset$ )。删除时,第一步和普通二叉搜索树相同;如果被删除节点是黑色的,接下来进行修复:

$$\text{delete } x = \text{makeBlack} \circ \text{del } x \quad (4.6)$$

这一定义是柯里化的。如果树中只有一个元素,删除后它变为空。为了处理这一情况,我们需要修改 *makeBlack* 的定义如下:

$$\begin{aligned} \text{makeBlack } \emptyset &= \emptyset \\ \text{makeBlack } (C, l, k, r) &= (B, l, k, r) \end{aligned} \quad (4.7)$$

*del* 接受要删除的元素  $x$  和一棵树:

$$\text{del } x \emptyset = \emptyset$$

$$\text{del } x (C, l, k, r) = \begin{cases} x < k: & \text{fixB}^2(C, (\text{del } x \ l), k, r) \\ x > k: & \text{fixB}^2(C, l, k, (\text{del } x \ r)) \\ x = k: & \begin{cases} l = \emptyset: & \text{if } C = B \text{ then } \text{shiftB } r \text{ else } r \\ r = \emptyset: & \text{if } C = B \text{ then } \text{shiftB } l \text{ else } l \\ \text{否则}: & \text{fixB}^2(C, l, m, (\text{del } m \ r)), \text{ 其中: } m = \min(r) \end{cases} \end{cases} \quad (4.8)$$

如果树为空,结果为  $\emptyset$ ;否则我们比较  $x$  和树中的  $k$ ,如果  $x < k$ ,我们递归地从左侧分支删除  $x$ ;如果  $x > k$ ,则从右侧删除。由于递归结果中可能含有双重黑色节点,需要用 *fixB*<sup>2</sup> 修复。当  $x = k$  时,我们定位到了要删除的节点。如果任一子树为空,用另一棵替换掉当前节点。如果节点是黑色的,还需要将黑色移动到子树中。如果两棵子树都不为空,我们将右子树中的最小值  $m = \min r$  切下,用  $m$  替换  $k$ 。为了保持黑色节点个数, *shiftB* 将红色节点变为黑色,将黑色节点变为双重黑色。如果再次应用到双重黑色节点上,则变回黑色。

$$\begin{aligned} \text{shiftB } (B, l, k, r) &= (B^2, l, k, r) \\ \text{shiftB } (C, l, k, r) &= (B, l, k, r) \\ \text{shiftB } \emptyset &= \emptyset \\ \text{shiftB } \emptyset &= \emptyset \end{aligned} \quad (4.9)$$

下面是相应的例子程序(不包含双重黑色的修复部分):

```

delete x = makeBlack o (del x) where
  del x Empty = Empty
  del x (Node color l k r)
    | x < k = fixDB color (del x l) k r
    | x > k = fixDB color l k (del x r)

```

```

| isEmpty l = if color == B then shiftBlack r else r
| isEmpty r = if color == B then shiftBlack l else l
| otherwise = fixDB color l m (del m r) where m = min r
makeBlack (Node _ l k r) = Node B l k r
makeBlack _ = Empty

```

```
isEmpty Empty = True
```

```
isEmpty _ = False
```

```
shiftBlack (Node B l k r) = Node BB l k r
```

```
shiftBlack (Node _ l k r) = Node B l k r
```

```
shiftBlack Empty = BBEmpty
```

```
shiftBlack BBEmpty = Empty
```

函数  $fixB^2$  通过旋转操作和重新染色消除双重黑色。双重黑色节点既可能是分枝节点,也可能是空节点  $\emptyset$ 。有三种情况:

**情况 1:** 双重黑色的兄弟节点为黑色, 并且该兄弟节点有一个红色子节点。可以通过旋转修复这种情况。共有四种子情况, 全部可以变换到一种统一形式。如图 A.1 所示。

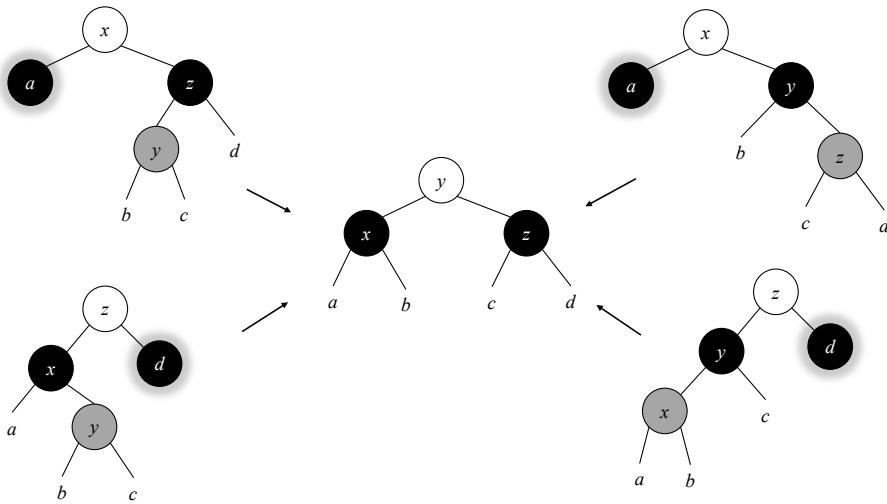


图 4.8: 4 种子情况可以修复为统一的形式

$$\begin{aligned}
 fixB^2 C_{a_{B^2}} x (\mathcal{B}, (\mathcal{R}, b, y, c), z, d) &= (C, (\mathcal{B}, shiftB(a), x, b), y, (\mathcal{B}, c, z, d)) \\
 fixB^2 C_{a_{B^2}} x (\mathcal{B}, b, y, (\mathcal{R}, c, z, d)) &= (C, (\mathcal{B}, shiftB(a), x, b), y, (\mathcal{B}, c, z, d)) \\
 fixB^2 C (\mathcal{B}, a, x, (\mathcal{R}, b, y, c)) z_{d_{B^2}} &= (C, (\mathcal{B}, a, x, b), y, (\mathcal{B}, c, z, shiftB(d))) \\
 fixB^2 C (\mathcal{B}, (\mathcal{R}, a, x, b), y, c) z_{d_{B^2}} &= (C, (\mathcal{B}, a, x, b), y, (\mathcal{B}, c, z, shiftB(d)))
 \end{aligned}
 \tag{4.10}$$

其中  $a_{B^2}$  表示节点  $a$  是双重黑色。

**情况 2:** 双重黑色节点的兄弟节点为红色。可以通过旋转, 将其变换为情况 1 或 3。如图 A.2 所示。我们在式 (4.10) 的基础上增加情况 2 的修复:

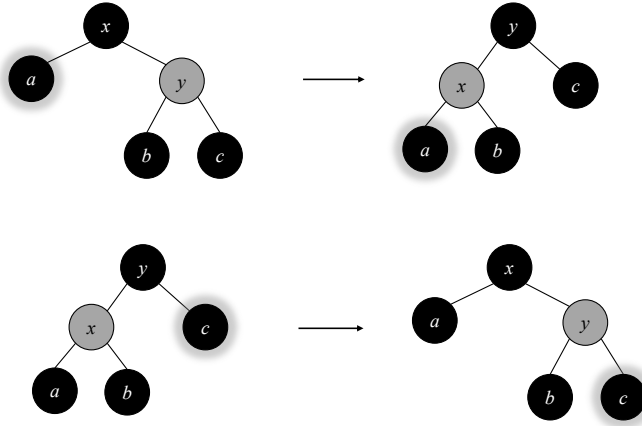


图 4.9: 双重黑色节点的兄弟节点为红色

$$\begin{aligned}
 & \dots \\
 & \text{fix}B^2 \mathcal{B} a_{\mathcal{B}^2} x (\mathcal{R}, b, y, c) = \text{fix}B^2 \mathcal{B} (\text{fix}B^2 \mathcal{R} a x b) y c \quad (4.11) \\
 & \text{fix}B^2 \mathcal{B} (\mathcal{R}, a, x, b) y c_{\mathcal{B}^2} = \text{fix}B^2 \mathcal{B} a x (\text{fix}B^2 \mathcal{R} b y c)
 \end{aligned}$$

情况 3: 双重黑色的兄弟节点, 该兄弟节点的两个子节点都是黑色。这种情况下, 我们将兄弟节点染成红色, 将双重黑色变回黑色, 然后将双重黑色属性向上传递一层到父节点。如图 A.3 所示。有两种对称情况: 对于上方的情况, 如果  $x$  是红色, 则变为黑色, 否则变为双重黑色; 对于下方的情况,  $y$  的变化与此类似。我们继续在式 (4.11) 的基础上增加此种修复:

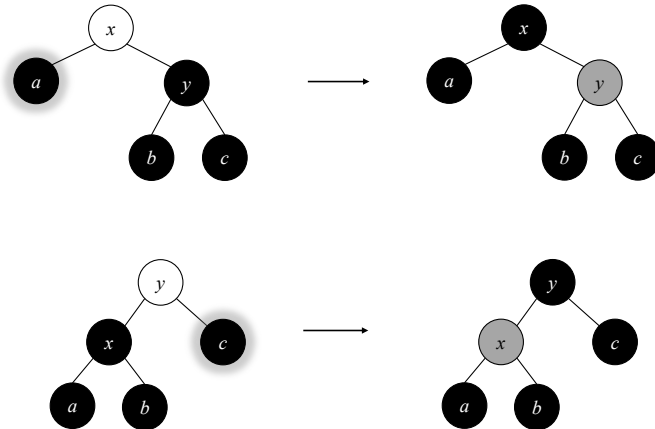


图 4.10: 将双重黑色向上传递

$$\begin{aligned}
 & \dots \\
 & \text{fix}B^2 \mathcal{C} a_{\mathcal{B}^2} x (\mathcal{B}, b, y, c) = \text{shift}B (\mathcal{C}, (\text{shift}B a), x, (\mathcal{R}, b, y, c)) \quad (4.12) \\
 & \text{fix}B^2 \mathcal{C} (\mathcal{B}, a, x, b) y c_{\mathcal{B}^2} = \text{shift}B (\mathcal{C}, (\mathcal{R}, a, x, b), y, (\text{shift}B c)) \\
 & \text{fix}B^2 \mathcal{C} l k r = (\mathcal{C}, l, k, r)
 \end{aligned}$$

如果没有匹配到上述三种模式,最后一行保持节点不变。双重黑色的修复是递归的。它中止于两种情况:一个是**情况 1**,双重黑色节点被消除了;另外一个双重黑色向上移动,直到根节点,并最终恢复为黑色。下面的例子程序将以上情况汇总到一起:

```
fixDB color a@(Node BB _ _ _) x (Node B (Node R b y c) z d)
    = Node color (Node B (shiftBlack a) x b) y (Node B c z d)
fixDB color BBEEmpty x (Node B (Node R b y c) z d)
    = Node color (Node B Empty x b) y (Node B c z d)
fixDB color a@(Node BB _ _ _) x (Node B b y (Node R c z d))
    = Node color (Node B (shiftBlack a) x b) y (Node B c z d)
fixDB color BBEEmpty x (Node B b y (Node R c z d))
    = Node color (Node B Empty x b) y (Node B c z d)
fixDB color (Node B a x (Node R b y c)) z d@(Node BB _ _ _)
    = Node color (Node B a x b) y (Node B c z (shiftBlack d))
fixDB color (Node B a x (Node R b y c)) z BBEEmpty
    = Node color (Node B a x b) y (Node B c z Empty)
fixDB color (Node B (Node R a x b) y c) z d@(Node BB _ _ _)
    = Node color (Node B a x b) y (Node B c z (shiftBlack d))
fixDB color (Node B (Node R a x b) y c) z BBEEmpty
    = Node color (Node B a x b) y (Node B c z Empty)
fixDB B a@(Node BB _ _ _) x (Node R b y c)
    = fixDB B (fixDB R a x b) y c
fixDB B a@BBEEmpty x (Node R b y c)
    = fixDB B (fixDB R a x b) y c
fixDB B (Node R a x b) y c@(Node BB _ _ _)
    = fixDB B a x (fixDB R b y c)
fixDB B (Node R a x b) y c@BBEEmpty
    = fixDB B a x (fixDB R b y c)
fixDB color a@(Node BB _ _ _) x (Node B b y c)
    = shiftBlack (Node color (shiftBlack a) x (Node R b y c))
fixDB color BBEEmpty x (Node B b y c)
    = shiftBlack (Node color Empty x (Node R b y c))
fixDB color (Node B a x b) y c@(Node BB _ _ _)
    = shiftBlack (Node color (Node R a x b) y (shiftBlack c))
fixDB color (Node B a x b) y BBEEmpty
    = shiftBlack (Node color (Node R a x b) y Empty)
fixDB color l k r = Node color l k r
```

删除算法的复杂度为  $O(h)$ , 其中  $h$  为树的高度。由于红黑树保持平衡性, 对于  $n$  个节点的树,  $h = O(\lg n)$ 。

## 练习 4.5

4.5.1. 实现“标记——重建”删除算法: 标记被删除的节点, 但不进行真正的移除。当被标记的节点数目超过 50% 时重建树。

答案: 366 页

## 4.5 命令式算法 \*

通过模式匹配和递归,我们简化了红黑树的实现。为了完整,我们给出命令式的实现。插入算法的第一步和二叉搜索树相同,接下来通过旋转操作修复平衡。

```

1: function INSERT( $T, k$ )
2:    $root \leftarrow T$ 
3:    $x \leftarrow \text{CREATE-LEAF}(k)$ 
4:    $\text{COLOR}(x) \leftarrow \text{RED}$ 
5:    $p \leftarrow \text{NIL}$ 
6:   while  $T \neq \text{NIL}$  do
7:      $p \leftarrow T$ 
8:     if  $k < \text{KEY}(T)$  then
9:        $T \leftarrow \text{LEFT}(T)$ 
10:    else
11:       $T \leftarrow \text{RIGHT}(T)$ 
12:    $\text{PARENT}(x) \leftarrow p$ 
13:   if  $p = \text{NIL}$  then ▷ 树  $T$  为空
14:     return  $x$ 
15:   else if  $k < \text{KEY}(p)$  then
16:      $\text{LEFT}(p) \leftarrow x$ 
17:   else
18:      $\text{RIGHT}(p) \leftarrow x$ 
19:   return INSERT-FIX( $root, x$ )

```

新节点为红色,接下来修复平衡。共有 3 种基本情况,每种都有左右对称的情况,总计 6 种情况。其中有两种可以合并,它们都有红色的“叔父”节点,我们可将父节点和叔父节点都变为黑色,将祖父节点变为红色:

```

1: function INSERT-FIX( $T, x$ )
2:   while  $\text{PARENT}(x) \neq \text{NIL}$  and  $\text{COLOR}(\text{PARENT}(x)) = \text{RED}$  do
3:     if  $\text{COLOR}(\text{UNCLE}(x)) = \text{RED}$  then ▷ 情况 1:  $x$  的叔父节点是红色
4:        $\text{COLOR}(\text{PARENT}(x)) \leftarrow \text{BLACK}$ 
5:        $\text{COLOR}(\text{GRAND-PARENT}(x)) \leftarrow \text{RED}$ 
6:        $\text{COLOR}(\text{UNCLE}(x)) \leftarrow \text{BLACK}$ 
7:        $x \leftarrow \text{GRAND-PARENT}(x)$ 
8:     else ▷  $x$  的叔父节点是黑色
9:       if  $\text{PARENT}(x) = \text{LEFT}(\text{GRAND-PARENT}(x))$  then
10:        if  $x = \text{RIGHT}(\text{PARENT}(x))$  then ▷ 情况 2:  $x$  是右子树
11:           $x \leftarrow \text{PARENT}(x)$ 

```



```

12:          T ← LEFT-ROTATE(T, x)
                                                    ▷ 情况 3: x 是左子树
13:          COLOR(PARENT(x)) ← BLACK
14:          COLOR(GRAND-PARENT(x)) ← RED
15:          T ← RIGHT-ROTATE(T, GRAND-PARENT(x))
16:      else
17:          if x = LEFT(PARENT(x)) then
                                                    ▷ 情况 2 的对称
18:              x ← PARENT(x)
19:              T ← RIGHT-ROTATE(T, x)
                                                    ▷ 情况 3 的对称
20:          COLOR(PARENT(x)) ← BLACK
21:          COLOR(GRAND-PARENT(x)) ← RED
22:          T ← LEFT-ROTATE(T, GRAND-PARENT(x))
23:      COLOR(T) ← BLACK
24:      return T

```

插入算法的复杂度为  $O(\lg n)$ , 其中  $n$  是节点数。和 *balance* 函数对比, 它们的处理逻辑并不相同。即使输入相同序列, 也会构造出不同的红黑树。图 4.11 给出了两棵红黑树, 它们是使用和图 4.7 中相同的序列构造出的。使用模式匹配的函数式算法存在一些性能损失<sup>[13]</sup>。

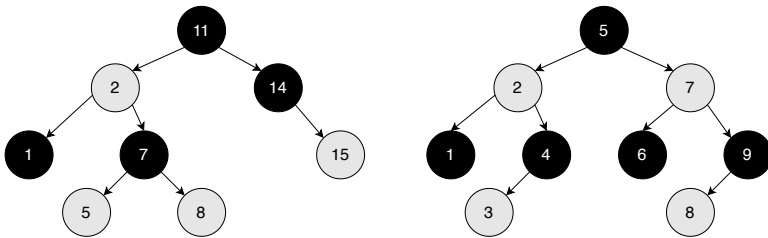


图 4.11: 命令式算法构建出的红黑树

命令式删除更为复杂, 参见本书附录 A。红黑树是一种常见的自平衡二叉树。可看作其它复杂数据结构的基础: 将分枝扩展到  $k$  个并保持平衡, 就演化到 B 树。将数据存储存储在边上而非节点中, 就演化出基数树。在红黑树的实现中, 为了修复平衡性, 需要处理多种情况。冈崎给出了一种简化方法, 并激发了多种类似的实现<sup>[16]</sup>。本书中的 AVL 树、Splay 树都是基于模式匹配实现的。

## 4.6 附录:例子程序

带有父节点引用的红黑树定义, 默认节点为红色。

```

data Node<T> {
    T key

```

```

Color color
Node<T> left, right, parent

Node(T x) = Node(null, x, null, Color.RED)

Node(Node<T> l, T k, Node<T> r, Color c) {
    left = l, key = k, right = r, color = c
    if left ≠ null then left.parent = this
    if right ≠ null then right.parent = this
}

Self setLeft(l) {
    left = l
    if l ≠ null then l.parent = this
}

Self setRight(r) {
    right = r
    if r ≠ null then r.parent = this
}

Node<T> sibling() = if parent.left == this then parent.right
                  else parent.left

Node<T> uncle() = parent.sibling()

Node<T> grandparent() = parent.parent
}

```

红黑树的插入:

```

Node<T> insert(Node<T> t, T key) {
    root = t
    x = Node(key)
    parent = null
    while (t ≠ null) {
        parent = t
        t = if (key < t.key) then t.left else t.right
    }
    if (parent == null) { //tree is empty
        root = x
    } else if (key < parent.key) {
        parent.setLeft(x)
    } else {
        parent.setRight(x)
    }
    return insertFix(root, x)
}

```

插入后的平衡修复:

```

// Fix the red→red violation
Node<T> insertFix(Node<T> t, Node<T> x) {
    while (x.parent ≠ null and x.parent.color == Color.RED) {

```

```

if (x.uncle().color == Color.RED) {
    // case 1: ((a:R x:R b) y:B c:R) ==> ((a:R x:B b) y:R c:B)
    x.parent.color = Color.BLACK
    x.grandparent().color = Color.RED
    x.uncle().color = Color.BLACK
    x = x.grandparent()
} else {
    if (x.parent == x.grandparent().left) {
        if (x == x.parent.right) {
            // case 2: ((a x:R b:R) y:B c) ==> case 3
            x = x.parent
            t = leftRotate(t, x)
        }
        // case 3: ((a:R x:R b) y:B c) ==> (a:R x:B (b y:R c))
        x.parent.color = Color.BLACK
        x.grandparent().color = Color.RED
        t = rightRotate(t, x.grandparent())
    } else {
        if (x == x.parent.left) {
            // case 2': (a x:B (b:R y:R c)) ==> case 3'
            x = x.parent
            t = rightRotate(t, x)
        }
        // case 3': (a x:B (b y:R c:R)) ==> ((a x:R b) y:B c:R)
        x.parent.color = Color.BLACK
        x.grandparent().color = Color.RED
        t = leftRotate(t, x.grandparent())
    }
}
}
t.color = Color.BLACK
return t
}

```



# 第五章 AVL 树

为了解决平衡问题,红黑树限制在某一路径上的节点数。AVL 树采用了更为直接方法:度量分枝间的差异。对于树  $T$ , 定义:

$$\delta(T) = |r| - |l| \tag{5.1}$$

其中  $|T|$  表示树  $T$  的高度,  $l, r$  为左右子树。定义空树  $\delta(\emptyset) = 0$ 。如果每棵子树  $T$  都有  $\delta(T) = 0$ , 则树是完全平衡的。例如, 一棵高度为  $h$  的完全二叉树有  $n = 2^h - 1$  个节点。除了叶子节点外, 所有节点都不为空。  $\delta(T)$  的绝对值越小, 树越平衡。我们称  $\delta(T)$  为二叉树的“平衡因子”。

## 5.1 定义

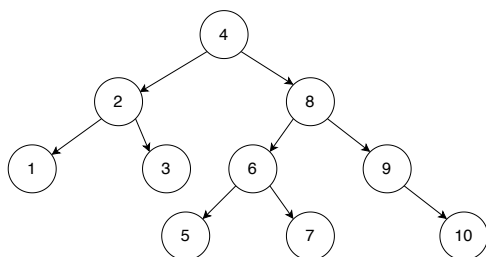


图 5.1: AVL 树

一棵二叉搜索树称为 AVL 树, 如果所有子树  $T$  都满足如下条件:

$$|\delta(T)| \leq 1 \tag{5.2}$$

平衡因子  $\delta(T)$  只能是  $\pm 1, 0$ 。图 5.1 给出了一棵 AVL 树的例子。如果树中有  $n$  个节点, 这一定义保证了树的高度  $h = O(\lg n)$ 。我们可以证明这个结论。一棵高为  $h$  的 AVL 树, 其节点数目并不是一个固定的值。当它是完全二叉树时, 含有的节点数目最多, 为  $2^h - 1$ 。我们关心它至少包含多少节点。定义  $N(h)$  代表高度为  $h$  的 AVL 树的最小节点数目。我们有:

- 空树  $\emptyset: h = 0, N(0) = 0$ ;

- 只有一个节点的树:  $h = 1, N(1) = 1$ ;

图 5.2 中给出了一个高度为  $h$  的 AVL 树  $T$ 。它包含三部分: 元素  $k$  和左右子树  $l$ 、 $r$ 。树的高度  $h$  和子树高度之间满足下面的关系:

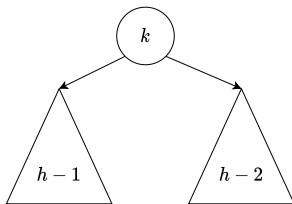


图 5.2: 高度为  $h$  的 AVL 树, 其中一棵子树高  $h - 1$ , 另外一棵的高度不小于  $h - 2$

$$h = \max(|l|, |r|) + 1 \quad (5.3)$$

因此必然存在一个子树的高度为  $h - 1$ 。根据 AVL 树的定义, 有  $||l| - |r|| \leq 1$ 。所以另外一棵子树的高度不会小于  $h - 2$ 。而  $T$  所包含的节点数为两个子树的节点数和加 1 (节点  $T$  本身):

$$N(h) = N(h - 1) + N(h - 2) + 1 \quad (5.4)$$

这一递归形式让我们联想起斐波那契数列。如果定义  $N'(h) = N(h) + 1$ , 我们就可以将式 (5.4) 转换成斐波那契数列的递归关系:

$$N'(h) = N'(h - 1) + N'(h - 2) \quad (5.5)$$

**引理 5.1.1.** 若  $N(h)$  表示高为  $h$  的 AVL 树的节点数目最小值, 令  $N'(h) = N(h) + 1$ , 则:

$$N'(h) \geq \phi^h \quad (5.6)$$

其中  $\phi = \frac{\sqrt{5} + 1}{2}$ , 称为黄金分割比。

**证明.** 使用数学归纳法。当  $h = 0$  或  $1$  时:

- $h = 0, N'(0) = 1 \geq \phi^0 = 1$
- $h = 1, N'(1) = 2 \geq \phi^1 = 1.618\dots$

对于递推情况, 设  $N'(h) \geq \phi^h$ 。

$$\begin{aligned} N'(h + 1) &= N'(h) + N'(h - 1) && \{\text{斐波那契}\} \\ &\geq \phi^h + \phi^{h-1} && \{\text{递推假设}\} \\ &= \phi^{h-1}(\phi + 1) && \{\phi + 1 = \phi^2 = \frac{\sqrt{5} + 3}{2}\} \\ &= \phi^{h+1} \end{aligned}$$

□

由定理 5.1.1, 我们立即得到下面的结果:

$$h \leq \log_{\phi}(n+1) = \log_{\phi} 2 \cdot \lg(n+1) \approx 1.44 \lg(n+1) \quad (5.7)$$

这一不等式说明 AVL 树的高度为  $O(\lg n)$ , 从而保证了平衡性。插入和删除会改变树的结构, 导致平衡因子的绝对值超出 1, 需要通过修复使得  $|\delta| < 1$ 。传统的修复方法是树旋转。我们给出一种基于模式匹配的方法简化实现。思路类似于函数式的红黑树<sup>[13]</sup>。由于这种“改变——恢复”的策略, AVL 树也是一种自平衡二叉树。我们复用二叉搜索树的定义, 尽管平衡因子  $\delta$  可以递归地求出, 为了方便, 我们在每个非空节点  $T = (l, k, r, \delta)$  中保存平衡因子的值, 并在改变树结构时更新它<sup>1</sup>。下面的例子程序增加了一个整型变量  $\delta$ :

```
data AVLTree a = Empty | Br (AVLTree a) a (AVLTree a) Int
```

AVL 树的 *lookup*、*max*、*min* 等操作和二叉搜索树相同, 而插入和删除操作是特殊的。

## 5.2 插入

向 AVL 树中插入一个新元素时, 平衡因子的绝对值  $|\delta(T)|$  可能超过 1。我们用类似红黑树修复的模式匹配方法恢复平衡。插入元素  $x$  后, 包含它的子树高度最多增加 1。我们需要沿着插入路径递归地更新平衡因子。定义插入结果为一对值  $(T', \Delta H)$ , 其中  $T'$  为插入后的树,  $\Delta H$  为高度的增加值。我们将二叉搜索树的插入算法修改如下:

$$\text{insert } x = \text{fst} \circ \text{ins } x \quad (5.8)$$

其中  $\text{fst}(a, b) = a$  返回一对值中的第一个。 $\text{ins } x T$  将元素  $x$  插入到树  $T$  中:

$$\begin{aligned} \text{ins } x \ \emptyset &= ((\emptyset, x, \emptyset, 0), 1) \\ \text{ins } x \ (l, k, r, \delta) &= \begin{cases} x < k : \text{tree}(\text{ins } x \ l) \ k \ (r, 0) \ \delta \\ x > k : \text{tree}(l, 0) \ k \ (\text{ins } x \ r) \ \delta \end{cases} \end{aligned} \quad (5.9)$$

如果树为空  $\emptyset$ , 结果为包含  $x$  的叶子节点, 平衡因子为 0, 高度增加 1。否则, 令  $T = (l, k, r, \delta)$ 。我们比较  $x$  和  $k$ , 如果  $x < k$ , 我们递归地将  $x$  插入到左子树  $l$  中, 否则插入右子树  $r$ 。递归插入的结果也是一对值  $(l', \Delta l)$  或  $(r', \Delta r)$ 。我们通过函数 *tree* 调整平衡因子并更新高度, 它接受 4 个参数:  $(l', \Delta l)$ 、 $k$ 、 $(r', \Delta r)$ 、 $\delta$ , 并产生结果  $(T', \Delta H)$ 。其中  $T'$  为新树,  $\Delta H$  定义如下:

$$\Delta H = |T'| - |T| \quad (5.10)$$

它可以进一步分解为 4 种情况:

<sup>1</sup>也可以保存树的高度而非  $\delta$ <sup>[20]</sup>。

$$\begin{aligned}
\Delta H &= |T'| - |T| \\
&= 1 + \max(|r'|, |l'|) - (1 + \max(|r|, |l|)) \\
&= \max(|r'|, |l'|) - \max(|r|, |l|) \\
&= \begin{cases} \delta \geq 0, \delta' \geq 0: & \Delta r \\ \delta \leq 0, \delta' \geq 0: & \delta + \Delta r \\ \delta \geq 0, \delta' \leq 0: & \Delta l - \delta \\ \text{否则}: & \Delta l \end{cases} \quad (5.11)
\end{aligned}$$

其中  $\delta' = \delta(T') = |r'| - |l'|$ , 是变化后的平衡因子。附录 B 给出了相关证明。在平衡调整前, 还需要确定新的平衡因子  $\delta'$ 。

$$\begin{aligned}
\delta' &= |r'| - |l'| \\
&= |r| + \Delta r - (|l| + \Delta l) \\
&= |r| - |l| + \Delta r - \Delta l \\
&= \delta + \Delta r - \Delta l
\end{aligned} \quad (5.12)$$

使用树的高度变化和平衡因子, 就可进一步定义式 (5.9) 中的函数 *tree*:

$$tree(l', \Delta l) k (r', \Delta r) \delta = balance(l', k, r', \delta') \Delta H \quad (5.13)$$

下面的例子程序实现了目前给出的结论:

```

insert x = fst o ins x where
  ins x Empty = (Br Empty x Empty 0, 1)
  ins x (Br l k r d)
    | x < k = tree (ins x l) k (r, 0) d
    | x > k = tree (l, 0) k (ins x r) d

tree (l, dl) k (r, dr) d = balance (Br l k r d') deltaH where
  d' = d + dr - dl
  deltaH | d ≥ 0 && d' ≥ 0 = dr
         | d ≤ 0 && d' ≥ 0 = d + dr
         | d ≥ 0 && d' ≤ 0 = dl - d
         | otherwise = dl

```

### 5.2.1 平衡调整

共有 4 种情况需要修复, 如图 5.3 所示。平衡因子为  $\pm 2$  超出了  $[-1, 1]$  范围。我们将其统一调整为图中心的结构, 使得  $\delta(y) = 0$ 。

我们称这 4 种情况为左-左、右-右、右-左、左-右。记调整前的平衡因子为  $\delta(x)$ 、 $\delta(y)$ 、 $\delta(z)$ ; 调整后的平衡因子为  $\delta'(x)$ 、 $\delta'(y) = 0$ 、 $\delta'(z)$ 。它们的关系如下。附录 B 给出了证明。

左-左          右-右



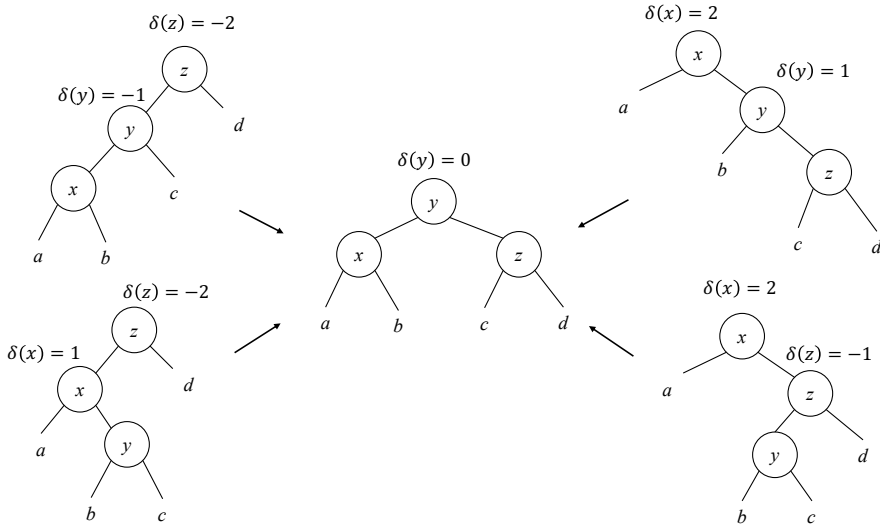


图 5.3: 将 4 种情况修复为统一形式

$$\begin{cases} \delta'(x) = \delta(x) \\ \delta'(y) = 0 \\ \delta'(z) = 0 \end{cases} \quad \begin{cases} \delta'(x) = 0 \\ \delta'(y) = 0 \\ \delta'(z) = \delta(z) \end{cases} \quad (5.14)$$

右-左和左-右相同:

$$\begin{aligned} \delta'(x) &= \begin{cases} \delta(y) = 1 : & -1 \\ \text{否则} : & 0 \end{cases} \\ \delta'(y) &= 0 \\ \delta'(z) &= \begin{cases} \delta(y) = -1 : & 1 \\ \text{否则} : & 0 \end{cases} \end{aligned} \quad (5.15)$$

利用模式匹配, 定义修复为:

$$\begin{aligned} \text{balance } ((a, x, b, \delta(x)), y, c, -1), z, d, -2) \Delta H &= ((a, x, b, \delta(x)), y, (c, z, d, 0), 0, \Delta H - 1) \\ \text{balance } (a, x, (b, y, c, z, d, \delta(z)), 1), 2) \Delta H &= ((a, x, b, 0), y, (c, z, d, \delta(z)), 0, \Delta H - 1) \\ \text{balance } ((a, x, (b, y, c, \delta(y)), 1), z, d, -2) \Delta H &= ((a, x, b, \delta'(x)), y, (c, z, d, \delta'(z)), 0, \Delta H - 1) \\ \text{balance } (a, x, ((b, y, c, \delta(y)), z, d, -1), 2) \Delta H &= ((a, x, b, \delta'(x)), y, (c, z, d, \delta'(z)), 0, \Delta H - 1) \\ \text{balance } T \Delta H &= (T, \Delta H) \end{aligned} \quad (5.16)$$

其中  $\delta'(x)$  和  $\delta'(z)$  按照式 (B.17) 定义。如果没有匹配任何模式, 则保持树不变。下面是相应的例子程序:

```
balance (Br (Br (Br a x b dx) y c (-1)) z d (-2)) dH =
    (Br (Br a x b dx) y (Br c z d 0) 0, dH-1)
```

```

balance (Br a x (Br b y (Br c z d dz) 1) 2) dH =
    (Br (Br a x b 0) y (Br c z d dz) 0, dH-1)
balance (Br (Br a x (Br b y c dy) 1) z d (-2)) dH =
    (Br (Br a x b dx') y (Br c z d dz') 0, dH-1) where
    dx' = if dy == 1 then -1 else 0
    dz' = if dy == -1 then 1 else 0
balance (Br a x (Br (Br b y c dy) z d (-1)) 2) dH =
    (Br (Br a x b dx') y (Br c z d dz') 0, dH-1) where
    dx' = if dy == 1 then -1 else 0
    dz' = if dy == -1 then 1 else 0
balance t d = (t, d)

```

插入算法的复杂度和树的高度成正比, 根据式 (5.7), 对于  $n$  个节点的树, *insert* 的复杂度为  $O(\lg n)$ 。

### 5.2.2 检验

判断一棵树是否是 AVL 树需要检查两点: (1) 是否是二叉搜索树; (2) 对于每棵子树  $T$ , 式 (5.2):  $\delta(T) \leq 1$  是否成立。下面的函数递归地检查子树间的高度差:

$$\begin{aligned}
 \text{avl? } \emptyset &= \text{True} \\
 \text{avl? } (l, k, r, \delta) &= \text{avl? } l \text{ 且 } \text{avl? } r \text{ 且 } \delta = |r| - |l| \text{ 且 } |\delta| \leq 1
 \end{aligned}
 \tag{5.17}$$

其中高度递归计算如下:

$$\begin{aligned}
 |\emptyset| &= 0 \\
 |(l, k, r, \delta)| &= 1 + \max(|r|, |l|)
 \end{aligned}
 \tag{5.18}$$

下面的例子程序实现了 AVL 树的高度检查:

```

isAVL Empty = True
isAVL (Br l _ r d) = isAVL l && isAVL r &&
    d == (height r - height l) && abs d <= 1

height Empty = 0
height (Br l _ r _) = 1 + max (height l) (height r)

```

### 练习 5.1

5.1.1. 我们只验证了 AVL 树的高度性质, 完成验证程序检查一棵二叉树是否是 AVL 树。

答案: 368 页

## 5.3 命令式算法 ★

为了完整, 本节给出 AVL 树的命令式算法。和红黑树的命令式算法相似, 我们先按二叉搜索树将新元素插入, 然后再通过旋转操作恢复平衡。

```

1: function INSERT( $T, k$ )
2:    $root \leftarrow T$ 
3:    $x \leftarrow \text{CREATE-LEAF}(k)$ 
4:    $\delta(x) \leftarrow 0$ 
5:    $parent \leftarrow \text{NIL}$ 
6:   while  $T \neq \text{NIL}$  do
7:      $parent \leftarrow T$ 
8:     if  $k < \text{KEY}(T)$  then
9:        $T \leftarrow \text{LEFT}(T)$ 
10:    else
11:       $T \leftarrow \text{RIGHT}(T)$ 
12:     $\text{PARENT}(x) \leftarrow parent$ 
13:    if  $parent = \text{NIL}$  then
14:      return  $x$ 
15:    else if  $k < \text{KEY}(parent)$  then
16:       $\text{LEFT}(parent) \leftarrow x$ 
17:    else
18:       $\text{RIGHT}(parent) \leftarrow x$ 
19:    return AVL-INSERT-FIX( $root, x$ )

```

▷ 树  $T$  为空

插入新元素后, 树的高度可能增加, 因此平衡因子  $\delta$  也会变化。插入到右侧可能使  $\delta$  增加 1, 插入左侧可能使  $\delta$  减少 1。我们从  $x$  开始, 自底向上修复平衡, 直到根节点。记新的平衡因子为  $\delta'$ , 共有 3 种情况:

- $|\delta| = 1, |\delta'| = 0$ 。插入后树处于平衡状态。父节点的高度没有变化。
- $|\delta| = 0, |\delta'| = 1$ 。左右子树之一的高度增加了, 需要继续向上检查平衡。
- $|\delta| = 1, |\delta'| = 2$ 。需要旋转以修复平衡。

```

1: function AVL-INSERT-FIX( $T, x$ )
2:   while  $\text{PARENT}(x) \neq \text{NIL}$  do
3:      $\delta \leftarrow \delta(\text{PARENT}(x))$ 
4:     if  $x = \text{LEFT}(\text{PARENT}(x))$  then
5:        $\delta' \leftarrow \delta - 1$ 
6:     else
7:        $\delta' \leftarrow \delta + 1$ 
8:      $\delta(\text{PARENT}(x)) \leftarrow \delta'$ 
9:      $P \leftarrow \text{PARENT}(x)$ 
10:     $L \leftarrow \text{LEFT}(x)$ 
11:     $R \leftarrow \text{RIGHT}(x)$ 

```



```

49:         else
50:              $\delta(P) \leftarrow 0$ 
51:             LEFT-ROTATE( $T, L$ )
52:             RIGHT-ROTATE( $T, P$ )
53:         break
54:     return  $T$ 

```

除了旋转,还需要更新平衡因子  $\delta$ 。右-右和左-左情况需要进行一次旋转;而右-左和左-右需要进行两次旋转。我们略过了 AVL 树的删除算法,附录 B 给出了删除的实现。

AVL 树是 1962 年(早于红黑树)由 Adelson-Velskii 和 Landis<sup>[18]</sup>、<sup>[19]</sup> 提出的,并以两位作者的名字命名。大多数操作的复杂度都是  $O(\lg n)$ 。式 (5.7) 使得 AVL 树的平衡性更为严格。在大量查询的情况下,其表现要好于红黑树<sup>[18]</sup>。但红黑树在频繁插入和删除的情况下性能更佳。很多程序库使用红黑树作为自平衡二叉搜索树的内部实现,AVL 树同样也可以直观、高效地解决平衡问题。

## 5.4 附录:例子程序

AVL 树的定义:

```

data Node<T> {
    int delta
    T key
    Node<T> left, right, parent
}

```

平衡修复:

```

Node<T> insertFix(Node<T> t, Node<T> x) {
    while (x.parent  $\neq$  null) {
        var (p, l, r) = (x.parent, x.parent.left, x.parent.right)
        var d1 = p.delta
        var d2 = if x == parent.left then d1 - 1 else d1 + 1
        p.delta = d2

        if abs(d1) == 1 and abs(d2) == 0 {
            return t
        } else if abs(d1) == 0 and abs(d2) == 1 {
            x = p
        } else if abs(d1) == 1 and abs(d2) == 2 {
            if d2 == 2 {
                if r.delta == 1 { //Right-right
                    p.delta = 0
                    r.delta = 0
                    t = rotateLeft(t, p)
                } else if r.delta == -1 { //Right-Left
                    var dy = r.left.delta

```

```
        p.delta = if dy == 1 then -1 else 0
        r.left.delta = 0
        r.delta = if dy == -1 then 1 else 0
        t = rotateRight(t, r)
        t = rotateLeft(t, p)
    }
} else if d2 == -2 {
    if l.delta == -1 { //Left-left
        p.delta = 0
        l.delta = 0
        t = rotateRight(t, p)
    } else if l.delta == 1 { //Left-right
        var dy = l.right.delta
        l.delta = if dy == 1 then -1 else 0
        l.right.delta = 0
        p.delta = if dy == -1 then 1 else 0
        t = rotateLeft(t, l)
        t = rotateRight(t, p)
    }
}
break
}
}
return t
}
```

## 第六章 基数树

二叉搜索树将信息存储在节点中。我们可以用边 (edge) 来携带信息么? 基数树 (Radix tree), 包括 trie、前缀树、后缀树就是根据这一思路设计出的数据结构。它们产生于 1960 年代, 被广泛用于编译器<sup>[21]</sup> 和生物信息处理 (如 DNA 模式匹配)<sup>[23]</sup> 等领域。

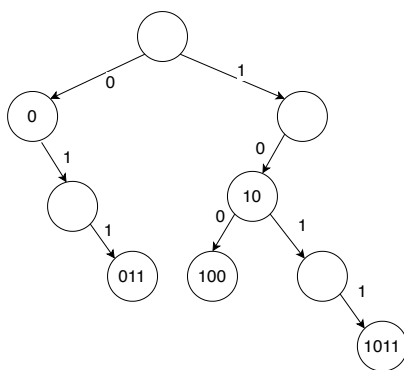


图 6.1: 基数树

图 6.1 展示了一棵基数树。它包含了二进制串 1011、10、011、100、0。如果查找二进制数  $k = (b_0b_1\dots b_n)_2$ , 我们首先检查左侧的最高位  $b_0$ 。若为 0, 则转向左子树继续查找; 若为 1, 则转向右子树。接着, 我们检查第二位, 并重复这一过程直到处理完所有的  $n$  位或到达某一叶子节点。我们并不需要在节点中存储键 (key), 这一信息由边来代表。图 6.1 标注在节点中的键仅仅是为了示意。

### 6.1 整数 trie

我们称图 6.1 所示的数据结构为 *binary trie*。Trie 是 Edward Fredkin 在 1960 年提出的。它来自英文单词 *retrieval*。Fredkin 将其读作 /'tri:/, 但其他人读作 /'traɪ/(和英文单词 *try* 的发音相同)<sup>[24]</sup>。有些情况下 trie 也被称为前缀树。本章中 trie 和前缀树分指不同的数据结构。一棵 *binary trie* 是一种特殊的二叉树, 每个键的位置由它的二进制位来决定。0 表示“向左”, 1 表示“向右”<sup>[21]</sup>。考虑图 6.2 中的 trie, “11”、“011”、“0011”代表同一个十进制整数 3。

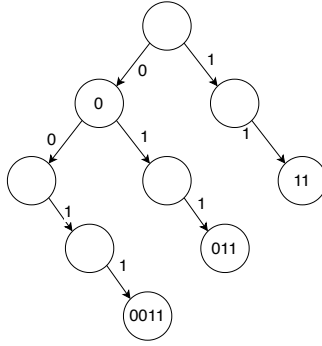


图 6.2: 大端(big-endian)trie

在 32 位整数系统中,如果把前面的 0 也当作有效位,插入 1 的结果将是一棵 32 层的树。为了解决这一问题,冈崎建议使用小端整数<sup>[21]</sup>。二进制数的最高位(MSB)通常在左边,最低位(LSB)在右边。这种形式称为大端整数,反之最高位在右边称为小端整数。使用小端,1 表示为  $(1)_2$ 、2 表示为  $(01)_2$ 、3 表示为  $(11)_2$ ……

### 6.1.1 定义

重用二叉树的定义,一个节点或者为空,或者包含左右子树和一个值(值可为空)左子树编码为 0,右子树编码为 1。

```
data IntTrie a = Empty | Branch (IntTrie a) (Maybe a) (IntTrie a)
```

对于 binary trie 中的任一节点,其对应的整数键是由节点的位置唯一确定的。因此我们无需在节点中存储键,而只需存储值。键的类型被固定为整数,如果值的类型为  $A$ ,则树的类型为  $IntTrie A$ 。

### 6.1.2 插入

当插入整数键  $k$  和值  $x$  时,我们将  $k$  转换成二进制。如果  $k$  是偶数,最低位是 0,我们递归向左子树插入;如果  $k$  是奇数,最低位是 1,我们递归向右子树插入。接下来我们将  $k$  除以 2 取整以去掉最低位。对于非空树  $T = (l, v, r)$ ,其中  $l, r$  是左右子树, $v$  是值(可为空),定义函数  $insert$  如下:

$$\begin{aligned}
 insert\ k\ x\ \emptyset &= insert\ k\ x\ (\emptyset, Nothing, \emptyset) \\
 insert\ 0\ x\ (l, v, r) &= (l, Just\ x, r) \\
 insert\ k\ x\ (l, v, r) &= \begin{cases} even(k): & (insert\ \frac{k}{2}\ x\ l, v, r) \\ odd(k): & (l, v, insert\ \lfloor \frac{k}{2} \rfloor\ x\ r) \end{cases} \quad (6.1)
 \end{aligned}$$

如果  $k = 0$ ,我们将  $x$  存入节点。如果  $k$  已经存在,这一算法覆盖以前的值。我们也可以用一个列表存储多个值,并将  $x$  添加到列表中。只要  $k \neq 0$ ,我们就根据  $k$



的奇偶性前进, 遇到  $\emptyset$  就建立一个空叶子节点 ( $\emptyset, \text{Nothing}, \emptyset$ )。图 6.3 是插入映射  $\{1 \rightarrow a, 4 \rightarrow b, 5 \rightarrow c, 9 \rightarrow d\}$  的结果。下面的例子程序实现了 *insert* 函数:

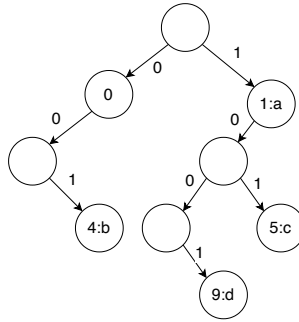


图 6.3: 小端整数 trie, 包含映射:  $\{1 \rightarrow a, 4 \rightarrow b, 5 \rightarrow c, 9 \rightarrow d\}$

```

insert k x Empty = insert k x (Branch Empty Nothing Empty)
insert 0 x (Branch l v r) = Branch l (Just x) r
insert k x (Branch l v r) | even k    = Branch (insert (k `div` 2) x l) v r
                          | otherwise = Branch l v (insert (k `div` 2) x r)
  
```

为了判定奇偶, 我们可以检查除以 2 的余数 (模 2) 是否为 0:  $even(k) = (k \bmod 2 = 0)$ , 或者使用位运算, 如:  $(k \& 0x1) == 0$ 。我们可以消除递归用循环实现插入算法:

```

1: function INSERT(T, k, x)
2:   if T = NIL then
3:     T ← EMPTY-NODE                                     ▷ (NIL, Nothing, NIL)
4:   p ← T
5:   while k ≠ 0 do
6:     if EVEN?(k) then
7:       if LEFT(p) = NIL then
8:         LEFT(p) ← EMPTY-NODE
9:         p ← LEFT(p)
10:      else
11:        if RIGHT(p) = NIL then
12:          RIGHT(p) ← EMPTY-NODE
13:          p ← RIGHT(p)
14:        k ←  $\lfloor k/2 \rfloor$ 
15:      VALUE(p) ← x
16:    return T
  
```

INSERT 接受 3 个参数: 树 *T*、要插入的键 *k* 和相应的数据 *x*。对于有 *m* 位的二进制整数 *k*, 这一算法访问 trie 中 *m* 层, 时间复杂度为  $O(m)$ 。我们将  $insert\ k\ x\ T$  和

INSERT( $T, k, x$ ) 的参数顺序设计为对称的。前者用 *foldr*、后者用 *foldl* (或 for 循环) 从若干键、值对构造树。如:

$$fromList = foldr (uncurry insert) \emptyset \quad (6.2)$$

用法为: *fromList* [(1, a), (4, b), (5, c), (9, d)], 其中 *uncurry* 为柯里化的逆运算, 它把键、值对拆开送入 *insert*, 其定义为:

$$uncurry f (a, b) = f a b \quad (6.3)$$

### 6.1.3 查找

在小端整数 trie 中查找  $k$  时, 若  $k = 0$ , 则返回根节点中存储的数据。否则根据最后一位是 0 还是 1, 对左右子树进行递归查找。

$$\begin{aligned} lookup\ k\ \emptyset &= Nothing \\ lookup\ 0\ (l, v, r) &= v \\ lookup\ k\ (l, v, r) &= \begin{cases} even(k) : lookup\ \frac{k}{2}\ l \\ odd(k) : lookup\ \lfloor \frac{k}{2} \rfloor\ r \end{cases} \end{aligned} \quad (6.4)$$

我们也可以消除递归实现迭代式的查找算法:

```

1: function LOOKUP( $T, k$ )
2:   while  $k \neq 0$  and  $T \neq \text{NIL}$  do
3:     if EVEN?( $k$ ) then
4:        $T \leftarrow \text{LEFT}(T)$ 
5:     else
6:        $T \leftarrow \text{RIGHT}(T)$ 
7:      $k \leftarrow \lfloor k/2 \rfloor$ 
8:   if  $T \neq \text{NIL}$  then
9:     return VALUE( $T$ )
10:  else
11:    return NIL

```

对于有  $m$  位的整数  $k$ , *lookup* 函数的复杂度为  $O(m)$ 。

## 练习 6.1

6.1.1. 是否可以将定义 Branch (IntTrie a) (Maybe a) (IntTrie a) 变为 Branch (IntTrie a) a (IntTrie a), 如果值不存在返回 *Nothing*, 否则返回 *Just v*?

## 6.2 整数前缀树

Trie 的缺点是空间消耗大。在图 6.3 的例子中,只有 4 个节点存有数据,其它 5 个节点都是空的。空间利用率不足 50%。为了提高空间利用率,我们可以将链接在一起的节点压缩成一个。前缀树就是这样的数据结构,由 Donald R. Morrison 在 1968 年提出。他称之为 Patricia,是 **P**raticeal **A**lgorithm **T**o **R**etrieve **I**nformation **C**oded **I**n **A**lphanumeric 的缩写<sup>[22]</sup>。当键是整数时,我们称之为整数前缀树,在不引起歧义的情况下简称为整数树。冈崎给出了整数树的实现<sup>[21]</sup>。将图 6.3 中的节点合并后,得到如图 6.4 所示的树。在整数树中,分枝节点对应的键是它所有子树的公共前缀。子树的键在公共前缀的下一位开始不同。这样整数树消除了不必要的存储空间。

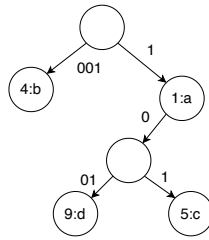


图 6.4: 小端整数树实现的映射 $\{1 \rightarrow a, 4 \rightarrow b, 5 \rightarrow c, 9 \rightarrow d\}$

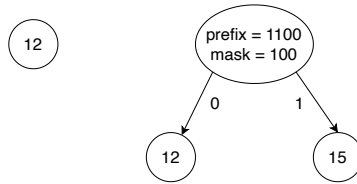
### 6.2.1 定义

整数前缀树是一种特殊的二叉树。或者为空  $\emptyset$ ,或者是叶子节点  $(k, v)$ ,包含整数键  $k$  和值  $v$ ;或者是分枝节点  $(p, m, l, r)$ ,其左右子树  $l, r$  共有个二进制**最长公共前缀**  $p$ ,左子树的下一位是 0,而右子树的下一位是 1。掩码  $m$  表明从哪一位开始分枝出子树。掩码为  $m = 2^n$  的形式,其中整数  $n \geq 0$ 。所有低于  $n$  位的二进制位都不属于公共前缀。下面的例子代码定义了整数前缀树。

```
data IntTree a = Empty
    | Leaf Int a
    | Branch Int Int (IntTree a) (IntTree a)
```

### 6.2.2 插入

当向树  $T$  插入整数  $y$  时,若  $T$  为空,我们用  $y$  创建一个叶子节点;如果  $T$  本身只包含一个叶子节点  $x$ ,我们创建一个分枝,并将两个叶子节点  $x$  和  $y$  设为子树。为了确定左右,我们需要找到  $x$  和  $y$  的最长公共前缀  $p$ 。例如,  $x = 12 = (1100)_2$ 、 $y = 15 = (1111)_2$ ,则  $p = (1100)_2$ ,其中  $o$  表示我们不关心的二进制位,我们使用掩码  $m$  来去掉(mask)这些位。在本例中,  $m = 4 = (100)_2$ ,最长公共前缀  $p$  后面的一位代表  $2^1$ 。 $x$  中这一位是 0,而  $y$  中这一位是 1。因此  $x$  是左子树,而  $y$  是右子树。如图 6.5 所示。

图 6.5: 左:  $T$  是叶子节点 12; 右: 插入 15 后

如果树  $T$  既不为空, 也不是叶子节点, 我们需要先比较  $y$  是否和  $T$  的最长公共前缀  $p$  匹配。如果匹配, 则根据下一位是 0 或 1 递归地在左、右子树中插入。例如, 若将  $y = 14 = (1110)_2$  插入图 6.5 所示的树, 由于最长公共前缀  $p = (1100)_2$ , 并且接下来的一位 ( $2^1$  位) 是 1, 我们递归地将  $y$  插入右子树。如果  $y$  和最长公共前缀  $p$  不匹配, 我们需要分出一个新叶子节点。如图 6.6 所示。

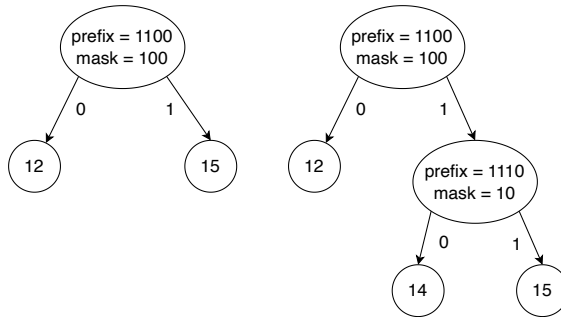
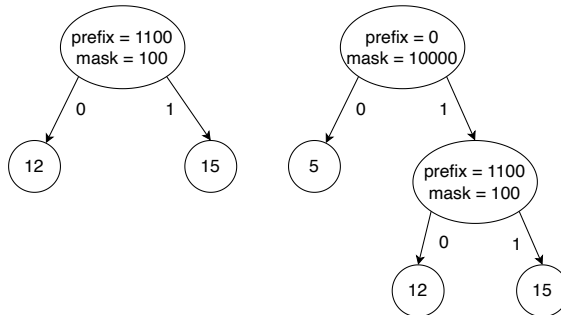
(a) 插入  $14 = (1110)_2$ , 它和最长公共前缀  $p = (1100)_2$  匹配, 递归插入右子树。(b) 插入  $5 = (101)_2$ , 它和最长公共前缀  $p = (1100)_2$  不匹配, 分出一个新叶子节点。

图 6.6: 根节点为分枝

$$\begin{aligned}
\text{insert } k \ v \ \emptyset &= (k, v) \\
\text{insert } k \ v \ (k, v') &= (k, v) \\
\text{insert } k \ v \ (k', v') &= \text{join } k \ (k, v) \ k' \ (k', v') \\
\text{insert } k \ v \ (p, m, l, r) &= \begin{cases} \text{match}(k, p, m) : & \begin{cases} \text{zero}(k, m) : & (p, m, \text{insert } k \ v \ l, r) \\ \text{otherwise} : & (p, m, l, \text{insert } k \ v \ r) \end{cases} \\ \text{否则} : & \text{join } k \ (k, v) \ p \ (p, m, l, r) \end{cases}
\end{aligned} \tag{6.5}$$

当  $T = \emptyset$  时, 创建一个叶子节点  $(k, v)$ ; 如果键相同则覆盖。函数  $\text{match}(k, p, m)$  检查整数  $k$  和最长公共前缀  $p$  在掩码  $m$  下是否相同:  $\text{mask}(k, m) = p$ , 其中  $\text{mask}(k, m) = \overline{m-1} \& k$ 。它先对  $m-1$  按位取反, 然后和  $k$  按位与。函数  $\text{zero}(k, m)$  检查掩码  $m$  之后的二进制位是 0 还是 1。我们将  $m$  向右移动 1 位, 然后和  $k$  按位与:

$$\text{zero}(k, m) = k \& (m \gg 1) \tag{6.6}$$

函数  $\text{join}(p_1, T_1, p_2, T_2)$  接受两个前缀和两棵树。从  $p_1$ 、 $p_2$  抽出最长公共前缀  $(p, m) = \text{LCP}(p_1, p_2)$ , 创建一个新分枝, 并将  $T_1$ 、 $T_2$  设为子树:

$$\text{join}(p_1, T_1, p_2, T_2) = \begin{cases} \text{zero}(p_1, m) : & (p, m, T_1, T_2) \\ \text{否则} : & (p, m, T_2, T_1) \end{cases} \tag{6.7}$$

为了计算最长公共前缀, 我们先对  $p_1$ 、 $p_2$  按位计算异或, 然后数出最高位  $h = \text{highest}(\text{xor}(p_1, p_2))$ :

$$\begin{aligned}
\text{highest}(0) &= 0 \\
\text{highest}(n) &= 1 + \text{highest}(n \gg 1)
\end{aligned}$$

接下来我们产生掩码  $m = 2^h$ 。最长公共前缀  $p$  可以用掩码  $m$  和  $p_1$ 、 $p_2$  中的任何一个得出。例如  $p = \text{mask}(p_1, m)$ 。下面的例子程序实现了  $\text{insert}$  函数:

```

insert k x t
= case t of
  Empty → Leaf k x
  Leaf k' x' → if k == k' then Leaf k x
                else join k (Leaf k x) k' t
  Branch p m l r
    | match k p m → if zero k m
                    then Branch p m (insert k x l) r
                    else Branch p m l (insert k x r)
    | otherwise → join k (Leaf k x) p t

join p1 t1 p2 t2 = if zero p1 m then Branch p m t1 t2
                  else Branch p m t2 t1

where
  (p, m) = lcp p1 p2

lcp p1 p2 = (p, m) where

```

```

    m = bit (highestBit (p1 `xor` p2))
    p = mask p1 m

highestBit x = if x == 0 then 0 else 1 + highestBit (shiftR x 1)

mask x m = x .&. complement (m - 1)

zero x m = x .&. (shiftR m 1) == 0

match k p m = (mask k m) == p

```

我们也可以使用命令式方法实现 *insert*:

```

1: function INSERT( $T, k, v$ )
2:   if  $T = \text{NIL}$  then
3:     return CREATE-LEAF( $k, v$ )
4:    $y \leftarrow T$ 
5:    $p \leftarrow \text{NIL}$ 
6:   while  $y$  is not leaf, and MATCH( $k, \text{PREFIX}(y), \text{MASK}(y)$ ) do
7:      $p \leftarrow y$ 
8:     if ZERO?( $k, \text{MASK}(y)$ ) then
9:        $y \leftarrow \text{LEFT}(y)$ 
10:    else
11:       $y \leftarrow \text{RIGHT}(y)$ 
12:    if  $y$  is leaf, and  $k = \text{KEY}(y)$  then
13:      VALUE( $y$ )  $\leftarrow v$ 
14:    else
15:       $z \leftarrow \text{BRANCH}(y, \text{CREATE-LEAF}(k, v))$ 
16:      if  $p = \text{NIL}$  then
17:         $T \leftarrow z$ 
18:      else
19:        if LEFT( $p$ ) =  $y$  then
20:          LEFT( $p$ )  $\leftarrow z$ 
21:        else
22:          RIGHT( $p$ )  $\leftarrow z$ 
23:    return  $T$ 

```

其中  $\text{BRANCH}(T_1, T_2)$  创建一个新分枝，抽出最长公共前缀，并将  $T_1$  和  $T_2$  设为子树。

```

1: function BRANCH( $T_1, T_2$ )
2:    $T \leftarrow \text{EMPTY-NODE}$ 
3:   ( $\text{PREFIX}(T), \text{MASK}(T)$ )  $\leftarrow \text{LCP}(\text{PREFIX}(T_1), \text{PREFIX}(T_2))$ 

```

```

4:   if ZERO?(PREFIX( $T_1$ ), MASK( $T$ )) then
5:     LEFT( $T$ )  $\leftarrow T_1$ 
6:     RIGHT( $T$ )  $\leftarrow T_2$ 
7:   else
8:     LEFT( $T$ )  $\leftarrow T_2$ 
9:     RIGHT( $T$ )  $\leftarrow T_1$ 
10:  return  $T$ 

```

```

11: function ZERO?( $x, m$ )
12:  return ( $x \& \lfloor \frac{m}{2} \rfloor$ ) = 0

```

函数 LCP 获取两个整数的最长公共前缀：

```

1: function LCP( $a, b$ )
2:    $d \leftarrow xor(a, b)$ 
3:    $m \leftarrow 1$ 
4:   while  $d \neq 0$  do
5:      $d \leftarrow \lfloor \frac{d}{2} \rfloor$ 
6:      $m \leftarrow 2m$ 
7:   return (MASKBIT( $a, m$ ),  $m$ )

```

```

8: function MASKBIT( $x, m$ )
9:   return  $x \& \overline{m - 1}$ 

```

图 6.7 展示了使用插入算法构造的前缀树。虽然整数树压缩了节点,但仍然需要线性扫描二进制位以获取最长前缀,对  $m$  位整数,插入算法的复杂度是  $O(m)$ 。

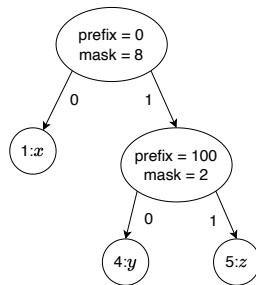


图 6.7: 插入映射  $1 \rightarrow x, 4 \rightarrow y, 5 \rightarrow z$  到大端整数前缀树

### 6.2.3 查找

当查找整数  $k$  时,如果树  $T = \emptyset$  为空,或者是一个叶子节点  $T = (k', v)$  但  $k \neq k'$ , 则  $k$  不存在;如果  $k = k'$ , 则  $v$  为查找结果。如果  $T = (p, m, l, r)$  是一个分枝节点, 我们需要检查公共前缀  $p$  和  $k$  在掩码  $m$  下是否匹配, 并根据下一位是 0/1 递归地在子树  $l$  或  $r$  中查找。如果不匹配公共前缀  $p$ , 则  $k$  不存在。

$$\begin{aligned}
 \text{lookup } k \ \emptyset &= \text{Nothing} \\
 \text{lookup } k \ (k', v) &= \begin{cases} k = k' : & \text{Just } v \\ \text{otherwise} : & \text{Nothing} \end{cases} \\
 \text{lookup } k \ (p, m, l, r) &= \begin{cases} \text{match}(k, p, m) : & \begin{cases} \text{zero}(k, m) : & \text{lookup } k \ l \\ \text{otherwise} : & \text{lookup } k \ r \end{cases} \\ \text{otherwise} : & \text{Nothing} \end{cases} \quad (6.8)
 \end{aligned}$$

我们也可以消除递归改用迭代的方式实现查找。

```

1: function LOOK-UP( $T, k$ )
2:   if  $T = \text{NIL}$  then
3:     return NIL
4:   while  $T$  is not leaf, and MATCH( $k, \text{PREFIX}(T), \text{MASK}(T)$ ) do
5:     if ZERO?( $k, \text{MASK}(T)$ ) then
6:        $T \leftarrow \text{LEFT}(T)$ 
7:     else
8:        $T \leftarrow \text{RIGHT}(T)$ 
9:   if  $T$  is leaf, and KEY( $T$ ) =  $k$  then
10:    return VALUE( $T$ )
11:  else
12:    return NIL

```

对于有  $m$  位的二进制整数,  $\text{lookup}$  算法的复杂为  $O(m)$ 。

### 练习 6.2

6.2.1. 编写程序实现整数前缀树的  $\text{lookup}$  算法。

6.2.2. 实现整数 trie 和整数树的前序遍历, 仅输出值不为空的节点键。结果有何规律?

答案: 369 页

## 6.3 Trie

我们可以把键从整数扩展到列表。其中一个特例是字符串。前缀树和 trie 可以作为文本处理的有力工具。当键从二进制 0/1 扩展到通用列表时, 树结构也自然从二叉



树扩展为多分枝树。拿英语来说,一共有 26 个字符,如果忽略大小写,分枝的个数可以达到 26,如图 6.8 所示。

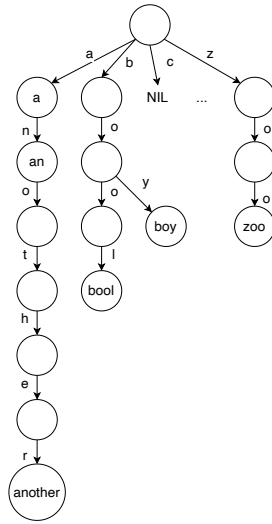


图 6.8: 含有多达 26 个分枝的 trie, 包含键: a、an、another、bool、boy、zoo

并非所有的 26 棵子树都包含数据。在图 6.8 的根节点分枝中,只有代表 a、b、z 的 3 棵子树不为空。其它分枝,例如代表 c 的子树,全部是空的。我们可以将这些空子树隐藏起来。如果区分大小写或者需要进一步将类型从字符串扩展到抽象列表,可以使用 map 等数据结构来处理动态数目的分枝。一棵类型为  $Trie\ K\ V$  的树或者为空  $\emptyset$ , 或者是一个节点,包含以下两种情况:

1. 值为  $v$  的叶子节点,不含有子树,记为  $(v, \emptyset)$ ,其中  $v$  的类型为  $V$ ;
2. 分枝节点,包含值  $v$  和多棵子树。每棵子树对应类型  $K$  中的某个值  $k$ , 记为  $(v, ts)$ 。其中  $ts = \{k_1 \mapsto T_1, k_2 \mapsto T_2, \dots, k_m \mapsto T_m\}$  包含  $k_i$  和子树  $T_i$  间的映射。可以使用关联列表或者自平衡树实现映射(见第 4、5 章)。

定义内容为空的树为:  $(Nothing, \emptyset)$ 。下面的例子程序定义了 trie:

```
data Trie k v = Trie { value :: Maybe v
                    , subTrees :: Map k (Trie k v)}
```

### 6.3.1 插入

考虑插入一对键、值到 trie 中,其中键为若干元素的列表。令 trie 为  $T = (v, ts)$ 。记  $ts[k]$  在映射  $ts$  中查找键  $k$  对应的值。如果没有找到返回空树;记  $ts[k] \leftarrow t$  在映射  $ts$  中插入键  $k$ 、树  $t$ ,如果存在则覆盖,并返回更新后的映射。

$$\begin{aligned} \text{insert } []\ v\ (v', ts) &= (\text{Just } v, ts) \\ \text{insert } (k:ks)\ v\ (v', ts) &= (v', ts[k] \leftarrow \text{insert } ks\ v\ ts[k]) \end{aligned} \quad (6.9)$$

对应的例子程序:

```

insert [] x (Trie _ ts) = Trie (Just x) ts
insert (k:ks) x (Trie v ts) = Trie v (Map.insert k (insert ks x t) ts) where
  t = case Map.lookup k ts of
    Nothing → Trie Nothing Map.empty
    (Just t) → t

```

我们也可以消除递归,实现迭代方式的 *insert* 函数:

```

1: function INSERT(T, k, v)
2:   if T = NIL then
3:     T ← EMPTY-NODE
4:   p ← T
5:   for each c in k do
6:     if SUB-TREES(p)[c] = NIL then
7:       SUB-TREES(p)[c] ← EMPTY-NODE
8:     p ← SUB-TREES(p)[c]
9:   VALUE(p) ← v
10:  return T

```

若键的类型为  $[K]$  ( $K$  的列表),  $K$  是包含  $m$  个元素的有限集,键的长度为  $n$ ,则插入算法的复杂度为  $O(n \lg m)$ 。当键是小写英文字符串时,  $m = 26$ ,插入操作的复杂度和字符串长度成正比。

### 6.3.2 查找

在  $T = (v, ts)$  中查找一个非空键 ( $k:ks$ ) 时,我们从第一个元素  $k$  开始,如果存在映射到  $k$  的子树  $T'$ ,则接下来递归地在  $T'$  中查找  $ks$ 。当键为空时,返回当前节点的值作为结果:

$$\begin{aligned}
 \text{lookup } [] (v, ts) &= v \\
 \text{lookup } (k:ks) (v, ts) &= \begin{cases} ts[k] = \text{Nothing} : \text{Nothing} \\ ts[k] = \text{Just } t : \text{lookup } ks t \end{cases} \quad (6.10)
 \end{aligned}$$

下面是相应的迭代实现:

```

1: function LOOK-UP(T, key)
2:   if T = NIL then
3:     return Nothing
4:   for each c in key do
5:     if SUB-TREES(T)[c] = NIL then
6:       return Nothing
7:     T ← SUB-TREES(T)[c]

```

8: **return** VALUE( $T$ )

查找算法的复杂度为  $O(n \lg m)$ , 其中  $n$  是键的长度,  $m$  是元素所在集合的大小。

## 6.4 前缀树

Trie 的空间利用率很低。我们可以用同样的方法压缩链接在一起的节点, 这样就得到了前缀树。前缀树节点  $t$  包含两部分: 一个可为空的值  $v$ 、零个或若干子前缀树。每棵子树  $t_i$  对应到一个列表  $s_i$ 。列表和子树的映射关系记为  $[s_i \mapsto t_i]$ 。这些列表拥有共同的前缀  $s$ , 而  $s$  映射到节点  $t$ 。也就是说  $s$  是  $s + s_1, s + s_2, \dots$  的最长公共前缀。对于任何  $i \neq j$ , 列表  $s_i, s_j$  不存在非空的公共前缀。将图 6.8 中链接在一起的节点压缩起来, 可以得到如图 6.9 的前缀树。

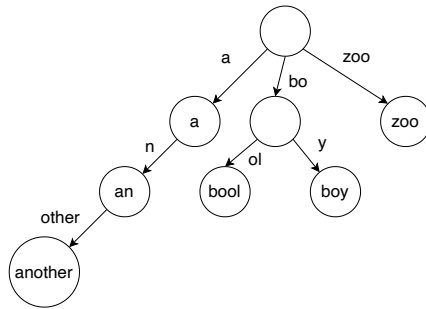


图 6.9: 一棵前缀树, 包含键: a、an、another、bool、boy、zoo

下面的例子程序定义了前缀树<sup>1</sup>:

```
data PrefixTree k v = PrefixTree { value :: Maybe v
                                , subTrees :: [[k], PrefixTree k v]}
```

我们记前缀树为  $t = (v, ts)$ 。特别地,  $(Nothing, [])$  表示内容为空的节点;  $(Just v, [])$  表示值为  $v$  的叶子节点。

### 6.4.1 插入

插入字符串  $s$  时, 若前缀树为空, 我们为  $s$  创建一个叶子节点, 如图 6.10(a) 所示。如果  $s$  和某个子树  $t_i$  相对应的  $s_i$  存在公共前缀, 我们创建一个新叶子节点  $t_j$ , 抽出公共前缀, 并将其映射到一个新分枝  $t'$  上, 然后令  $t_i, t_j$  分别为  $t'$  的两棵子树。如图 6.10(b) 所示。这里有两种特殊情况:  $s$  为  $s_i$  的前缀, 如图 6.10(c)  $\rightarrow$  (e); 以及  $s_i$  为  $s$  的前缀, 如图 6.10(d)  $\rightarrow$  (e)。

下面的函数向前缀树  $t = (v', ts)$  中插入键  $s$  和值  $v$ :

$$\begin{aligned} \text{insert } [] v (v', ts) &= (Just v, ts) \\ \text{insert } s v (v', ts) &= (v', \text{ins } ts) \end{aligned} \tag{6.11}$$

<sup>1</sup>也可以用 `Map [k] PrefixTree k v` 存储子树。

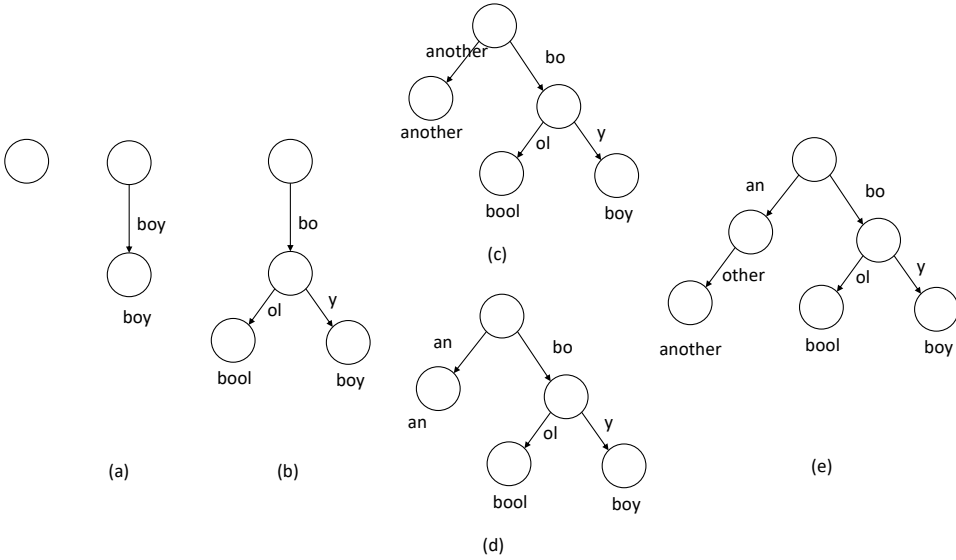


图 6.10: (a) 向空树插入 boy; (b) 插入 bool, 创建新分枝; (c) 向 (b) 插入 another (d) 向 (b) 插入 an (e) 向 (c) 插入 an, 结果与向 (d) 插入 another 相同

如果键  $s$  为空, 我们用  $v$  覆盖掉以前的值; 否则调用  $ins$  检查子树和它们的前缀。

$$\begin{aligned}
 ins [] &= [s \mapsto (Just\ v, [])] \\
 ins ((s' \mapsto t):ts') &= \begin{cases} match\ s\ s' : (branch\ s\ v\ s'\ t) : ts' & (6.12) \\ \text{否则} : (s' \mapsto t) : ins\ ts' \end{cases}
 \end{aligned}$$

如果节点不含任何子树, 我们创建一个值为  $v$  的子树, 并将  $s$  映射到其上; 否则, 对于每个子树映射  $s' \mapsto t$ , 我们比较  $s'$  和  $s$ 。如果它们有公共前缀(通过  $match$  函数检测), 则利用  $branch$  函数分枝出子树。我们定义两个列表匹配如果它们有公共前缀:

$$\begin{aligned}
 match [] B &= True \\
 match A [] &= True \\
 match (a:as) (b:bs) &= a = b
 \end{aligned} \tag{6.13}$$

我们定义函数  $(C, A', B') = lcp\ A\ B$  来提取列表  $A$  和  $B$  的最长公共前缀, 其中  $C \# A' = A$  且  $C \# B' = B$  成立。如果  $A, B$  中的任何一个为空或者它们的第一个元素不同, 则公共前缀  $C = []$ ; 否则, 我们递归地从子列表中提取公共前缀, 并将头部元素附加在前:

$$\begin{aligned}
 lcp [] B &= ([], [], B) \\
 lcp A [] &= ([], A, []) \\
 lcp (a:as) (b:bs) &= \begin{cases} a \neq b : ([], a:as, b:bs) \\ \text{否则} : (a:cs, as', bs'), \text{其中 } (cs, as', bs') = lcp\ as\ bs \end{cases}
 \end{aligned} \tag{6.14}$$

函数  $branch\ A\ v\ B\ t$  接受两个键  $A, B$ , 一个值  $v$  和树  $t$ , 它提取出  $A, B$  的最长公共前缀  $C$ , 将其映射到新分枝节点上, 并设置好子树:

$$branch\ A\ v\ B\ t = \begin{cases} (C, [], B') : (C, (Just\ v, [B' \mapsto t])) \\ (C, A', []) : (C, insert\ A'\ v\ t) \\ (C, A', B') : (C, (Nothing, [A' \mapsto (Just\ v, []), B' \mapsto t])) \end{cases} \quad (6.15)$$

如果  $A$  是  $B$  的前缀, 则将  $A$  映射到  $v$  所在的节点, 列表的剩余部分被重新映射到分枝的唯一子树  $t$ ; 如果  $B$  是  $A$  的前缀, 我们递归地将剩余列表和值插入到  $t$  中; 否则, 我们创建一个值为  $v$  的叶子节点, 将其和  $t$  作为分枝的两棵子树。下面的例子程序实现了  $insert$  算法:

```

insert [] v (PrefixTree _ ts) = PrefixTree (Just v) ts
insert k v (PrefixTree v' ts) = PrefixTree v' (ins ts) where
  ins [] = [(k, leaf v)]
  ins ((k', t) : ts) | match k k' = (branch k v k' t) : ts
                    | otherwise = (k', t) : ins ts

leaf v = PrefixTree (Just v) []

match [] _ = True
match _ [] = True
match (a:_) (b:_) = a == b

branch a v b t = case lcp a b of
  (c, [], b') → (c, PrefixTree (Just v) [(b', t)])
  (c, a', []) → (c, insert a' v t)
  (c, a', b') → (c, PrefixTree Nothing [(a', leaf v), (b', t)])

lcp [] bs = ([], [], bs)
lcp as [] = ([], as, [])
lcp (a:as) (b:bs) | a ≠ b = ([], a:as, b:bs)
                  | otherwise = (a:cs, as', bs') where
                      (cs, as', bs') = lcp as bs

```

我们也可以消除递归, 用循环实现插入算法:

```

1: function INSERT( $T, k, v$ )
2:   if  $T = \text{NIL}$  then
3:      $T \leftarrow \text{EMPTY-NODE}$ 
4:    $p \leftarrow T$ 
5:   loop
6:      $match \leftarrow \text{FALSE}$ 
7:     for each  $s_i \mapsto T_i$  in SUB-TREES( $p$ ) do
8:       if  $k = s_i$  then
9:         VALUE( $T_i$ )  $\leftarrow v$ 

```

▷ 覆盖

```

10:         return  $T$ 
11:      $c \leftarrow \text{LCP}(k, s_i)$ 
12:      $k_1 \leftarrow k - c, k_2 \leftarrow s_i - c$ 
13:     if  $c \neq \text{NIL}$  then
14:          $match \leftarrow \text{TRUE}$ 
15:         if  $k_2 = \text{NIL}$  then ▷  $s_i$  是  $k$  的前缀
16:              $p \leftarrow T_i, k \leftarrow k_1$ 
17:             break
18:         else ▷ 新分枝
19:              $\text{ADD}(\text{SUB-TREES}(p), c \mapsto \text{BRANCH}(k_1, \text{LEAF}(v), k_2, T_i))$ 
20:              $\text{DELETE}(\text{SUB-TREES}(p), s_i \mapsto T_i)$ 
21:             return  $T$ 
22:     if not  $match$  then ▷ 新叶子
23:          $\text{ADD}(\text{SUB-TREES}(p), k \mapsto \text{LEAF}(v))$ 
24:         break
25:     return  $T$ 

```

函数 LCP 提取出两个列表的公共前缀：

```

1: function LCP( $A, B$ )
2:      $i \leftarrow 1$ 
3:     while  $i \leq |A|$  and  $i \leq |B|$  and  $A[i] = B[i]$  do
4:          $i \leftarrow i + 1$ 
5:     return  $A[1 \dots i - 1]$ 

```

$\text{BRANCH}(s_1, T_1, s_2, T_2)$  中需要处理特殊情况。如果  $s_1$  为空, 说明待插入的键是某个子树的前缀, 我们将  $T_2$  设置为  $T_1$  的子树。否则, 我们创建一个新的分枝节点, 并将  $T_1, T_2$  设置为子树。

```

1: function BRANCH( $s_1, T_1, s_2, T_2$ )
2:     if  $s_1 = \text{NIL}$  then
3:          $\text{ADD}(\text{SUB-TREES}(T_1), s_2 \mapsto T_2)$ 
4:         return  $T_1$ 
5:      $T \leftarrow \text{EMPTY-NODE}$ 
6:      $\text{SUB-TREES}(T) \leftarrow \{s_1 \mapsto T_1, s_2 \mapsto T_2\}$ 
7:     return  $T$ 

```

虽然前缀树提高了空间利用率, 但其复杂度仍然是  $O(mn)$ , 其中  $n$  是键的长度,  $m$  是列表元素集合的大小。

## 6.4.2 查找

查找键  $k$  时,我们从根节点开始,如果  $k = []$  为空,则返回根节点的值;否则我们检查子树的映射,找到映射  $s_i \mapsto t_i$ ,使得  $s_i$  是  $k$  的前缀,然后再递归地在子树  $t_i$  中查找  $k - s_i$ 。如果所有的  $s_i$  都不是  $k$  的前缀,则树中不存在要查找的键。

$$\begin{aligned} \text{lookup } [] (v, ts) &= v \\ \text{lookup } k (v, ts) &= \text{find } ((s, t) \mapsto s \sqsubseteq k) ts = \\ &\begin{cases} \text{Nothing} : & \text{Nothing} \\ \text{Just } (s, t) : & \text{lookup } (k - s) t \end{cases} \end{aligned} \quad (6.16)$$

其中  $A \sqsubseteq B$  表示  $A$  是  $B$  的前缀。函数  $\text{find}$  的定义见第一章,它在列表中查找满足指定条件的元素。下面的例子程序实现了查找算法。

```
lookup [] (PrefixTree v _) = v
lookup ks (PrefixTree v ts) =
  case find (\(s, t) → s `isPrefixOf` ks) ts of
    Nothing → Nothing
    Just (s, t) → lookup (drop (length s) ks) t
```

前缀检查所需的时间和列表的长度成比例,  $\text{lookup}$  算法的复杂度为  $O(mn)$ , 其中  $m$  是列表元素集合的大小,  $n$  是列表的长度。

### 练习 6.3

6.3.1. 消除  $\text{lookup}$  算法中的递归,用循环实现前缀树的查找。

答案: 371 页

## 6.5 Trie 和前缀树的应用

我们可以用 trie 和前缀树来解决许多有趣的问题,包括词典、自动输入补齐、数字键盘输入法。与商业实现不同,本节给出的例子都是示意性的。

### 6.5.1 词典和自动补齐

如图 6.11 所示,当用户输入某些字符后,词典会搜索词库,列出候选单词。

词典通常存有数十万单词,全部查找的开销很大。商业词典软件会使用多种工程方法以提高性能,包括缓存、索引等。图 6.12 是一个带有自动补齐功能的输入框。当输入内容后,会列出一些可能的候选项。它们以用户输入的内容为前缀。

这两个例子都展示了自动补齐的功能。我们可以用前缀树来实现它。简单起见,我们在例子中限定字符为英文,候选列表不超过  $n$  个条目。一个词典保存了多个键、值对,其中键是英文单词或词组,值是对应的意思和解释。当用户输入字符串  $s$  时,我们在前缀树实现的词典中查找所有以  $s$  开头的键。如果  $s$  为空,就扩展出所有子树直

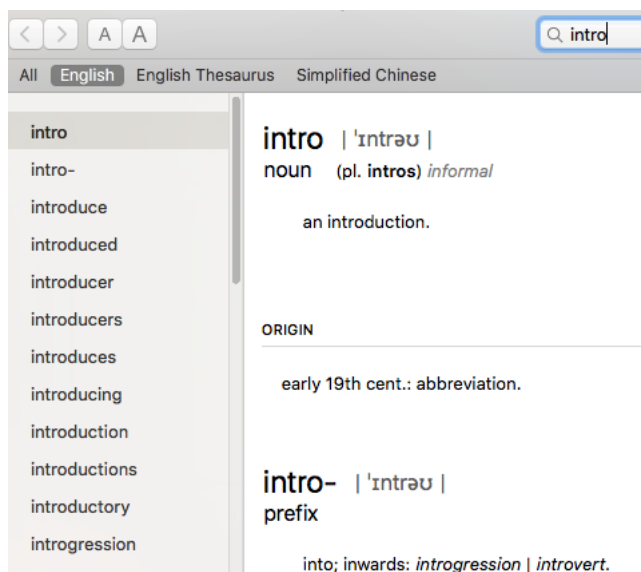


图 6.11: 词典

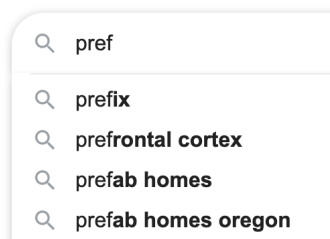


图 6.12: 带有自动补齐的输入框



到达到  $n$  条结果;否则,我们根据匹配的子树递归查找。在支持惰性求值的环境中,我们可以扩展出全部候选项,然后按需取得前  $n$  个:  $take\ n\ (startsWith\ s\ t)$ , 其中  $t$  是前缀树。

$$\begin{aligned}
 startsWith\ []\ (Nothing, ts) &= enum\ ts \\
 startsWith\ []\ (Just\ x, ts) &= ([], x) : enum\ ts \\
 startsWith\ s\ (v, ts) &= find\ ((k, t) \mapsto s \sqsubseteq k \text{ or } k \sqsubseteq s)\ ts = \quad (6.17) \\
 &\begin{cases} Nothing: & [] \\ Just\ (k, t): & [(k \# a, b) | (a, b) \in startsWith\ (s - k)\ t] \end{cases}
 \end{aligned}$$

给定一个列表  $s$ , 函数  $startsWith$  在前缀树中搜索所有以  $s$  为前缀的结果。如果  $s$  为空, 它枚举所有子树。如果根节点中的值  $x$  不为空, 则将  $([], x)$  附加到结果之前。函数  $enum\ ts$  定义如下:

$$enum = concatMap\ (k, t) \mapsto [(k \# a, b) | (a, b) \in startsWith\ []\ t] \quad (6.18)$$

其中  $concatMap$  (也称为  $flatMap$ ) 是列表计算中的一个重要概念。效果上相当于先对每个元素进行映射, 然后将结果连接起来。通常使用  $build\ foldr$  融合律来实现, 以消除计算中产生的中间结果列表 ([99] 第 5 章)。如果  $s$  不为空, 我们检查子树映射, 对于每个映射  $(k, t)$ , 如果  $s$  或  $k$  是另外一个的前缀, 我们就递归地扩展子树  $t$ , 并将  $k$  附加到每个结果的键之前; 否则, 如果  $s$  不和任何子树的映射匹配, 则不存在以  $s$  为前缀的结果。下面的例子程序实现了这一算法:

```

startsWith [] (PrefixTree Nothing ts) = enum ts
startsWith [] (PrefixTree (Just v) ts) = ([], v) : enum ts
startsWith k (PrefixTree _ ts) =
  case find (\(s, t) -> s `isPrefixOf` k || k `isPrefixOf` s) ts of
    Nothing -> []
    Just (s, t) -> [(s # a, b) |
                      (a, b) <- startsWith (drop (length s) k) t]

enum = concatMap (\(k, t) -> [(k # a, b) | (a, b) <- startsWith [] t])

```

我们也可以命令式的方式实现  $STARTS-WITH(T, k, n)$ 。从根节点开始, 我们循环检查每个子树映射  $k_i \mapsto T_i$ 。如果  $k$  是某个子树  $T_i$  的前缀, 我们就将这棵子树扩展到最多  $n$  条结果; 如果  $k_i$  是  $k$  的前缀, 我们就去掉前缀部分, 用新键  $k - k_i$  在  $T_i$  中递归查找。

- 1: **function** STARTS-WITH( $T, k, n$ )
- 2:     **if**  $T = \text{NIL}$  **then**
- 3:         **return** NIL
- 4:      $s \leftarrow \text{NIL}$
- 5:     **repeat**
- 6:          $match \leftarrow \text{FALSE}$

```

7:      for  $k_i \mapsto T_i$  in SUB-TREES( $T$ ) do
8:          if  $k$  is prefix of  $k_i$  then
9:              return EXPAND( $s \# k_i, T_i, n$ )
10:         if  $k_i$  is prefix of  $k$  then
11:             match  $\leftarrow$  TRUE
12:              $k \leftarrow k - k_i$                                 ▷ 去掉前缀
13:              $T \leftarrow T_i$ 
14:              $s \leftarrow s \# k_i$ 
15:             break
16:     until not match
17:     return NIL

```

其中函数 EXPAND( $s, T, n$ ) 从  $T$  中扩展出  $n$  个结果, 并将  $s$  附加在每个键的前面。我们可以用广度优先搜索方法实现它(见 15.1):

```

1: function EXPAND( $s, T, n$ )
2:      $R \leftarrow$  NIL
3:      $Q \leftarrow [(s, T)]$ 
4:     while  $|R| < n$  and  $Q \neq$  NIL do
5:          $(k, T) \leftarrow$  POP( $Q$ )
6:          $v \leftarrow$  VALUE( $T$ )
7:         if  $v \neq$  NIL then
8:             INSERT( $R, (k, v)$ )
9:         for  $k_i \mapsto T_i$  in SUB-TREES( $T$ ) do
10:            PUSH( $Q, (k \# k_i, T_i)$ )

```

### 6.5.2 数字键盘输入法

2010 年前, 大多数手机上都有一个如图 6.13 所示的数字键盘, 称为 ITU-T 键盘。它将每个数字映射到 3 到 4 个英文字母上。如果要输入英文单词 home, 我们可以按照下面的顺序按键:

1. 按两次 4 键输入字符 h;
2. 按三次 6 键输入字符 o;
3. 按一次 6 键输入字符 m;
4. 按两次 3 键输入字符 e;

另外一种更快速的方法使用下面的按键顺序:

1. 依次按下 4、6、6、3, 候选单词 home 出现;



图 6.13: 手机 ITU-T 键盘

2. 按下 '\*' 号键以变换到下一个候选单词 good;
3. 按下 '\*' 号键再次变换到下一个候选单词 gone;
4. ……

后者称为预测式输入, 简称为 T9<sup>[25]、[26]</sup>。商业实现通常在内存和文件系统中使用多级缓存和索引。作为示例, 我们可以将单词存储在一个前缀树中来实现这种输入法。首先我们需要定义数字键盘映射:

$$M_{T9} = \{ \begin{array}{l} 2 \mapsto \text{"abc"}, 3 \mapsto \text{"def"}, 4 \mapsto \text{"ghi"}, \\ 5 \mapsto \text{"jkl"}, 6 \mapsto \text{"mno"}, 7 \mapsto \text{"pqrs"}, \\ 8 \mapsto \text{"tuv"}, 9 \mapsto \text{"wxyz"} \end{array} \} \quad (6.19)$$

$M_{T9}[i]$  就给出数字  $i$  对应的若干字符。我们也可以定义从字符到数字的逆映射。

$$M_{T9}^{-1} = \text{concatMap} ((d, s) \mapsto [(c, d) | c \leftarrow s]) M_{T9} \quad (6.20)$$

通过查找  $M_{T9}^{-1}$ , 我们可以将字符串转换成一组按键序列。

$$\text{digits}(s) = \{M_{T9}^{-1}[c] | c \leftarrow s\} \quad (6.21)$$

对于任何不属于  $[a..z]$  的字符, 我们将其映射到特殊字符 '#' 上。下面的例子程序定义了上述映射:

```
mapT9 = Map.fromList [(('2', "abc"), ('3', "def"), ('4', "ghi"),
                    ('5', "jkl"), ('6', "mno"), ('7', "pqrs"),
                    ('8', "tuv"), ('9', "wxyz"))]

rmapT9 = Map.fromList $ concatMap (\(d, s) -> [(c, d) | c <- s]) $
  Map.toList mapT9

digits = map (\c -> Map.findWithDefault '#' c rmapT9)
```

令  $(v, ts)$  是从所有候选单词构建出的前缀树。我们可以修改自动补齐算法来处理数字序列  $ds$ 。我们把每个子树映射  $(s \mapsto t) \in ts$  中的前缀  $s$  转换为  $\text{digits}(s)$ , 检查它是否和  $ds$  匹配(其中一个为另一个的前缀)。可能存在多个子树匹配  $ds$  的情况:

$$pfx = [(s, t) | (s \mapsto t) \in ts, digits(s) \sqsubseteq ds \text{ or } ds \sqsubseteq digits(s)]$$

$$find_{T9} t [] = [[]]$$

$$find_{T9} (v, ts) ds = concatMap find pfx \quad (6.22)$$

对  $pfx$  中的每个映射  $(s, t)$ , 函数  $find$  递归地在  $t$  中查找剩余数字  $ds'$ , 其中  $ds' = drop |s| ds$ , 然后将  $s$  附加到每个候选项前面。为了防止长度超出数字个数, 我们截取前  $n = |ds|$  个字符:

$$find (s, t) = [take n (s ++ s_i) | s_i \in find_{T9} t ds'] \quad (6.23)$$

下面的例子程序实现了预测输入法:

```
findT9 _ [] = [[]]
findT9 (PrefixTree _ ts) k = concatMap find pfx where
  find (s, t) = map (take (length k) o (s++)) $ findT9 t (drop (length s) k)
  pfx = [(s, t) | (s, t) ← ts, let ds = digits s in
         ds `isPrefixOf` k || k `isPrefixOf` ds]
```

用命令式方法实现广度优先搜索时, 可以用一个队列  $Q$ , 队列中的元素为三元组  $(prefix, D, t)$ 。每个三元组包含已搜索过的前缀  $prefix$ , 尚未搜索的数字  $D$ , 和待搜索的子树  $t$ 。队列初始的时候, 三元组包含空前缀, 全部数字, 以及前缀树的根节点。我们不断从队列中取出三元组, 检查子树的映射。对于每个映射  $(s \mapsto T')$ , 我们将  $s$  转换成  $digits(s)$ 。如果  $D$  是它的前缀, 就找到了一个候选词。我们将  $s$  附加到  $prefix$  的前面, 并记录下这一结果。如果  $digits(s)$  是  $D$  的前缀, 我们需要递归在子树  $T'$  中搜索, 我们新建一个三元组  $(prefix ++ s, D', T')$ , 其中  $D'$  是剩余的数字。然后将这一新三元组放回队列。

```
1: function LOOK-UP-T9( $T, D$ )
2:    $R \leftarrow NIL$ 
3:   if  $T = NIL$  or  $D = NIL$  then
4:     return  $R$ 
5:    $n \leftarrow |D|$ 
6:    $Q \leftarrow \{(NIL, D, T)\}$ 
7:   while  $Q \neq NIL$  do
8:      $(prefix, D, T) \leftarrow POP(Q)$ 
9:     for  $(s \mapsto T') \in SUB-TREES(T)$  do
10:       $D' \leftarrow DIGITS(s)$ 
11:      if  $D' \sqsubseteq D$  then ▷  $D'$  是  $D$  的前缀
12:        APPEND( $R, (prefix ++ s)[1..n]$ ) ▷ 限制长度为  $n$ 
13:      else if  $D \sqsubseteq D'$  then
14:        PUSH( $Q, (prefix ++ s, D - D', T')$ )
15:   return  $R$ 
```

我们从整数 trie 和整数前缀树开始,通过整数的二进制表示,我们复用二叉树实现了基于整数的映射 (map) 数据结构。接下来我们将键的类型从整数扩展到有限集元素的列表。其中一个特例就是字符串,trie 和前缀树可以用来进行文字处理。我们给出了两个应用的例子:自动补齐和预测式输入。基数树的另外一个应用是后缀树,它和 trie 与前缀树有着密切的关系,是文字和 DNA 处理的有力工具。

## 练习 6.4

- 6.4.1. 使用 trie 实现自动补齐和预测式输入。
- 6.4.2. 对于返回多个候选结果的前缀树查找算法,如何保证输出的结果按照字典顺序排序? 这会对性能产生怎样的影响?
- 6.4.3. 在没有惰性求值的环境中,如何按需返回最多  $n$  条结果?

答案: 371 页

## 6.6 附录:例子程序

复用二叉树定义整数 trie:

```
data IntTrie<T> {
  IntTrie<T> left = null, right = null
  Optional<T> value = Optional.Nothing
}
```

下面的 *insert* 例子程序用位运算实现了奇偶测试和向右移位:

```
IntTrie<T> insert(IntTrie<T> t, Int key,
  Optional<T> value = Optional.Nothing) {
  if t == null then t = IntTrie<T>()
  p = t
  while key != 0 {
    if key & 1 == 0 {
      p = if p.left == null then IntTrie<T>() else p.left
    } else {
      p = if p.right == null then IntTrie<T>() else p.right
    }
    key = key >> 1
  }
  p.value = Optional.of(value)
  return t
}
```

整数前缀树的定义:

```
data IntTree<T> {
  Int key
  T value
  Int prefix, mask = 1
  IntTree<T> left = null, right = null
}
```

```

IntTree(Int k, T v) {
    key = k, value = v, prefix = k
}

bool isLeaf = (left == null and right == null)

Self replace(IntTree<T> x, IntTree<T> y) {
    if left == x then left = y else right = y
}

bool match(Int k) = maskbit(k, mask) == prefix
}

Int maskbit(Int x, Int mask) = x & ~(mask - 1)

```

向整数前缀树插入键、值:

```

IntTree<T> insert(IntTree<T> t, Int key, T value) {
    if t == null then return IntTree(key, value)
    node = t
    Node<T> parent = null
    while (not node.isLeaf()) and node.match(key) {
        parent = node
        node = if zero(key, node.mask) then node.left else node.right
    }
    if node.isleaf() and key == node.key {
        node.value = value
    } else {
        p = branch(node, IntTree(key, value))
        if parent == null then return p
        parent.replace(node, p)
    }
    return t
}

IntTree<T> branch(IntTree<T> t1, IntTree<T> t2) {
    var t = IntTree<T>()
    (t.prefix, t.mask) = lcp(t1.prefix, t2.prefix)
    (t.left, t.right) = if zero(t1.prefix, t.mask) then (t1, t2)
                       else (t2, t1)
    return t
}

bool zero(int x, int mask) = (x & (mask >> 1) == 0)

Int lcp(Int p1, Int p2) {
    Int diff = p1 ^ p2
    Int mask = 1
    while diff != 0 {
        diff = diff >> 1
        mask = mask << 1
    }
}

```

```

return (maskbit(p1, mask), mask)
}

```

trie 的定义和插入:

```

data Trie<K, V> {
  Optional<V> value = Optional.Nothing
  Map<K, Trie<K, V>> subTrees = Map.empty()
}

Trie<K, V> insert(Trie<K, V> t, [K] key, V value) {
  if t == null then t = Trie<K, V>()
  var p = t
  for c in key {
    if p.subTrees[c] == null then p.subTrees[c] = Trie<K, V>()
    p = p.subTrees[c]
  }
  p.value = Optional.of(value)
  return t
}

```

前缀树的定义和插入:

```

data PrefixTree<K, V> {
  Optional<V> value = Optional.Nothing
  Map<[K], PrefixTree<K, V>> subTrees = Map.empty()

  Self PrefixTree(V v) {
    value = Optional.of(v)
  }
}

PrefixTree<K, V> insert(PrefixTree<K, V> t, [K] key, V value) {
  if t == null then t = PrefixTree()
  var node = t
  loop {
    bool match = false
    for var (k, tr) in node.subtrees {
      if key == k {
        tr.value = value
        return t
      }
    }
    prefix, k1, k2 = lcp(key, k)
    if prefix != [] {
      match = true
      if k2 == [] {
        node = tr
        key = k1
        break
      } else {
        node.subtrees[prefix] = branch(k1, PrefixTree(value),
                                         k2, tr)
        node.subtrees.delete(k)
      }
    }
  }
  return t
}

```

```

        }
    }
}
if !match {
    node.subtrees[key] = PrefixTree(value)
    break
}
}
return t
}

```

提取最长公共前缀 `lcp` 和分枝 `branch`:

```

([K], [K], [K]) lcp([K] s1, [K] s2) {
    j = 0
    while j < length(s1) and j < length(s2) and s1[j] == s2[j] {
        j = j + 1
    }
    return (s1[0..j-1], s1[j..], s2[j..])
}

```

```

PrefixTree<K, V> branch([K] key1, PrefixTree<K, V> tree1,
                        [K] key2, PrefixTree<K, V> tree2) {
    if key1 == []:
        tree1.subtrees[key2] = tree2
        return tree1
    t = PrefixTree()
    t.subtrees[key1] = tree1
    t.subtrees[key2] = tree2
    return t
}

```

枚举共同前缀的所有候选项:

```

[[[K], V]] startsWith(PrefixTree<K, V> t, [K] key, Int n) {
    if t == null then return []
    [T] s = []
    repeat {
        bool match = false
        for var (k, tr) in t.subtrees {
            if key.isPrefixOf(k) {
                return expand(s ++ k, tr, n)
            } else if k.isPrefixOf(key) {
                match = true
                key = key[length(k)..]
                t = tr
                s = s ++ k
                break
            }
        }
    }
    return []
}

```



```

[[[K], V]] expand([K] s, PrefixTree<K, V> t, Int n) {
  [[[K], V]] r = []
  var q = Queue([s, t])
  while length(r) < n and !q.isEmpty() {
    var (s, t) = q.pop()
    v = t.value
    if v.isPresent() then r.append((s, v.get()))
    for k, tr in t.subtrees {
      q.push((s ++ k, tr))
    }
  }
  return r
}

```

预测式输入:

```

var T9MAP={ '2':"abc", '3':"def", '4':"ghi", '5':"jkl",
            '6':"mno", '7':"pqrs", '8':"tuv", '9':"wxyz"}

var T9RMAP = { c : d for var (d, cs) in T9MAP for var c in cs }

string digits(string w) = ''.join([T9RMAP[c] for c in w])

[string] lookupT9(PrefixTree<char, V> t, string key) {
  if t == null or key == "" then return []
  res = []
  n = length(key)
  q = Queue("", key, t)
  while not q.isEmpty() {
    (prefix, key, t) = q.pop()
    for var (k, tr) in t.subtrees {
      ds = digits(k)
      if key.isPrefixOf(ds) {
        res.append((prefix ++ k)[:n])
      } else if ds.isPrefixOf(key) {
        q.append((prefix ++ k, key[length(k)..], tr))
      }
    }
  }
  return res
}

```



# 第七章 B 树

## 7.1 简介

上一章介绍的整数前缀树利用二叉树的边来表达信息。另一种扩展二叉树的方法是将分枝数目从 2 增加到  $k$ 。B 树是一种自平衡的  $k$  叉搜索树<sup>[39]</sup>。它被广泛用于计算机文件系统(基于 B+ 树, 一种 B 树的扩展形式)和数据库系统。图 7.1 展示了一棵 B 树, 我们可以观察它和二叉搜索树之间的异同。

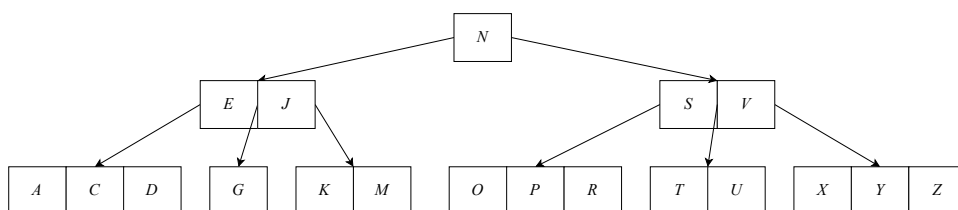


图 7.1: B 树

一棵二叉搜索树或为空, 或包含一个元素  $k$  和左右分枝  $l, r$ 。左子树  $l$  中的任何元素都小于  $k$ , 并且  $k$  小于右子树  $r$  中的任何元素<sup>1</sup>:

$$\forall x \in l, y \in r \Rightarrow x < k < y \quad (7.1)$$

B 树将这一思想推广到多个分枝。一棵 B 树或为空, 或包含  $n$  个元素和  $n+1$  个子分枝, 每个分枝也都是一棵 B 树。记这些元素为  $k_1, k_2, \dots, k_n$ , 分枝为  $t_1, t_2, \dots, t_n, t_{n+1}$ , 如图 7.2 所示。节点中的元素和分枝满足以下条件:

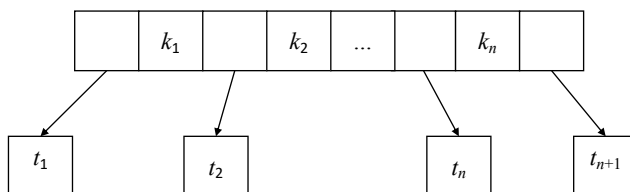


图 7.2: B 树节点

<sup>1</sup>严格来说, 节点中可以保存键(key)和对应的值(value)。值不是必需的。简单起见, 本章忽略了节点中的值, 称树中保存的内容为“元素”。

- 元素是递增的:  $k_1 \leq k_2 \leq \dots \leq k_n$ ;
- 对于任意  $k_i$ , 子树  $t_i$  中的所有元素都小于  $k_i$ , 并且  $k_i$  小于子树  $t_{i+1}$  中的任意元素。

$$\forall x_i \in t_i, i = 0, 1, \dots, n \Rightarrow x_1 < k_1 < x_2 < k_2 < \dots < x_n < k_n < x_{n+1} \quad (7.2)$$

叶子节点不包含子分枝。令元素的类型为  $K$ , 则 B 树的类型为  $BTree\ K$  或  $BTree\langle K \rangle$ 。此外, 我们还需定义一组规则以保持 B 树平衡:

- 所有的叶子节点都有相同的深度;
- 定义整数  $d$ , 称为 B 树的**最小度数**, 每个节点:
  - 最多含有  $2d - 1$  个元素;
  - 最少含有  $d - 1$  个元素, 根节点例外。

即:

$$d - 1 \leq |keys(t)| \leq 2d - 1 \quad (7.3)$$

我们接下来证明这些规则可以保证 B 树是平衡的。

**证明.** 考虑一棵含有  $n$  个元素的 B 树, 最小度数  $d \geq 2$ , 树的高度为  $h$ 。除根节点外, 其它节点至少含有  $d - 1$  个元素。根节点至少含有一个元素。如果它有子树, 则至少有两个深度为 1 的子分枝, 至少有  $2d$  个深度为 2 的子分枝, 至少有  $2d^2$  个深度为 3 的子分枝……最后, 至少有  $2d^{h-1}$  个深度为  $h$  的叶子节点。除根节点外, 将节点个数乘以  $d - 1$ , B 树中存储的元素个数满足下面的不等式:

$$\begin{aligned} n &\geq 1 + (d - 1)(2 + 2d + 2d^2 + \dots + 2d^{h-1}) \\ &= 1 + 2(d - 1) \sum_{k=0}^{h-1} d^k \\ &= 1 + 2(d - 1) \frac{d^h - 1}{d - 1} \\ &= 2d^h - 1 \end{aligned} \quad (7.4)$$

因此树的高度满足对数关系不等式:

$$h \leq \log_d \frac{n + 1}{2} \quad (7.5)$$

□

这就证明了 B 树的平衡性。最简单的 B 树称为 2-3-4 树。它的最小度数  $d = 2$ , 除根节点外的任何节点都包含 2 到 4 棵子分枝。任何红黑树本质上都可以转换为一棵 2-3-4 树。我们记度数为  $d$  的非空 B 树为  $(d, (ks, ts))$ , 其中  $ks$  是元素列表,  $ts$  是子树列表。下面的例子程序定义了 B 树:

```
data BTree a = BTree [a] [BTree a]
```

空节点记为  $(\emptyset, \emptyset)$  或 `BTree [] []`, 为了避免在每个节点中都存储一份  $d$ , 我们将其和 B 树  $t$  组成一对值  $(d, t)$ 。

## 7.2 插入

插入的思路和二叉搜索树类似, 只不过需要处理多个元素和分枝。当向 B 树  $t$  插入元素  $x$  时, 我们从根节点开始, 查找这样的位置<sup>2</sup>: 所有左侧的元素都小于  $x$ , 而右侧的元素大于  $x$ 。如果是未了的叶子节点 ( $|keys(t)| < 2d - 1$ ), 就将  $x$  插入到此位置。否则, 这一位置会指向一棵子树  $t'$ , 我们递归地将  $x$  插入到  $t'$ 。

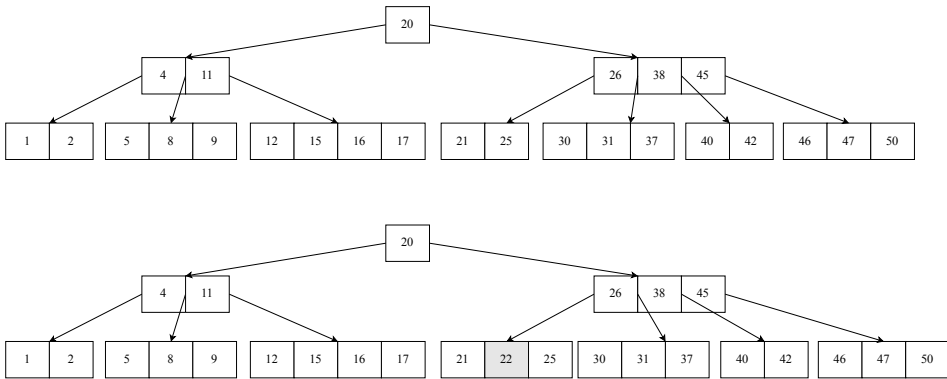


图 7.3: 将 22 插入到 2-3-4 树:  $22 > 20$ , 插入右子树;  $22 < 26$ , 插入第一棵子树。  $21 < 22 < 25$ , 插入到未了的叶子节点。

考虑向图 7.3 中的 2-3-4 树插入元素  $x = 22$ 。因为  $20 < 22$ , 我们转向右侧的子树。它包含 26、38、45。因为  $22 < 26$ , 所以接下来转向第一棵子树。它包含 21 和 25。这是一个未了的叶子节点, 将 22 插入到 21 和 25 中间。

但如果叶子节点已经含有  $2d - 1$  个元素, 插入  $x$  后就会因为元素过多破坏 B 树的规则。例如向图 7.3 插入 18 就会遇到这个问题。我们有两种解法: 先插入再分拆, 和先分拆再插入。

### 7.2.1 先插入再分拆

我们可以将红黑树中的“先插入再修复”方法扩展到 B 树。先不考虑 B 树的平衡性, 将元素插入到适当的位置。接下来, 如果树不再平衡了, 我们自下而上对含有过多元素的节点进行分拆。首先需要定义函数, 用以判断节点是否含有过多或过少的元素。

<sup>2</sup>实际上, 元素只需支持小于比较和等于比较。参见练习 7.1.1。

$$\begin{cases} \text{full } d(k_s, t_s) & = |k_s| > 2d - 1 \\ \text{low } d(k_s, t_s) & = |k_s| < d - 1 \end{cases} \quad (7.6)$$

如果含有过多元素和分枝, 我们定义 *split* 函数将其在位置  $m$  分拆为三部分, 如图 7.4 所示:

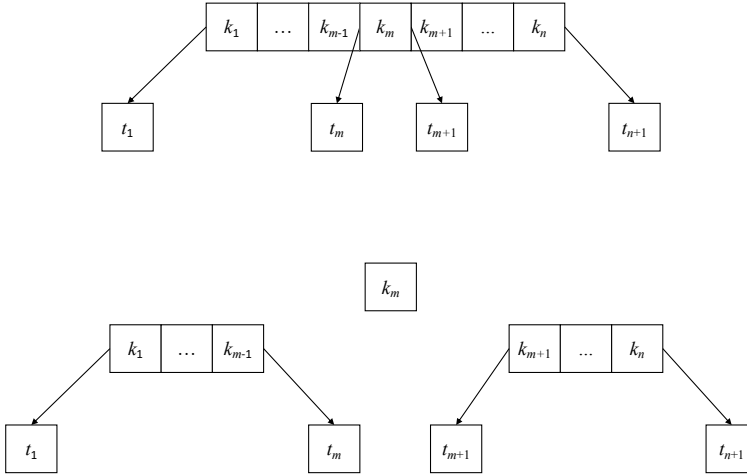


图 7.4: 在位置  $m$  将节点分拆为三部分。

$$\text{split } m(k_s, t_s) = ((k_{s_l}, t_{s_l}), k, (k_{s_r}, t_{s_r})) \quad (7.7)$$

我们使用第一章式 (1.45) 中定义的 *splitAt* 函数来实现:

$$\begin{cases} (k_{s_l}, (k : k_{s_r})) & = \text{splitAt } (m - 1) k_s \\ (t_{s_l}, t_{s_r}) & = \text{splitAt } m t_s \end{cases}$$

对称地, 我们可以定义 *unsplit* 函数, 将三个部分合并成一个 B 树节点:

$$\text{unsplit } (k_{s_l}, t_{s_l}) k (k_{s_r}, t_{s_r}) = (k_{s_l} \uparrow [k] \uparrow k_{s_r}, t_{s_l} \uparrow t_{s_r}) \quad (7.8)$$

下面的函数先将  $x$  插入树  $t$ , 然后使用 *fix* 修复平衡, 使其成为度数为  $d$  的合法 B 树:

$$\text{insert } x(d, t) = \text{fix } (d, \text{ins } t) \quad (7.9)$$

在 *ins* 之后, 如果根节点含有过多的元素, 函数 *fix* 使用 *split* 将其分拆, 并构建新的根节点。

$$\text{fix } (d, t) = \begin{cases} \text{full } d t : & (d, ([k], [l, r])), \text{ where } (l, k, r) = \text{split } d t \\ \text{otherwise} : & (d, t) \end{cases} \quad (7.10)$$

函数  $ins$  需要处理两种情况: 对于叶子节点, 我们可以重用第一章式 (1.11) 定义的列表插入函数  $insert$  来处理; 否则, 我们需要找到合适的位置, 递归地向子树插入。为此, 我们定义函数  $partition$ :

$$partition\ x\ (ks, ts) = (l, t', r) \tag{7.11}$$

其中  $l = (ks_l, ts_l), r = (ks_r, ts_r)$ 。它进一步使用第一章式 (1.47) 中定义的列表函数  $span$  进行划分:

$$\begin{cases} (ks_l, ks_r) & = span (< x) ks \\ (ts_l, (t' : ts_r)) & = splitAt |ks_l| ts \end{cases}$$

这样, 所有小于  $x$  的元素和所在的子分枝都在左侧  $l$ , 所有大于  $x$  的都在右侧  $r$ 。我们将最后一棵小于  $x$  的子树取出作为  $t'$ 。接下来我们递归地将  $x$  插入到  $t'$  中, 如图 7.5 所示。

$$\begin{aligned} ins\ (ks, \emptyset) &= (insert_L\ x\ ks, \emptyset) && \text{叶子节点, 列表插入} \\ ins\ (ks, ts) &= balance\ d\ l\ (ins\ t')\ r && \text{其中 } (l, t', r) = partition\ x\ t \end{aligned} \tag{7.12}$$

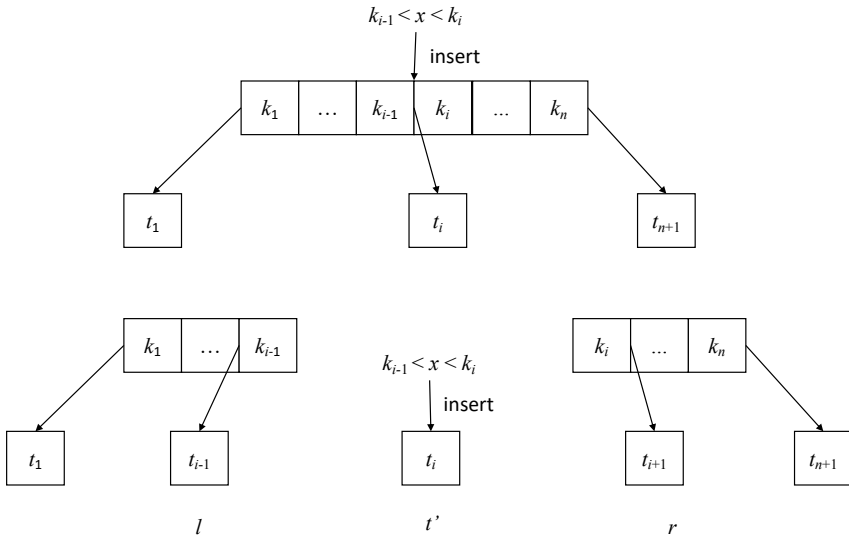


图 7.5: 用  $x$  划分节点

向  $t'$  插入  $x$  后, 它可能包含过多元素, 不再满足 B 树平衡条件。我们定义函数  $balance$  递归地进行分拆修复。

$$balance\ d\ (ks_l, ts_l)\ t\ (ks_r, ts_r) = \begin{cases} full\ d\ t : fix_f \\ otherwise : (ks_l + ks_r, ts_l + [t] + ts_r) \end{cases} \tag{7.13}$$

其中  $fix_f$  将度数为  $d$  的子分枝  $t$  分拆为  $(t_1, k, t_2) = split\ d\ t$ , 然后构建一个新的 B 树节点:

$$fix_f = (ks_l \# [k] \# ks_r, ts_l \# [t_1, t_2] \# ts_r) \quad (7.14)$$

下面的例子程序实现了 B 树的插入算法：

```

partition x (BTree ks ts) = (l, t, r) where
  l = (ks1, ts1)
  r = (ks2, ts2)
  (ks1, ks2) = span (< x) ks
  (ts1, (t:ts2)) = splitAt (length ks1) ts

split d (BTree ks ts) = (BTree ks1 ts1, k, BTree ks2 ts2) where
  (ks1, k:ks2) = splitAt (d - 1) ks
  (ts1, ts2) = splitAt d ts

insert x (d, t) = fixRoot (d, ins t) where
  ins (BTree ks []) = BTree (List.insert x ks) []
  ins t = balance d l (ins t') r where (l, t', r) = partition x t

fixRoot (d, t) | full d t = let (t1, k, t2) = split d t in
  (d, BTree [k] [t1, t2])
  | otherwise = (d, t)

balance d (ks1, ts1) t (ks2, ts2)
  | full d t = fixFull
  | otherwise = BTree (ks1 # ks2) (ts1 # [t] # ts2)
where
  fixFull = let (t1, k, t2) = split d t in
    BTree (ks1 # [k] # ks2) (ts1 # [t1, t2] # ts2)

```

图 7.6 给出了两棵 B 树的例子，它们都是依次将“GMPXACDEJKNORSTUVYZ”中的元素插入 B 树构造出的。

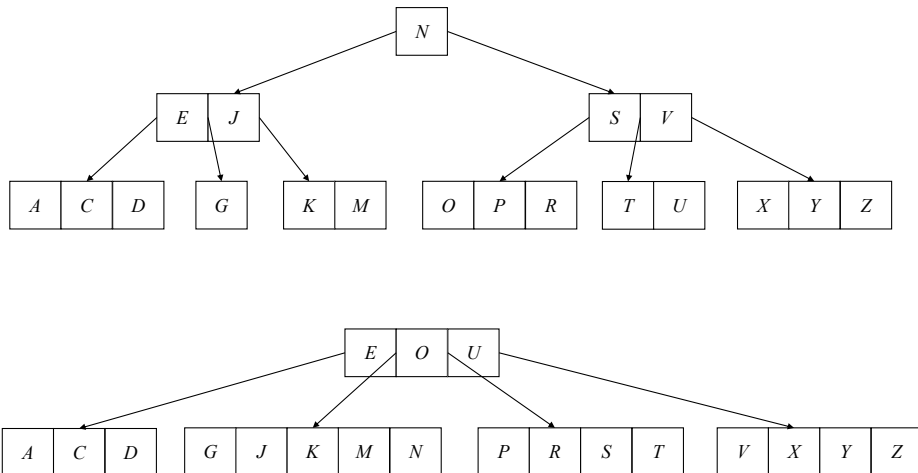


图 7.6: 依次插入“GMPXACDEJKNORSTUVYZ”。

上： $d = 2$ (2-3-4 树)，下： $d = 3$



### 7.2.2 先分拆再插入

第二种方法是在插入前先分拆节点以避免其含有过多元素。命令式实现常用这一方法。自顶向下递归插入时,当遇到含有  $2d - 1$  个元素的节点时,我们将其分拆为三部分,如图 7.4 所示。每个新节点都只含有  $d - 1$  个元素,即使插入元素后也仍然是合法的 B 树节点。对于节点  $x$ ,令  $K(x)$  表示它包含的元素, $T(x)$  表示它包含的子分枝。记  $x$  中的第  $i$  个元素为  $k_i(x)$ ,第  $j$  棵子分枝为  $t_j(x)$ 。下面的算法在第  $i$  个位置对节点  $z$  进行分拆:

```

1: procedure SPLIT( $z, i$ )
2:    $d \leftarrow \text{DEG}(z)$ 
3:    $x \leftarrow t_i(z)$ 
4:    $y \leftarrow \text{CREATE-NODE}$ 
5:    $K(y) \leftarrow [k_{d+1}(x), k_{d+2}(x), \dots, k_{2d-1}(x)]$ 
6:    $K(x) \leftarrow [k_1(x), k_2(x), \dots, k_d(x)]$ 
7:   if  $x$  is not leaf then
8:      $T(y) \leftarrow [t_{d+1}(x), t_{d+2}(x), \dots, t_{2d}(x)]$ 
9:      $T(x) \leftarrow [t_1(x), t_2(x), \dots, t_d(x)]$ 
10:  INSERT-AT( $K(z), i, k_d(x)$ )
11:  INSERT-AT( $T(z), i + 1, y$ )

```

分拆节点  $x = t_i(z)$  时,我们将第  $d$  个元素  $k_d(x)$  向上推入父节点  $z$ 。如果  $z$  已经满了,推入元素后就会违反 B 树规则。为此,我们需要从根节点起,自顶向下沿着插入的路径进行检查,分拆所有含有  $2d - 1$  个元素的节点。因为所有的父节点都这样被处理过,所以可以接受推上来的元素。这一方法只需要一轮自顶向下的处理,无需回溯。如果根节点已满,则需要新建一个节点,并将原来的根节点作为它的唯一子树。下面是插入算法的实现:

```

1: function INSERT( $t, k$ )
2:    $r \leftarrow t$ 
3:   if  $r$  is full then ▷ 根节点已满
4:      $s \leftarrow \text{CREATE-NODE}$ 
5:      $T(s) \leftarrow [r]$ 
6:     SPLIT( $s, 1$ )
7:      $r \leftarrow s$ 
8:   return INSERT-NONFULL( $r, k$ )

```

其中算法 INSERT-NONFULL 假设传入的节点  $r$  不满。如果  $r$  是叶子节点,我们按照  $k$  的大小将其插入到相应位置(练习 7.1.3 要求使用二分查找进行插入)。否则,我们找到一个位置,使得  $k_i(r) < k < k_{i+1}(r)$ ,如果分枝  $t_i(r)$  满了,就进行分拆。然后继续向子分枝插入。

```

1: function INSERT-NONFULL( $r, k$ )
2:    $n \leftarrow |K(r)|$ 
3:   if  $r$  is leaf then
4:      $i \leftarrow 1$ 
5:     while  $i \leq n$  and  $k > k_i(r)$  do
6:        $i \leftarrow i + 1$ 
7:     INSERT-AT( $K(r), i, k$ )
8:   else
9:      $i \leftarrow n$ 
10:    while  $i > 1$  and  $k < k_i(r)$  do
11:       $i \leftarrow i - 1$ 
12:    if  $t_i(r)$  is full then
13:      SPLIT( $r, i$ )
14:    if  $k > k_i(r)$  then
15:       $i \leftarrow i + 1$ 
16:    INSERT-NONFULL( $t_i(r), k$ )
17:  return  $r$ 

```

这一算法是递归的。练习 7.1.2 要求使用循环消除递归。图 7.7 给出了依次插入“GMPXACDEJKNORSTUVYZ”时的结果。

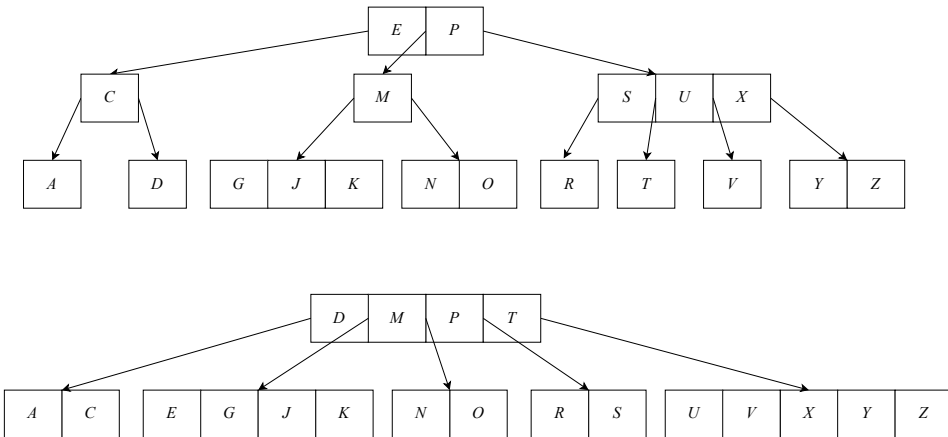


图 7.7: 依次插入“GMPXACDEJKNORSTUVYZ”。上:  $d = 2$  (2-3-4 树), 下:  $d = 3$

### 7.2.3 列表对

用列表存储元素时, 我们需要从第一个元素开始, 扫描列表找到插入位置。如果用数组存储, 我们可以使用二分查找。可否从节点中的某个位置开始, 根据元素的大小向左或向右前进呢? 我们可以将 B 树节点表示为三部分: 某棵子分枝  $t'$ , 它的左侧

$l$ , 右侧  $r$ 。其中左右侧都是“元素/子分枝”对  $(k_i, t_i)$  的列表。特别的, 左侧  $l$  是逆序的。 $l$  和  $r$  经由  $t'$  头对头地连接起来, 组成一个如图 7.8 所示的马蹄形。我们可以用常数时间前后移动。

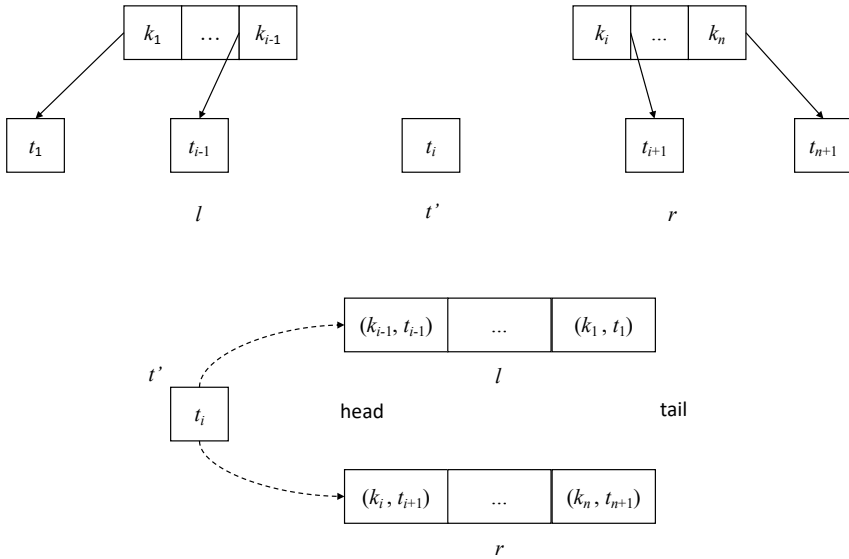


图 7.8: 将 B 树表示为某个子分枝和它两侧的一对列表

下面的例子程序用列表对定义了 B 树节点。它或者为空, 或者包含三部分: 左侧逆序的(元素, 子分枝)列表, 中间的某个子分枝, 右侧的(元素、子分枝)列表。我们记非空的节点为  $(l, t', r)$ 。

```

data BTree a = Empty
           | BTree [(a, BTree a)] (BTree a) [(a, BTree a)]
    
```

向右移动一步时, 我们从  $r$  中取出第一对值  $(k, t)$ , 组成另一对  $(k, t')$  置于  $l$  的最前面。然后用  $t$  替换  $t'$ 。向左移动的步骤与此对称。它们都只需要常数时间。

$$\begin{aligned}
 \text{step}_l ((k, t):l, t', r) &= (l, t, (k, t'):r) \\
 \text{step}_r (l, t', (k, t):r) &= ((k, t'):l, t, r)
 \end{aligned}
 \tag{7.15}$$

利用左右移动, 我们可以实现划分函数  $\text{partition } p \ t$ , 根据条件  $p$  把 B 树  $t$  分成左中右三部分:  $(l, m, r)$ 。所有  $l$  中的分枝和  $m$  都满足  $p$ , 而  $r$  中的分枝不满足。定义函数  $\text{hd} = \text{fst} \circ \text{head}$ , 它从列表中取出第一对值  $(a, b)$ , 然后再获取  $a$ 。

$$\begin{aligned}
\text{partition } p (\emptyset, m, r) &= \begin{cases} p(\text{hd}(r)) : & \text{partition } p (\text{step}_r t) \\ \text{otherwise} : & (\emptyset, m, r) \end{cases} \\
\text{partition } p (l, m, \emptyset) &= \begin{cases} (\text{not} \circ p)(\text{hd}(l)) : & \text{partition } p (\text{step}_l t) \\ \text{otherwise} : & (l, m, \emptyset) \end{cases} \\
\text{partition } p (l, m, r) &= \begin{cases} p(\text{hd}(l)) \text{ and } (\text{not} \circ p)(\text{hd}(r)) : & (l, m, r) \\ p(\text{hd}(r)) : & \text{partition } p (\text{step}_r t) \\ (\text{not} \circ p)(\text{hd}(l)) : & \text{partition } p (\text{step}_l t) \end{cases}
\end{aligned} \tag{7.16}$$

例如  $\text{partition} (< k) t$  将  $t$  中所有不小于  $k$  的元素和分枝留在右侧。下面的例子程序实现了  $\text{partition}$  函数：

```

partition p t@(BTree [] m r)
  | p (hd r) = partition p (stepR t)
  | otherwise = ([], m, r)
partition p t@(BTree l m [])
  | (not ∘ p) (hd l) = partition p (stepL t)
  | otherwise = (l, m, [])
partition p t@(BTree l m r)
  | p (hd l) && (not ∘ p) (hd r) = (l, m, r)
  | p (hd r) = partition p (stepR t)
  | (not ∘ p) (hd l) = partition p (stepL t)

```

我们也可以利用  $\text{step}_l/\text{step}_r$  把含有过多元素的节点在位置  $d$  拆分。令  $n = |l|$  表示左侧的“元素/子分枝”数量。 $f^n(x)$  表示对变量  $x$  重复应用函数  $f$  共  $n$  次。

$$\text{split } d t = \begin{cases} n < d : & sp(\text{step}_r^{d-n}(t)) \\ n > d : & sp(\text{step}_r^{n-d}(t)) \\ \text{otherwise} : & sp(t) \end{cases} \tag{7.17}$$

其中  $sp$  进行如下的分拆：

$$sp (l, t, (k, t') : r) = ((l, t, \emptyset), k, (\emptyset, t', r)) \tag{7.18}$$

利用  $\text{partition}$  和  $\text{split}$ ，对于列表对表示的 B 树，我们可以定义出插入算法。首先我们需要修改 B 树含有过多、过少元素的判断：

$$\begin{aligned}
\text{full } d \emptyset &= \text{False} \\
\text{full } d (l, t', r) &= |l| + |r| > 2d - 1
\end{aligned} \tag{7.19}$$

和

$$\begin{aligned}
\text{low } d \emptyset &= \text{False} \\
\text{low } d (l, t', r) &= |l| + |r| < d - 1
\end{aligned} \tag{7.20}$$

向度数为  $d$  的 B 树  $t$  插入元素  $x$  时，我们首先递归地插入，然后再修复元素过多的问题：

$$\text{insert } x (d, t) = \text{fix } (d, \text{ins } t) \quad (7.21)$$

如果根节点含有过多元素, 函数  $\text{fix}$  在位置  $d$  将其分拆:

$$\text{fix } (d, t) = \begin{cases} \text{full } d \ t: & (d, (\emptyset, t_1, [(k, t_2)]) \text{ 其中 } (t_1, k, t_2) = \text{split } d \ t \\ \text{otherwise}: & (d, t) \end{cases} \quad (7.22)$$

函数  $\text{ins}$  需要处理  $t = \emptyset$  和  $t \neq \emptyset$  两种情况。对于空树, 我们新建一个单独的叶子节点; 否则调用  $(l, t', r) = \text{partition } (< x) \ t$  定位到递归插入的位置:

$$\begin{aligned} \text{ins } \emptyset &= (\emptyset, \emptyset, [(x, \emptyset)]) \\ \text{ins } t &= \begin{cases} t' = \emptyset: & \text{balance } d \ l \ \emptyset \ ((x, \emptyset):r) \\ t' \neq \emptyset: & \text{balance } d \ l \ (\text{ins } t') \ r \end{cases} \end{aligned} \quad (7.23)$$

函数  $\text{balance}$  检查子分枝  $t$  是否包含过多元素并进行分拆。

$$\text{balance } d \ l \ t \ r = \begin{cases} \text{full } d \ t: & \text{fixFull} \\ \text{otherwise}: & (l, t, r) \end{cases} \quad (7.24)$$

其中  $\text{fixFull} = (l, t_1, ((k, t_2):r), (t_1, k, t_2) = \text{split } d \ t$ 。下面的例子程序实现了插入算法:

```

insert x (d, t) = fixRoot (d, ins t) where
  ins Empty = BTree [] Empty [(x, Empty)]
  ins t = let (l, t', r) = partition (< x) t in
    case t' of
      Empty → balance d l Empty ((x, Empty):r)
      _     → balance d l (ins t') r

fixRoot (d, t) | full d t = let (t1, k, t2) = split d t in
  (d, BTree [] t1 [(k, t2)])
  | otherwise = (d, t)

balance d l t r | full d t = fixFull
  | otherwise = BTree l t r

where
  fixFull = let (t1, k, t2) = split d t in BTree l t1 ((k, t2):r)

split d t@(BTree l _ _) | n < d = sp $ iterate stepR t !! (d - n)
  | n > d = sp $ iterate stepL t !! (n - d)
  | otherwise = sp t

where
  n = length l
  sp (BTree l t ((k, t'):r)) = (BTree l t [], k, BTree [] t' r)

```

## 练习 7.1

7.1.1. 我们是否可以用  $\leq$  使得 B 树含有重复元素?

7.1.2. 使用循环消除“先分拆再插入”算法中的递归。

7.1.3. 我们使用线性查找获得元素插入的位置。请使用二分查找对命令式实现进行改进。算法复杂度会提升么？

答案: 372 页

## 7.3 查找

我们可以将二叉搜索树的查找算法扩展到含有多个分枝的 B 树。二叉树查找只有左右两个方向,但 B 树有多个方向。考虑在 B 树  $t = (ks, ts)$  中查找元素  $k$ , 如果  $t$  是叶子节点 ( $ts$  为空), 则问题简化为列表查找; 否则, 我们用  $k$  将树  $t$  划分为三部分:  $l = (ks_l, ts_l)$ 、 $t'$ 、 $r = (ks_r, ts_r)$ , 其中  $l$  和子分枝  $t'$  中的所有元素都小于  $k$ , 而  $r$  中的所有元素大于等于  $k$ 。如果  $r$  中的第一个元素  $ks_r$  等于  $k$ , 我们就找到了结果; 否则我们递归地在子分枝  $t'$  中查找。

$$\begin{aligned} \text{lookup } k \ (ks, \emptyset) &= \begin{cases} k \in ks : & \text{Just } (ks, \emptyset) \\ \text{otherwise} : & \text{Nothing} \end{cases} \\ \text{lookup } k \ (ks, ts) &= \begin{cases} \text{Just } k = \text{safeHd } ks_r : & \text{Just } (ks, ts) \\ \text{otherwise} : & \text{lookup } k \ t' \end{cases} \end{aligned} \quad (7.25)$$

其中  $((ks_l, ts_l), t', (ks_r, ts_r)) = \text{partition } k \ t$ 。函数  $\text{safeHd}$  定义为:

$$\begin{aligned} \text{safeHd } [] &= \text{Nothing} \\ \text{safeHd } (x:xs) &= \text{Just } x \end{aligned}$$

下面的例子程序<sup>3</sup>实现了查找算法。

```
lookup k t@(BTree ks []) = if k `elem` ks then Just t else Nothing
lookup k t = if (Just k) == safeHd ks then Just t
           else lookup k t' where
           (_, t', (ks, _)) = partition k t
```

对于列表的实现,思路是类似的。如果树不为空,我们用条件“ $< k$ ”进行划分。然后检查右侧部分的第一个元素是否等于  $k$ , 否则再递归地进行查找:

$$\begin{aligned} \text{lookup } k \ \emptyset &= \text{Nothing} \\ \text{lookup } k \ t &= \begin{cases} \text{Just } k = \text{safeFst } (\text{safeHd } r) : & \text{Just } (l, t', r) \\ \text{otherwise} : & \text{lookup } k \ t' \end{cases} \end{aligned} \quad (7.26)$$

其中  $(l, t', r) = \text{partition } (< k) \ t$  是对非空树的划分。 $\text{safeFst}$  将函数  $\text{fst}$  应用到“Maybe”的值上, 下面的例子程序使用了  $\text{fmap}$  来实现:

<sup>3</sup>`safeHd` 在某些程序库中以 `listToMaybe` 提供

```

lookup x Empty = Nothing
lookup x t = let (l, t', r) = partition (< x) t in
  if (Just x) == fmap fst (safeHd r) then Just (BTree l t' r)
  else lookup x t'

```

对于命令式实现,我们从根节点  $r$  开始,找到位置  $i$  使得  $k_i(r) \leq k < k_{i+1}(r)$ 。如果  $k_i(r) = k$ ,则返回节点  $r$  和索引  $i$  组成的值对;否则,我们继续在子分枝  $t_i(r)$  中继续查找。如果  $r$  是叶子节点,并且  $k$  不在其中,则返回空结果。

```

1: function LOOKUP( $r, k$ )
2:   loop
3:      $i \leftarrow 1, n \leftarrow |K(r)|$ 
4:     while  $i \leq n$  and  $k > k_i(r)$  do
5:        $i \leftarrow i + 1$ 
6:       if  $i \leq n$  and  $k = k_i(r)$  then
7:         return ( $r, i$ )
8:       if  $r$  is leaf then
9:         return Nothing ▷  $k$  不存在
10:      else
11:         $r \leftarrow t_i(r)$  ▷ 继续查找第  $i$  棵分枝

```

## 练习 7.2

7.2.1. 使用二分查找改进命令式查找算法。

答案: 373 页

## 7.4 删除

删除元素后,节点可能因为元素不足无法满足 B 树的要求。除根节点外,元素数不能小于  $d-1$ ,其中  $d$  是最小度数。对称于插入算法,我们也有两种解法:先删除再修复、先合并再删除。

### 7.4.1 先删除再修复

我们首先扩展二叉搜索树的删除算法到多个分枝,然后再修复 B 树的平衡性。算法包含两步:

$$\text{delete } x(d, t) = \text{fix}(d, \text{del } x \ t) \quad (7.27)$$

其中函数  $\text{del}$  是对多分枝扩展的删除操作。如果  $t$  是叶子节点,我们从节点元素中删除  $x$ ;否则,我们用  $x$  将树划分为三部分:  $(l, t', r)$ 。其中  $l$  和  $t'$  的所有元素小于  $x$ ,

而  $r$  中的其余元素大于等于 ( $\geq$ )  $x$ 。如果  $r$  不为空, 我们取出其中的第一个元素  $k_i$ , 若它等于  $x$  (即  $k_i = x$ ), 我们接下来用子分枝  $t'$  中的最大元素  $k'$  (即  $k' = \max(t')$ ) 取代  $k_i$ 。然后递归地从  $t'$  中删除  $k'$ , 如图 7.9 所示。否则 ( $r$  为空或  $k_i \neq x$ ), 我们递归地从  $t'$  中删除  $x$ 。

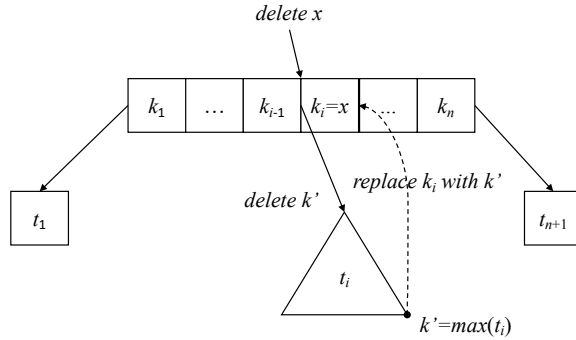


图 7.9: 用  $k' = \max(t')$  替换  $k_i$ , 然后递归地从  $t'$  删除  $k'$

$$\begin{aligned} \text{del } x (ks, \emptyset) &= (\text{delete}_l x ks, \emptyset) \\ \text{del } x t &= \begin{cases} \text{Just } x = \text{safeHd } ks' : \text{balance } d l (\text{del } k' t') (k' : (\text{tail } ks'), ts') \\ \text{otherwise} : \text{balance } d l (\text{del } x t') (ks', ts') \end{cases} \end{aligned} \quad (7.28)$$

其中  $(l, t', (ks', ts')) = \text{partition } x t$ , 是用  $x$  进行划分的三部分。我们可以进一步从  $t'$  中获得最大元素  $k'$ 。函数  $\max$  定义如下:

$$\begin{aligned} \max (ks, \emptyset) &= \text{last } ks \\ \max (ks, ts) &= \max (\text{last } ts) \end{aligned} \quad (7.29)$$

函数  $\text{last}$  返回列表中的最后一个元素 (式 (1.4))。  $\text{delete}_l$  是式 (1.14) 中定义的列表删除函数。  $\text{tail}$  将列表中的第一个元素去掉, 并返回剩下的元素 (式 (1.1))。 我们还需要修改此前在插入算法中定义的  $\text{balance}$  函数。 如果节点中的元素太少, 就进行合并。

$$\text{balance } d (ks_l, ts_l) t (ks_r, ts_r) = \begin{cases} \text{full } d t : \text{fix}_f \\ \text{low } d t : \text{fix}_l \\ \text{otherwise} : (ks_l \# ks_r, ts_l \# [t] \# ts_r) \end{cases} \quad (7.30)$$

如果  $t$  中的元素不足 ( $< d - 1$ ), 我们调用  $\text{fix}_l$  与左侧  $(ks_l, ts_l)$  或右侧  $(ks_r, ts_r)$  合并 (选择一个不为空的)。 以左侧为例: 我们从  $ks_l, ts_l$  中取出最后的元素  $k_m, t_m$ 。 然后调用  $\text{unsplit}$  (式 (7.8)) 和  $t$  合并:  $\text{unsplit } t_m k_m t$ 。 构造一个含有更多元素的新分枝。 最后, 我们再次调用  $\text{balance}$  函数构造最终的 B 树。



$$fix_l = \begin{cases} ks_l \neq \emptyset : & balance\ d\ (init\ ks_l, init\ ts_l)\ (unsplit\ t_m\ k_m\ t)\ (ks_r, ts_r) \\ ks_r \neq \emptyset : & balance\ d\ (ks_l, ts_l)\ (unsplit\ t\ k_1\ t_1)\ (tail\ ks_r, tail\ ts_r) \\ otherwise : & t \end{cases} \quad (7.31)$$

上式最后一种情况中  $ks_l = ks_r = \emptyset$ , 两侧都为空。这是一棵只有一个叶子的树, 无需进一步修复。 $k_1$  和  $t_1$  分别是  $ks_r$  和  $ts_r$  中的第一个元素。最后我们修改此前插入算法中定义的  $fix$  函数, 加入删除的处理逻辑:

$$fix\ (d, (\emptyset, [t])) = (d, t)$$

$$fix\ (d, t) = \begin{cases} full\ d\ t : & (d, ([k], [l, r])), \text{ 其中 } (l, k, r) = split\ d\ t \\ otherwise : & (d, t) \end{cases} \quad (7.32)$$

上式中, 我们在最前面加入一条: 如果删除后, 根节点只包含一棵子树, 我们可以缩减高度, 将此唯一的子树作为新的根。下面的例子程序实现了删除算法。

```

delete x (d, t) = fixRoot (d, del x t) where
  del x (BTree ks []) = BTree (List.delete x ks) []
  del x t = if (Just x) == safeHd ks' then
    let k' = max t' in
      balance d l (del k' t') (k':(tail ks'), ts')
    else balance d l (del x t') r
  where
    (l, t', r@(ks', ts')) = partition x t

fixRoot (d, BTree [] [t]) = (d, t)
fixRoot (d, t) | full d t = let (t1, k, t2) = split d t in
  (d, BTree [k] [t1, t2])
  | otherwise = (d, t)

balance d (ks1, ts1) t (ks2, ts2)
  | full d t = fixFull
  | low d t = fixLow
  | otherwise = BTree (ks1 # ks2) (ts1 # [t] # ts2)
where
  fixFull = let (t1, k, t2) = split d t in
    BTree (ks1 # [k] # ks2) (ts1 # [t1, t2] # ts2)
  fixLow | not $ null ks1 = balance d (init ks1, init ts1)
    (unsplit (last ts1) (last ks1) t)
    (ks2, ts2)
  | not $ null ks2 = balance d (ks1, ts1)
    (unsplit t (head ks2) (head ts2))
    (tail ks2, tail ts2)
  | otherwise = t

```

我们将列表对 B 树的删除算法留作练习。图 7.10、图 7.11、图 7.12 描述了删除的例子。

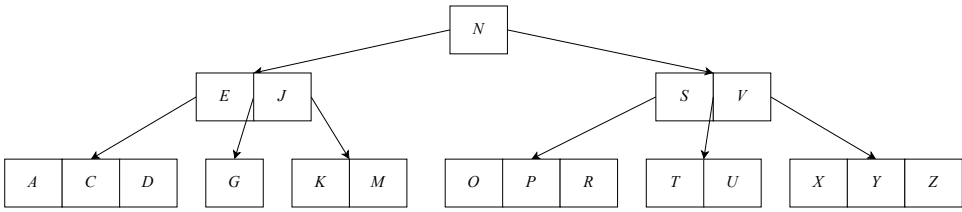


图 7.10: 删除前

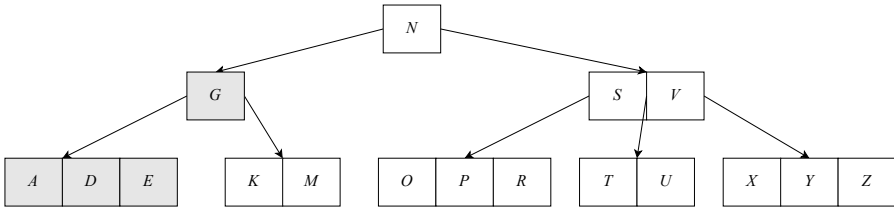
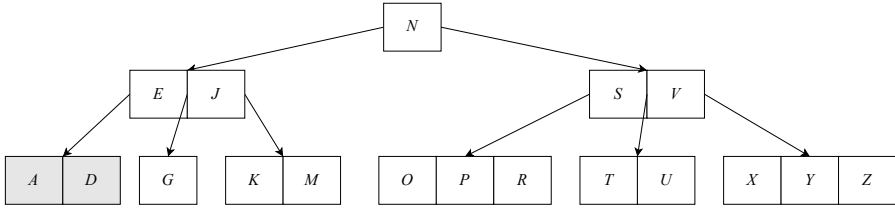


图 7.11: 删除 'C', 然后删除 'J'

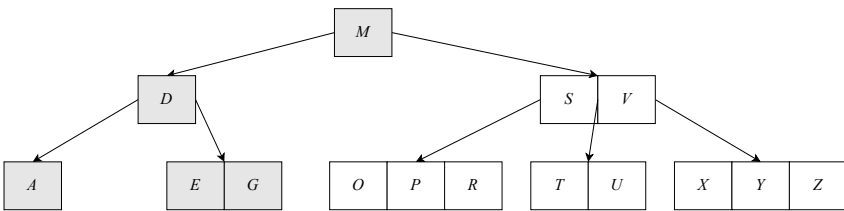
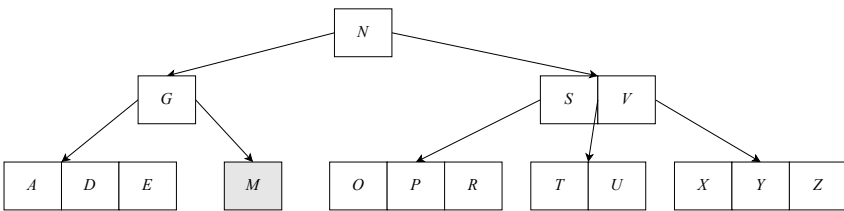


图 7.12: 删除 'K', 然后删除 'N'

### 7.4.2 先合并再删除

另一种方法是先把元素不足的节点合并,然后再删除。考虑从树  $t$  中删除元素  $x$ 。我们先从最简单的情况入手。

**情况 1:** 如果  $x$  存在于  $t$  的元素中,并且  $t$  是叶子节点。我们可以直接将  $t$  中的  $x$  删除。如果  $t$  是树中的唯一节点(根),则无需进一步修复。

**情况 2:** 如果  $x$  存在于  $t$  的元素中,但  $t$  不是叶子节点。则存在三种子情况:

**情况 2a:** 如图 7.9 所示,令  $k_i = x$  的前驱元素为  $k'$ ,其中  $k' = \max(t_i)$ 。如果  $t_i$  含有足够的元素( $\geq d$ ),我们用  $k'$  替换  $k_i$ ,然后递归地从  $t_i$  中删除  $k'$ 。

**情况 2b:** 如果  $t_i$  中的元素不足,但是子分枝  $t_{i+1}$  含有足够的元素( $\geq d$ ),对称地,我们用后继元素  $k'' = \min(t_{i+1})$  替换  $k_i$ ,然后递归地从  $t_{i+1}$  中删除  $k''$ 。如图 7.13 所示。

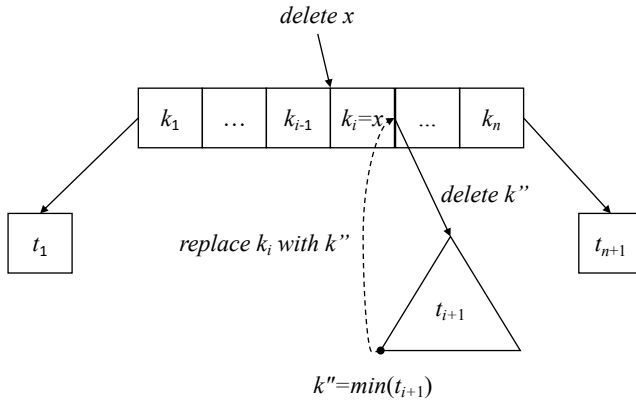


图 7.13: 用  $k'' = \min(t_{i+1})$  替换  $k_i$ ,然后递归地从  $t_{i+1}$  中删除  $k''$

**情况 2c:** 如果  $t_i$  和  $t_{i+1}$  的元素都不足( $|t_i| = |t_{i+1}| = d - 1$ ),我们将  $t_i, x, t_{i+1}$  合并成一个新节点。它含有  $2d - 1$  个元素,可以安全地从中删除。如图 7.14 所示。

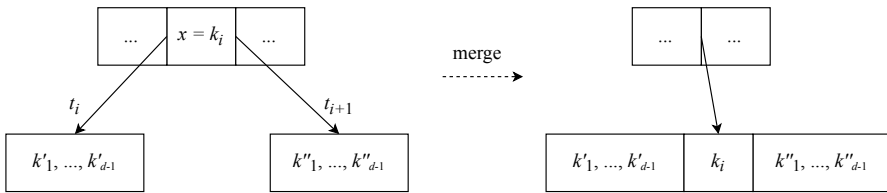


图 7.14: 先合并再删除

合并过程会将元素  $k_i$  推入子树。如果  $t$  因此变空(不再含有元素),说明  $k_i$  是  $t$  中的唯一元素,并且  $t_i, t_{i+1}$  是仅有的两棵子树。我们可以将树的高度缩减,如图 7.15 所示。

**情况 3:** 如果  $t$  的元素中不包含  $x$ ,我们需要递归地在某个子分枝  $t_i$  中删除  $x$ 。如果  $t_i$  中的元素不足,我们需要处理两种子情况:

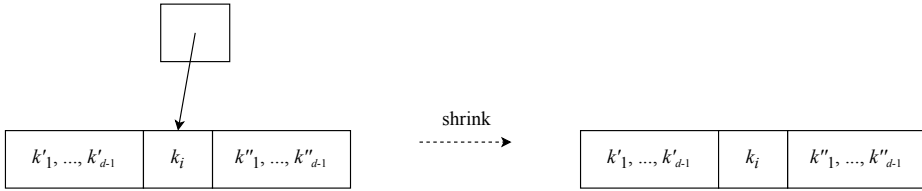


图 7.15: 缩减高度

**情况 3a:** 如果  $t_i$  的两个相邻节点  $t_{i-1}$ 、 $t_{i+1}$  中的任何一个含有足够的元素 ( $\geq d$ )，我们将  $t$  中的一个元素移到  $t_i$ ，然后将相邻节点中的一个元素向上移入  $t$ ，将相应的子分枝移入  $t_i$ 。如图 7.16 所示， $t_i$  获得一个元素。接下来递归地从  $t_i$  中删除  $x$ 。

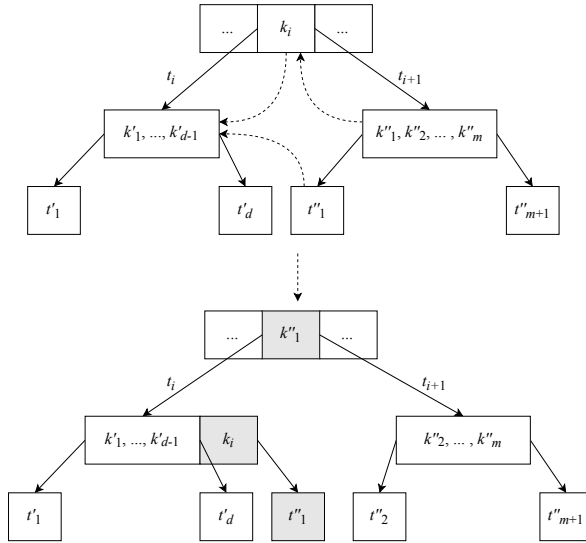


图 7.16: 从右侧移入一个元素

**情况 3b:** 如果两个相邻节点中的元素都不足 ( $|t_{i-1}| = |t_{i+1}| = d - 1$ )，我们将  $t_i$ 、 $t$  中的一个元素，和任一相邻节点合并成一个新节点，如图 7.17 所示。然后从中递归地删除  $x$ 。

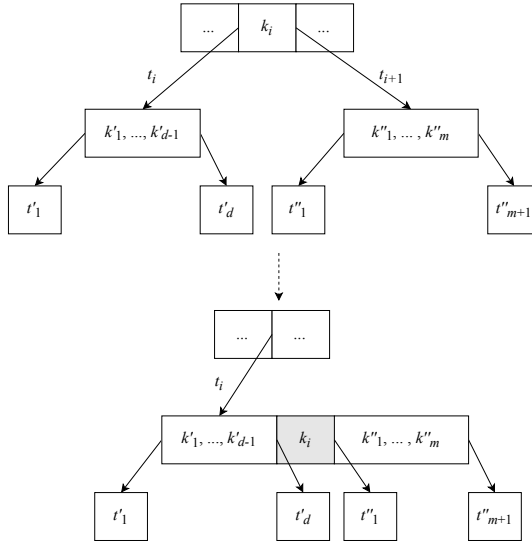
下面的 DELETE 函数实现了先合并再删除算法：

```

1: function DELETE( $t, k$ )
2:   if  $t$  is empty then
3:     return  $t$ 
4:    $i \leftarrow 1, n \leftarrow |K(t)|$ 
5:   while  $i \leq n$  and  $k > k_i(t)$  do
6:      $i \leftarrow i + 1$ 
7:   if  $k = k_i(t)$  then
8:     if  $t$  is leaf then

```

▷ 情况 1

图 7.17: 合并  $t_i, k, t_{i+1}$ 

```

9:         REMOVE( $K(t), k$ )
10:    else ▷ 情况 2
11:        if  $|K(t_i(t))| \geq d$  then ▷ 情况 2a
12:             $k_i(t) \leftarrow \text{MAX}(t_i(t))$ 
13:            DELETE( $t_i(t), k_i(t)$ )
14:        else if  $|K(t_{i+1}(t))| \geq d$  then ▷ 情况 2b
15:             $k_i(t) \leftarrow \text{MIN}(t_{i+1}(t))$ 
16:            DELETE( $t_{i+1}(t), k_i(t)$ )
17:        else ▷ 情况 2c
18:            MERGE-AT( $t, i$ )
19:            DELETE( $t_i(t), k$ )
20:            if  $K(T)$  is empty then
21:                 $t \leftarrow t_i(t)$  ▷ 缩减高度
22:    return  $t$ 
23: if  $t$  is not leaf then
24:     if  $k > k_n(t)$  then
25:          $i \leftarrow i + 1$ 
26:     if  $|K(t_i(t))| < d$  then ▷ 情况 3
27:         if  $i > 1$  and  $|K(t_{i-1}(t))| \geq d$  then ▷ 情况 3a:左
28:             INSERT( $K(t_i(t)), k_{i-1}(t)$ )
29:              $k_{i-1}(t) \leftarrow \text{POP-LAST}(K(t_{i-1}(t)))$ 
30:             if  $t_i(t)$  is not leaf then
31:                 INSERT( $T(t_i(t)), \text{POP-BACK}(T(t_{i-1}(t)))$ )

```

```

32:         else if  $i \leq n$  and  $|K(t_{i+1}(t))| \geq d$  then                                ▷ 情况 3a: 右
33:             APPEND( $K(t_i(t)), k_i(t)$ )
34:              $k_i(t) \leftarrow$  POP-FIRST( $K(t_{i+1}(t))$ )
35:             if  $t_i(t)$  is not leaf then
36:                 APPEND( $T(t_i(t)),$  POP-FIRST( $T(t_{i+1}(t))$ )))
37:         else                                                                    ▷ 情况 3b
38:             if  $i = n + 1$  then
39:                  $i \leftarrow i - 1$ 
40:                 MERGE-AT( $t, i$ )
41:             DELETE( $t_i(t), k$ )
42:             if  $K(t)$  is empty then                                            ▷ 缩减高度
43:                  $t \leftarrow t_1(t)$ 
44:         return  $t$ 

```

其中 MERGE-AT( $t, i$ ) 将分枝  $t_i(t)$ 、元素  $k_i(t)$ 、分枝  $t_{i+1}(t)$  合并成一个新分枝。

```

1: procedure MERGE-AT( $t, i$ )
2:      $x \leftarrow t_i(t)$ 
3:      $y \leftarrow t_{i+1}(t)$ 
4:      $K(x) \leftarrow K(x) \uplus [k_i(t)] \uplus K(y)$ 
5:      $T(x) \leftarrow T(x) \uplus T(y)$ 
6:     REMOVE-AT( $K(t), i$ )
7:     REMOVE-AT( $T(t), i + 1$ )

```

### 练习 7.3

7.3.1. 我们在本节中使用了前驱子分枝中的最大元素  $k' = \max(t')$  替换要删除的元素  $k$ ，然后递归地在  $t'$  中删除  $k'$ 。还有一种对称的处理方法：用后继分枝中的最小元素来替换  $k$ 。请实现这一方法

7.3.2. 实现列表对 B 树的删除算法。

答案: 374 页

## 7.5 小结

B 树将二叉搜索树扩展到多个分枝，并将分枝的数目限制在一个范围内。B 树被用来控制磁盘访问 ([4], 第 18 章)。B 树节点的分枝不会过多、过少，平衡性得以保障。大多数的操作都和树的高度成比例，对于含有  $n$  个元素的 B 树，其性能为  $O(\lg n)$ 。

## 7.6 附录:例子程序

B 树的定义:

```
data BTree<K, Int deg> {
    [K] keys
    [BTree<K>] subStreets;
}
```

分拆节点:

```
void split(BTree<K, deg> z, Int i) {
    var d = deg
    var x = z.subTrees[i]
    var y = BTree<K, deg>()
    y.keys = x.keys[d ...]
    x.keys = x.keys[ ... d - 1]
    if not isLeaf(x) {
        y.subTrees = x.subTrees[d ... ]
        x.subTrees = x.subTrees[... d]
    }
    z.keys.insert(i, x.keys[d - 1])
    z.subTrees.insert(i + 1, y)
}
```

```
Bool isLeaf(BTree<K, deg> t) = t.subTrees == []
```

插入:

```
BTree<K, deg> insert(BTree<K, deg> tr, K key) {
    var root = tr
    if isFull(root) {
        var s = BTree<K, deg>()
        s.subTrees.insert(0, root)
        split(s, 0)
        root = s
    }
    return insertNonfull(root, key)
}
```

插入到未了的节点。

```
BTree<K, deg> insertNonfull(BTree<K, deg> tr, K key) {
    if isLeaf(tr) {
        orderedInsert(tr.keys, key)
    } else {
        Int i = length(tr.keys)
        while i > 0 and key < tr.keys[i - 1] {
            i = i - 1
        }
        if isFull(tr.subTrees[i]) {
            split(tr, i)
            if key > tr.keys[i] then i = i + 1
        }
    }
}
```

```

        insertNonfull(tr.subTree[i], key)
    }
    return tr
}

```

其中 `orderedInsert` 按序插入元素到列表中。

```

void orderedInsert([K] lst, K x) {
    Int i = length(lst)
    lst.append(x)
    while i > 0 and lst[i] < lst[i-1] {
        (lst[i-1], lst[i]) = (lst[i], lst[i-1])
        i = i - 1
    }
}

Bool isFull(BTree<K, deg> x) = length(x.keys) ≥ 2 * deg - 1
Bool isLow(BTree<K, deg> x) = length(x.keys) ≤ deg - 1

```

迭代查找:

```

Optional<(BTree<K, deg>, Int)> lookup(BTree<K, deg> tr, K key) {
    loop {
        Int i = 0, n = length(tr.keys)
        while i < n and key > tr.keys[i] {
            i = i + 1
        }
        if i < n and key == tr.keys[i] then return Optional.of((tr, i))
        if isLeaf(tr) {
            return Optional.Nothing
        } else {
            tr = tr.subTrees[i]
        }
    }
}

```

命令式先合并再删除:

```

BTree<K, deg> delete(BTree<K, deg> t, K x) {
    if empty(t.keys) then return t
    Int i = 0, n = length(t.keys)
    while i < n and x > t.keys[i] { i = i + 1 }
    if x == t.keys[i] {
        if isLeaf(t) { // case 1
            removeAt(t.keys, i)
        } else {
            var tl = t.subtrees[i]
            var tr = t.subtrees[i + 1]
            if not low(tl) { // case 2a
                t.keys[i] = max(tl)
                delete(tl, t.keys[i])
            } else if not low(tr) { // case 2b
                t.keys[i] = min(tr)
                delete(tr, t.keys[i])
            }
        }
    }
}

```



```

        } else { // case 2c
            mergeSubtrees(t, i)
            delete(d, tl, x)
            if empty(t.keys) then t = tl // shrink height
        }
    return t
}
if not isLeaf(t) {
    if x > t.keys[n - 1] then i = i + 1
    if low(t.subtrees[i]) {
        var tl = if i == 0 then null else t.subtrees[i - 1]
        var tr = if i == n then null else t.subtrees[i + 1]
        if tl != null and (not low(tl)) { // case 3a, left
            insert(t.subtrees[i].keys, 0, t.keys[i - 1])
            t.keys[i - 1] = popLast(tl.keys)
            if not isLeaf(tl) {
                insert(t.subtrees[i].subtrees, 0, popLast(tl.subtrees))
            }
        } else if tr != null and (not low(tr)) { // case 3a, right
            append(t.subtrees[i].keys, t.keys[i])
            t.keys[i] = popFirst(tr.keys)
            if not isLeaf(tr) {
                append(t.subtrees[i].subtrees, popFirst(tr.subtrees))
            }
        } else { // case 3b
            mergeSubtrees(t, if i < n then i else (i - 1))
            if i == n then i = i - 1
        }
        delete(t.subtrees[i], x)
        if empty(t.keys) then t = t.subtrees[0] // shrink height
    }
}
return t
}
}

```

合并子分枝,获取最大、最小元素。

```

void mergeSubtrees(BTree<K, deg>, Int i) {
    t.subtrees[i].keys += [t.keys[i]] + t.subtrees[i + 1].keys
    t.subtrees[i].subtrees += t.subtrees[i + 1].subtrees
    removeAt(t.keys, i)
    removeAt(t.subtrees, i + 1)
}

K max(BTree<K, deg> t) {
    while not empty(t.subtrees) {
        t = last(t.subtrees)
    }
    return last(t.keys)
}

K min(BTree<K, deg> t) {
    while not empty(t.subtrees) {

```

```
        t = t.subtrees[0]
    }
    return t.keys[0]
}
```

# 第八章 二叉堆

## 8.1 定义

堆是一种常见的数据结构,可以用于排序、带有优先级的调度、图算法等<sup>[40]</sup>。堆通常用数组表示的二叉树<sup>[4]</sup>实现。弗洛伊德(Robert W. Floyd)由此给出了一种高效的堆排序算法<sup>[41][42]</sup>。堆本身的定义是抽象的,也可由其它数据结构实现。本章介绍用二叉树实现的堆,包括左偏堆(Leftist heap)、斜堆(Skew heap)、伸展堆(splay heap)<sup>[3]</sup>。一个堆或为空,或包含若干可比较的元素。它满足一条性质并定义了如下操作:

1. **性质**:堆顶总保存着最小元素;
2. **弹出操作**移除堆顶元素,并保持堆的性质:新的堆顶元素仍是剩余中最小的;
3. **插入操作**将新元素加入堆中,并保持堆的性质;
4. **其它操作**(如合并两个堆)也保持堆的性质。

由于元素可比较,也可令堆顶总保存最大元素。我们称顶部保存最小元素的堆为**小顶堆**,顶部保存最大元素的堆为**大顶堆**。可以用树来实现堆。将最小(或最大)元素置于根节点。获取堆顶元素时,返回根节点中的数据。弹出堆顶时,将根节点删除,然后从子节点重新构建树。我们称使用二叉树实现的堆为**二叉堆**。

## 8.2 用数组实现的二叉堆

可以用数组表示一棵**完全二叉树**。它是一种“几乎”满的二叉树。深度为 $k$ 的满二叉树含有 $2^k - 1$ 个节点。如果将每个节点从上到下、从左向右编号为 $1, 2, \dots, 2^k - 1$ ,则完全二叉树中编号为 $i$ 的节点和满二叉树中编号为 $i$ 的节点在树中的位置相同。完全二叉树的叶子节点仅出现在最下面一行和倒数第二行。图 8.1 给出了一棵完全二叉树及其数组表示。数组中第 $i$ 个元素代表的节点,它的父节点对应第 $\lfloor i/2 \rfloor$ 个元素、左子树对应第 $2i$ 个元素、右子树对应第 $2i + 1$ 个元素。如果子节点的索引超出了数组的长度,则不含相应的子树(例如叶子节点)。树和数组之间的映射可以定义如下(令数组的索引从1开始,附录给出了位运算实现的例子):

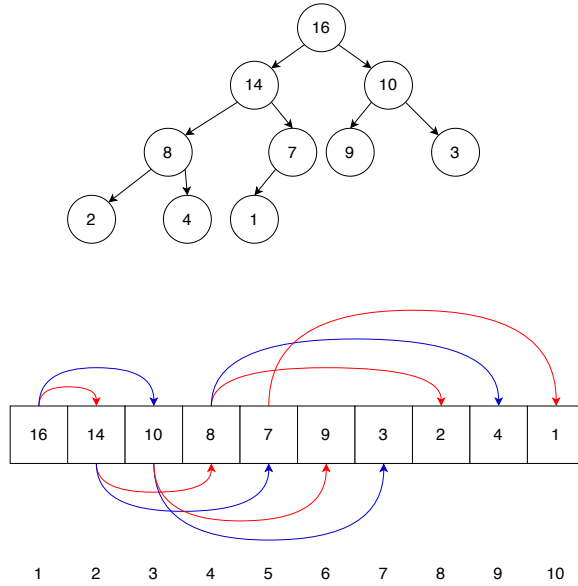


图 8.1: 完全二叉树到数组的映射

$$\begin{cases} \mathit{parent}(i) &= \lfloor \frac{i}{2} \rfloor \\ \mathit{left}(i) &= 2i \\ \mathit{right}(i) &= 2i + 1 \end{cases} \quad (8.1)$$

### 8.2.1 堆调整

堆调整<sup>1</sup>是维护堆性质的过程,使得堆顶元素为最小(或最大)。由于背后的数据模型是带有递归结构的二叉树,我们可以获得一个增强的堆性质:每棵子树的根节点都是最小(或最大)的。即任何子树都代表一个子堆。简单起见,我们考虑数组表示的小顶堆。任给数组下标  $i$ ,递归地检查  $i$  对应的子树中,左右子树的值是否大于等于 ( $\geq$ ) 根的值。不满足则交换,使得父节点保存最小值<sup>[4]</sup>。

```

1: function HEAPIFY( $A, i$ )
2:    $n \leftarrow |A|$ 
3:   loop
4:      $s \leftarrow i$ 
5:      $l \leftarrow \text{LEFT}(i), r \leftarrow \text{RIGHT}(i)$ 
6:     if  $l \leq n$  且  $A[l] < A[i]$  then
7:        $s \leftarrow l$ 
8:     if  $r \leq n$  且  $A[r] < A[s]$  then
9:        $s \leftarrow r$ 

```

▷  $s$  指向最小

<sup>1</sup>Heapify, 也译作堆化。

```

10:     if  $s \neq i$  then
11:         EXCHANGE  $A[i] \leftrightarrow A[s]$ 
12:          $i \leftarrow s$ 
13:     else
14:         return

```

由于自顶向下检查并调整堆,循环次数和完全二叉树的高度成正比。HEAPIFY 的时间复杂度为  $O(\lg n)$ ,其中  $n$  是元素个数。图 8.2 描述了小顶堆的调整步骤,索引从 2 开始,数组从  $[1, 13, 7, 3, 10, 12, 14, 15, 9, 16]$  变换为  $[1, 3, 7, 9, 10, 12, 14, 15, 13, 16]$ 。

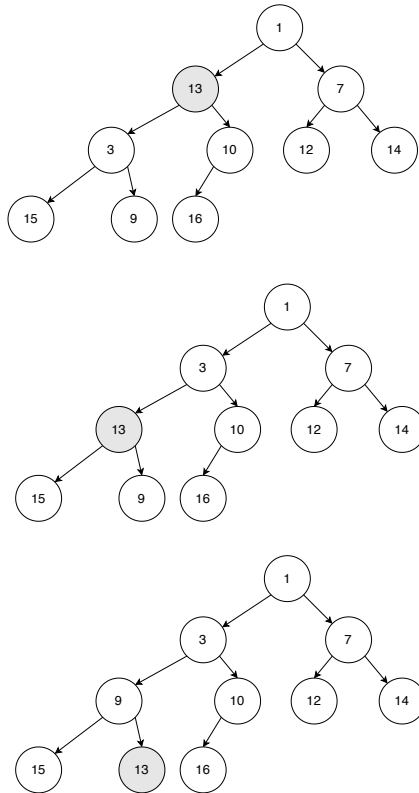


图 8.2: 堆调整。第一步:13、3、10 中最小值为 3,交换  $3 \leftrightarrow 13$ ;第二步:13、15、9 中最小值为 9,交换  $13 \leftrightarrow 9$ ;第三步:13 为叶子节点,调整结束。

### 8.2.2 构造堆

使用 HEAPIFY,我们可以从任意数组构造出堆。观察完全二叉树各层的节点数目:1, 2, 4, 8, ... 都是 2 的整数次幂,唯一例外是最后一层。由于树不一定满,最后一层最多含有  $2^{p-1}$  个节点,其中  $p$  是使  $2^p - 1 \geq n$  的最小整数, $n$  是数组长度。我们跳过所有叶子,从第一个分支节点(索引位于  $\lfloor n/2 \rfloor$ )开始不断向上执行 HEAPIFY。

```

1: function BUILD-HEAP( $A$ )
2:    $n \leftarrow |A|$ 
3:   for  $i \leftarrow \lfloor n/2 \rfloor$  down to 1 do
4:     HEAPIFY( $A, i$ )

```

虽然 HEAPIFY 的复杂度为  $O(\lg n)$ , 但是 BUILD-HEAP 的复杂度不是  $O(n \lg n)$ , 而是  $O(n)$ 。我们跳过了所有叶子, 最多有  $1/4$  的节点被比较并向下移动一次; 最多有  $1/8$  的节点被比较并向下移动两次; 最多有  $1/16$  的节点被比较并向下移动三次……总共比较和移动次数的上限为:

$$S = n\left(\frac{1}{4} + 2\frac{1}{8} + 3\frac{1}{16} + \dots\right) \quad (8.2)$$

将两侧都乘以 2:

$$2S = n\left(\frac{1}{2} + 2\frac{1}{4} + 3\frac{1}{8} + \dots\right) \quad (8.3)$$

用式 (8.3) 减去式 (8.2):

$$\begin{aligned}
 2S - S &= n\left[\frac{1}{2} + \left(2\frac{1}{4} - \frac{1}{4}\right) + \left(3\frac{1}{8} - 2\frac{1}{8}\right) + \dots\right] && \text{错开第一项两两相减} \\
 S &= n\left[\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots\right] && \text{等比级数和} \\
 &= n
 \end{aligned}$$

图 8.3 描述了从数组  $[4, 1, 3, 2, 16, 9, 10, 14, 8, 7]$  构造小顶堆的过程。黑色表示执行 HEAPIFY 的目标节点; 灰色表示进行交换的节点。

### 8.2.3 堆的基本操作

堆的基本操作包括获取顶部, 弹出顶部, 寻找最小(或最大)的前  $k$  个元素, 减小小顶堆中某一元素(或增大小顶堆中某一元素), 以及插入。二叉堆顶部对应的根节点, 是数组的第一个元素:  $Top(A) = A[1]$ 。

弹出堆顶后, 数组剩余元素顺次向前移动。除去根节点的剩余部分不再是一棵树。为了避免这种情况, 保持树结构, 我们将待移除的数组头部和末尾交换, 然后将数组的长度减一。这相当于删除叶子节点而非根节点。最后再用 HEAPIFY 恢复堆性质:

```

1: function POP( $A$ )
2:    $x \leftarrow A[1], n \leftarrow |A|$ 
3:   EXCHANGE  $A[1] \leftrightarrow A[n]$ 
4:   REMOVE( $A, n$ )
5:   if  $A$  is not empty then
6:     HEAPIFY( $A, 1$ )
7:   return  $x$ 

```

从数组的末尾删除元素仅需常数时间, 这样弹出的时间复杂度就取决于 HEAPIFY, 为  $O(\lg n)$ 。连续使用 pop, 可以找出一组元素中的前  $k$  个最小(或最大)元素:

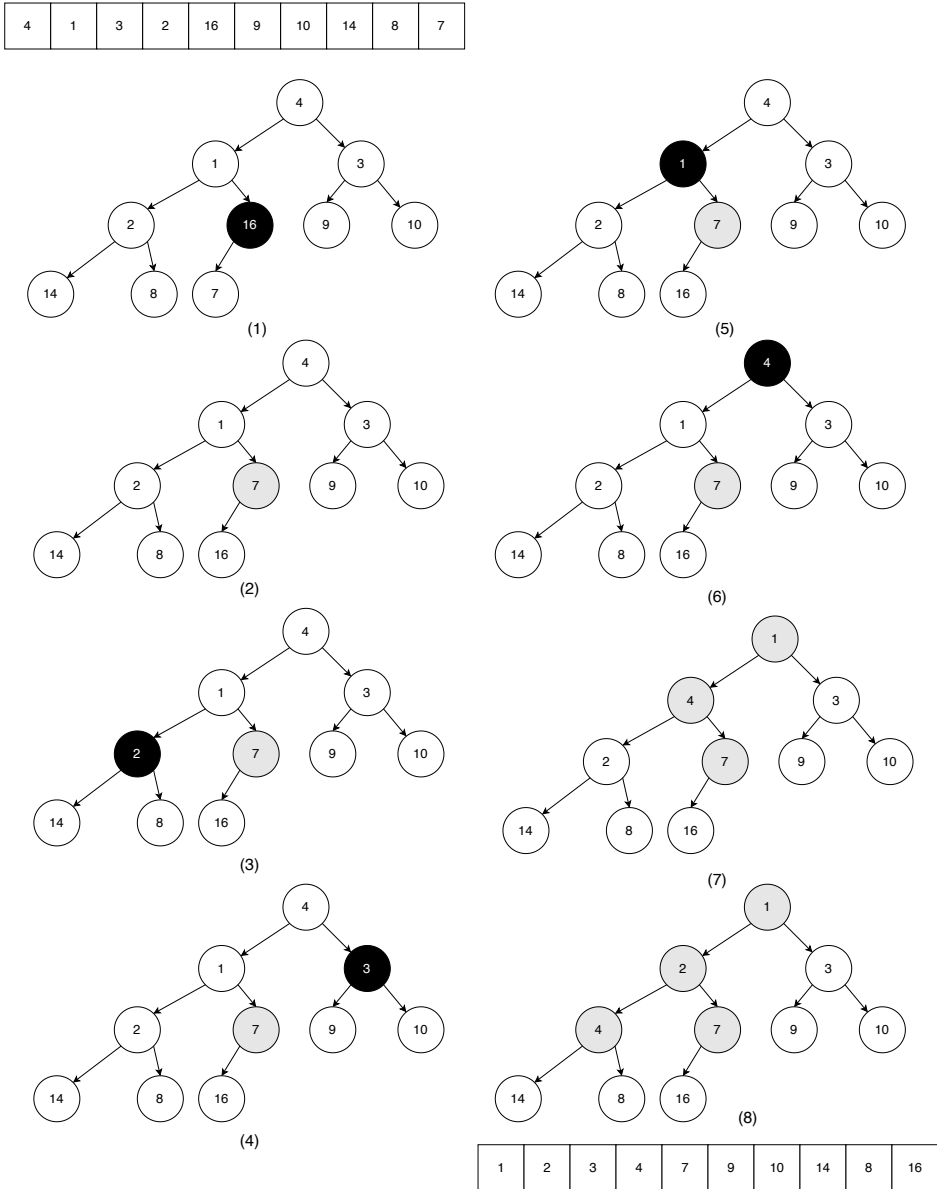


图 8.3: 构造堆。(1)检查 16, 它大于子节点 7; (2)交换  $16 \leftrightarrow 7$ ; (3)检查 2, 它比 14, 8 都小; (4)检查 3, 它比 9 和 10 都小; (5)检查 1, 它比 2 和 7 都小; (6)检查 4, 1 比 4 和 3 更小; (7)交换  $4 \leftrightarrow 1$ ; (8)交换  $4 \leftrightarrow 2$ , 结束。

```

1: function TOP-K( $A, k$ )
2:    $R \leftarrow []$ 
3:   BUILD-HEAP( $A$ )
4:   loop MIN( $k, |A|$ ) times ▷  $k$  超出长度则截断
5:     APPEND( $R, \text{POP}(A)$ )
6:   return  $R$ 

```

可以用堆实现“优先级队列”，进行任务调度。将若干任务放入堆中，每次从堆顶取出优先级最高的任务执行。为了尽早执行某个任务，可以提升它的优先级。对于小顶堆，这意味着减小某个元素的值，如图 8.4 所示。

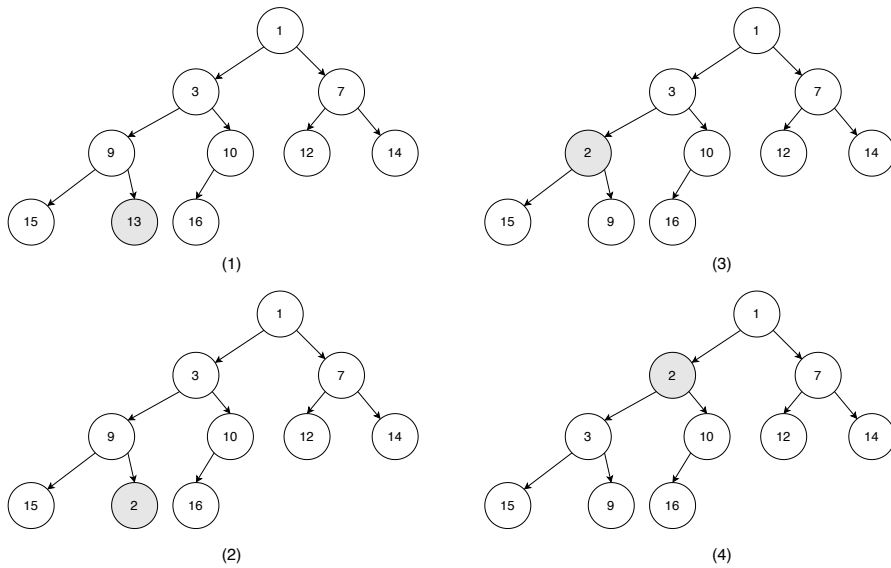


图 8.4: 将 13 减小为 2。2 先与 9 交换,然后再与 3 交换。

减小(小顶堆中的)某个元素可能会破坏堆性质。令元素索引为  $i$ , 下面的算法自底向上恢复堆性质, 时间复杂度为  $O(\lg n)$ 。

```

1: function HEAP-FIX( $A, i$ )
2:   while  $i > 1$  and  $A[i] < A[\text{PARENT}(i)]$  do
3:     EXCHANGE  $A[i] \leftrightarrow A[\text{PARENT}(i)]$ 
4:      $i \leftarrow \text{PARENT}(i)$ 

```

可以用 HEAP-FIX 实现插入<sup>[4]</sup>。先向数组末尾添加新元素  $k$ , 再使用 HEAP-FIX 调整:

```

1: function PUSH( $A, k$ )
2:   APPEND( $A, k$ )
3:   HEAP-FIX( $A, |A|$ )

```



### 8.2.4 堆排序

可以利用堆实现排序。从  $n$  个待排序元素构建一个小顶堆,不断弹出堆顶,就获得了升序结果。构建堆的时间复杂度是  $O(n)$ 。弹出操作的复杂度为  $O(\lg n)$ ,共执行了  $n$  次。总体时间复杂度为  $O(n \lg n)$ 。由于我们使用了另外一个列表存放排序结果,空间复杂度为  $O(n)$ 。

```

1: function HEAP-SORT( $A$ )
2:    $R \leftarrow []$ 
3:   BUILD-HEAP( $A$ )
4:   while  $A \neq []$  do
5:     APPEND( $R$ , POP( $A$ ))
6:   return  $R$ 

```

弗洛伊德给出了另一种实现。构造一个大顶堆。将堆顶和数组末尾交换,这样最大元素就存储到了排序后的正确位置。而原来的末尾元素变成了新的堆顶。将堆的大小减一,然后执行 HEAPIFY 恢复堆性质。重复这一过程,直到堆中仅剩一个元素。这一算法省去了额外的空间。

```

1: function HEAP-SORT( $A$ )
2:   BUILD-MAX-HEAP( $A$ )
3:    $n \leftarrow |A|$ 
4:   while  $n > 1$  do
5:     EXCHANGE  $A[1] \leftrightarrow A[n]$ 
6:      $n \leftarrow n - 1$ 
7:     HEAPIFY( $A[1..n]$ , 1)

```

### 练习 8.1

8.1.1. 考虑另外一种实现原地堆排序的想法: 第一步先从待排序数组构建一个最小堆  $A$ , 此时, 第一个元素  $a_1$  已经在正确的位置了。接下来, 将剩余的元素  $[a_2, a_3, \dots, a_n]$  当成一个新的堆, 并从  $a_2$  开始执行 HEAPIFY。重复这一从左向右的步骤完成排序。这一方法正确么?

```

1: function HEAP-SORT( $A$ )
2:   BUILD-HEAP( $A$ )
3:   for  $i = 1$  to  $n - 1$  do
4:     HEAPIFY( $A[i..n]$ , 1)

```

8.1.2. 类似地, 可以通过自左向右执行  $k$  遍 HEAPIFY 来实现 top- $k$  算法么?

```

1: function TOP-K( $A, k$ )
2:   BUILD-HEAP( $A$ )
3:    $n \leftarrow |A|$ 

```

```

4:   for  $i \leftarrow 1$  to  $\min(k, n)$  do
5:     HEAPIFY( $A[i..n], 1$ )

```

答案: 375 页

## 8.3 左偏堆和斜堆

弹出二叉堆的堆顶元素后, 剩余左右两棵子树, 它们也都是堆, 如图 8.5 所示。如何合并它们呢? 显然:

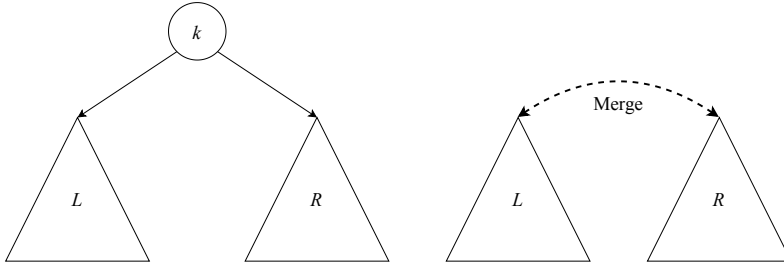


图 8.5: 弹出堆顶(删除根节点)后, 合并左右子树形成一个堆。

$$\begin{aligned}
 \text{merge } \emptyset R &= R \\
 \text{merge } L \emptyset &= L \\
 \text{merge } L R &= ?
 \end{aligned}$$

如果左右子树都不空, 为保持堆性质, 新的根节点必须是剩余元素中最小的。左右子树(它们也都是堆)的根节点分别保存了各自的最小元素, 我们选择较小的一个作为合并后的根。令  $L = (A, x, B)$ 、 $R = (A', y, B')$ , 其中  $A, A', B, B'$  都是子树。如果  $x < y$ ,  $x$  就是新的根。我们可以保留  $A$ , 递归地合并  $B$  和  $R$ ; 或者保留  $B$ , 递归地合并  $A$  和  $R$ 。新的堆为  $(\text{merge } A R, x, B)$  或  $(A, x, \text{merge } B R)$  之一。两个都可以, 为了简单, 我们总选择右子树进行合并。称这样的堆为**左偏堆**(*Leftist heap*)。

### 8.3.1 左偏堆

使用左偏树实现的堆称为左偏堆。克雷恩(C. A. Crane)于 1972 年<sup>[43]</sup>提出了左偏树。每个节点都定义了一个秩(也称作  $S$  值), 等于到最近的 NIL 节点的距离。NIL 节点的秩等于 0。如图 8.6 所示, 距离根节点 4 最近的叶子节点为 8, 所以根节点的秩为 2。节点 6 和 8 都是叶子, 它们的秩为 1。虽然节点 5 的左子树不为空, 但是它的右子树为空, 因此秩等于 1。记左右子树的秩分别为  $r_l, r_r$ , 定义合并策略如下:

1. 总是合并右子树;
2. 若  $r_l < r_r$ , 就交换左右子树。

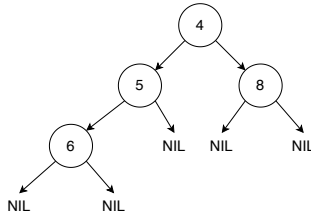


图 8.6:  $rank(4) = 2, rank(6) = rank(8) = rank(5) = 1$

我们称这样的合并策略为“左偏性质”。概括来说,在一棵左偏树中,到某个 NIL 的最短距离总在右侧。左偏树总是趋向不平衡,但是它可以维护一条重要的性质:

**定理 8.3.1.** 若一棵左偏树  $T$  包含  $n$  个节点,从根节点到达最右侧  $NIL$  的路径上最多含有  $\lfloor \log(n+1) \rfloor$  个节点。

我们省略了证明<sup>[44], [45]</sup>。根据此定理,沿着这一路径进行操作的算法,可以保证  $O(\lg n)$  的复杂度。记非空的左偏树为  $(r, L, k, R)$ , 分别表示秩、左子树、元素值、右子树:

```
data LHeap a = Empty | Node Int (LHeap a) a (LHeap a)
```

定义  $rank$  函数返回节点的秩。

$$\begin{aligned} rank \ \emptyset &= 0 \\ rank \ (r, L, k, R) &= r \end{aligned} \quad (8.4)$$

定义  $make$  函数比较左右子树的秩,并交换子树保持左偏性质。

$$make(A, k, B) = \begin{cases} rank(A) < rank(B) : & (1 + rank(A), B, k, A) \\ \text{否则} : & (1 + rank(B), A, k, B) \end{cases} \quad (8.5)$$

传入两棵子树  $A, B$  和元素  $k$ 。若  $A$  的秩较小,则用  $B$  作为左子树、 $A$  作为右子树。新节点的秩为  $1 + rank(A)$ ; 否则,若  $B$  的秩较小,就用  $A$  作为左子树、 $B$  作为右子树。新节点的秩为  $1 + rank(B)$ 。给定两个左偏堆  $H_1$  和  $H_2$ , 不空时记为  $(r_1, L_1, K_1, R_1)$ 、 $(r_2, L_2, k_2, R_2)$ , 合并操作定义为:

$$\begin{aligned} merge \ \emptyset \ H_2 &= H_2 \\ merge \ H_1 \ \emptyset &= H_1 \\ merge \ H_1 \ H_2 &= \begin{cases} k_1 < k_2 : & make(L_1, k_1, merge \ R_1 \ H_2) \\ \text{否则} : & make(L_2, k_2, merge \ H_1 \ R_2) \end{cases} \end{aligned} \quad (8.6)$$

$merge$  总在右子树上进行递归以保持左偏性质,保证了复杂度为  $O(\lg n)$ 。回顾上节,使用数组实现的堆在大多数情况下性能很好,但合并操作的时间复杂度为  $O(n)$ 。我们需要连接两个数组,再重新构建堆<sup>[50]</sup>。

- 1: **function** MERGE-HEAP( $A, B$ )
- 2:      $C \leftarrow \text{CONCAT}(A, B)$
- 3:     BUILD-HEAP( $C$ )

使用 *merge* 函数,可以实现基本的堆操作,例如在常数时间获取位于根节点的堆顶元素(设树不为空):

$$\text{top}(r, L, k, R) = k \quad (8.7)$$

弹出顶部后,将左右子树合并为一个新堆。弹出的复杂度和 *merge* 相同,也是  $O(\lg n)$ 。

$$\text{pop}(r, L, k, R) = \text{merge } L \ R \quad (8.8)$$

插入新元素  $k$  时,从  $k$  构造出一个叶子节点,然后将它合并:

$$\text{insert } k \ H = \text{merge}(1, \emptyset, k, \emptyset) \ H \quad (8.9)$$

或写成柯里化的形式  $\text{insert } k = \text{merge}(1, \emptyset, k, \emptyset)$ 。插入的复杂度也是  $O(\lg n)$ 。我们可以将一个列表中的元素依次插入,构造左偏堆(柯里化形式):  $\text{build} = \text{fold}_r \ \text{insert } \emptyset$ 。图 8.7 给出了一个构造左偏堆的例子。

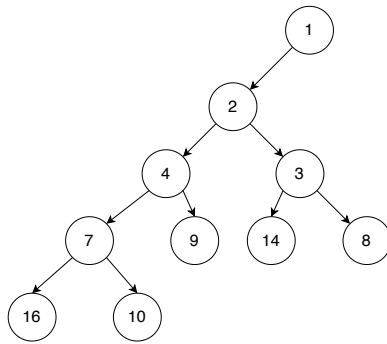


图 8.7: 从列表  $[9, 4, 16, 7, 10, 2, 14, 3, 8, 1]$  构造左偏堆

给定一个序列,将它转换成一个左偏堆,然后不断取出堆顶的最小元素就可以实现排序:

$$\text{sort} = \text{heapSort} \circ \text{build} \quad (8.10)$$

其中:

$$\begin{aligned} \text{heapSort } \emptyset &= [] \\ \text{heapSort } H &= (\text{top } H) : \text{heapSort}(\text{pop } H) \end{aligned} \quad (8.11)$$

弹出需要对数时间,并且被调用了  $n$  次,因此排序的总复杂度为  $O(n \lg n)$ 。

### 8.3.2 斜堆

左偏堆在某些情况下会产生不平衡的结构,如图 8.8 所示。斜堆(skew heap)是一种自调整堆,它简化了左偏堆并提高了平衡性<sup>[46][47]</sup>。构造左偏堆时,若左侧的秩小于右侧,则交换左右子树。但分支有一个 NIL 子节点时,不管树有多大,秩总为 1。斜堆简化了合并策略:每次都交换左右子树。

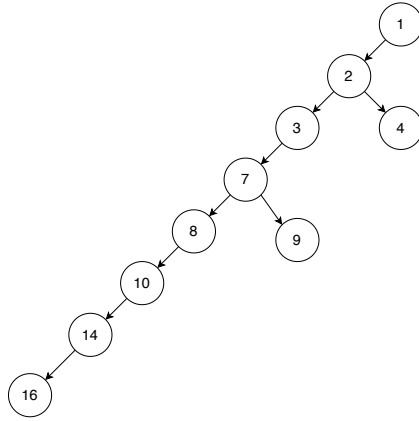


图 8.8: 从序列 [16, 14, 10, 8, 7, 9, 3, 2, 4, 1] 构造的左偏堆

斜堆由斜树(skew tree)实现。斜树是一种特殊的二叉树。最小的元素位于根节点,每棵子树也是一棵斜树。它无需秩,可直接复用二叉树的定义。树或者为空,或记为  $(L, k, R)$ 。

```
data SHeap a = Empty | Node (SHeap a) a (SHeap a)
```

合并两棵非空斜树  $H_1 = (L_1, k_1, R_1)$ 、 $H_2 = (L_2, k_2, R_2)$  时,若  $k_1 < k_2$ , 选择  $k_1$  作为新的根。我们既可以将  $H_2$  和  $L_1$  合并,也可以将  $H_2$  和  $R_1$  合并。我们选择合并到  $R_1$  上。然后交换左右子树,最后的结果为  $(merge\ R_1\ H_2, k_1, L_1)$ 。

$$\begin{aligned}
 merge\ \emptyset\ H_2 &= H_2 \\
 merge\ H_1\ \emptyset &= H_1 \\
 merge\ H_1\ H_2 &= \begin{cases} k_1 < k_2 : (merge\ R_1\ H_2, k_1, L_1) \\ \text{否则} : (merge\ H_1\ R_2, k_2, L_2) \end{cases} \quad (8.12)
 \end{aligned}$$

其它的操作,包括插入,获取和弹出顶部都和左偏树一样通过合并来实现。即使用斜堆处理已序序列,结果仍然是一棵较平衡的二叉树,如图 8.9 所示。

## 8.4 伸展堆

左偏堆和斜堆是直接用二叉树实现的堆。如果将二叉树换成二叉搜索树,则最小(或最大)元素就不再位于根节点。我们需要  $O(\lg n)$  时间来获取最小(或最大)元素。

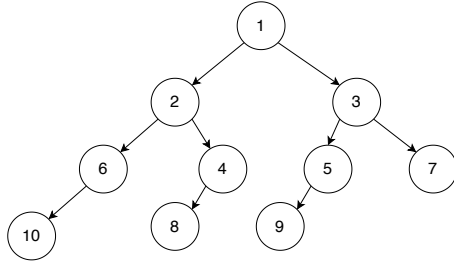


图 8.9: 用已序序列  $[1, 2, \dots, 10]$  构造的斜树

如果二叉搜索树不平衡,性能会下降。最坏情况下会退化为  $O(n)$ 。尽管可以用红黑树来保持平衡,伸展树提供了一种轻量级的实现方法,使得树不断趋向平衡。Sleator 和 Tarjan 在 1985 年引入了伸展树<sup>[48][49]</sup>。它采用类似缓存的策略,不断将正在访问的节点向根旋转,加快再次访问的速度。我们将这样的操作称为“伸展 (splay)”。不平衡的二叉搜索树经过若干次伸展后,会逐渐变得平衡。大多数伸展树操作的均摊性能是  $O(\lg n)$ 。

### 8.4.1 伸展操作

有两种方法可以实现伸展操作。第一种利用模式匹配,但需要处理较多的情况;第二种具备统一的形式,但是实现较为复杂。记正在访问的节点为  $x$ , 它的父节点为  $p$ , 如果存在祖父节点, 记为  $g$ 。伸展操作分为三个步骤, 每个步骤有两个对称的情况, 我们以每步中的一种情况举例说明, 如图 8.10 所示。

1. **zig-zig** 步骤,  $x$  和  $p$  都是左子树或都是右子树。通过两次旋转将  $x$  变成根节点。
2. **zig-zag** 步骤,  $x$  和  $p$  一左一右。经过旋转  $x$  变成根节点,  $p$  和  $g$  变成兄弟节点。
3. **zig** 步骤,  $p$  是根节点, 经过旋转  $x$  变成了根节点。这是伸展操作的最后一步。

共有 6 种不同的情况。记非空二叉树为  $T = (L, k, R)$ , 访问元素  $y$  时, 定义伸展操作如下:

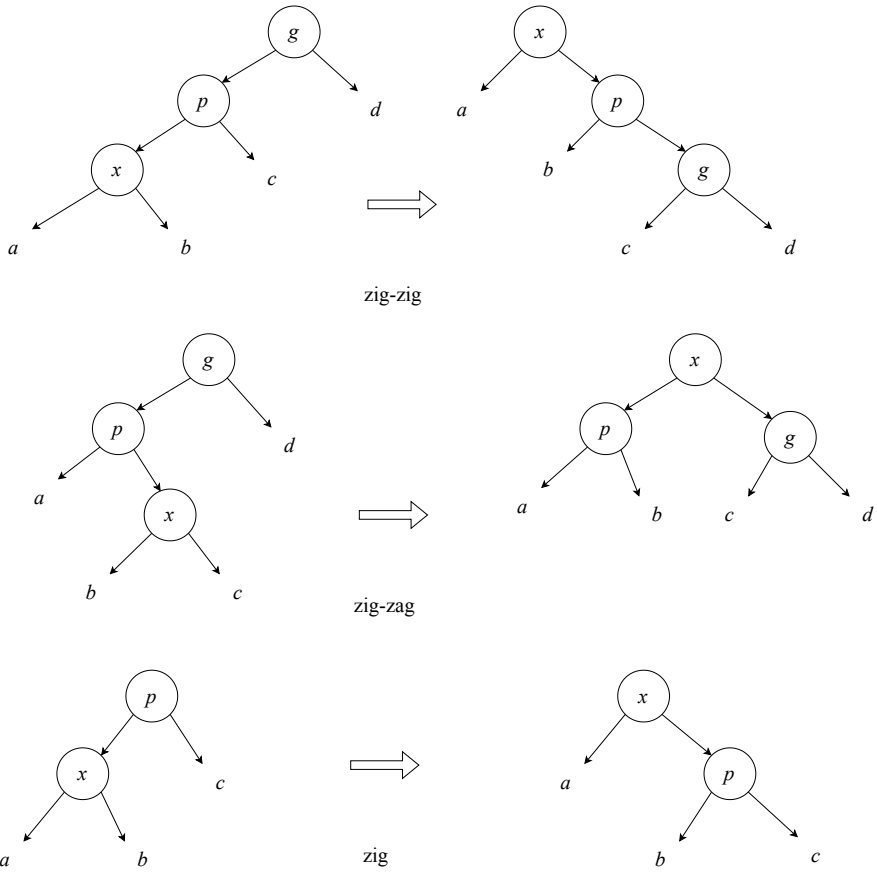


图 8.10: **zig-zig**:  $x$  和  $p$  都是左子树或都是右子树,  $x$  成为根节点。 **zig-zag**:  $x$  和  $p$  一左一右,  $x$  成为根节点,  $p$  和  $g$  变成兄弟节点。 **zig**:  $p$  是根节点, 旋转后  $x$  变为根节点。

$$\begin{aligned}
\text{splay } y ((a, x, b), p, c), g, d) &= \begin{cases} x = y : (a, x, (b, p, (c, g, d))) & \text{zig-zig} \\ \text{否则} : T & \end{cases} \\
\text{splay } y (a, g, (b, p, (c, x, d))) &= \begin{cases} x = y : (((a, g, b), p, c), x, d) & \text{zig-zig 对称} \\ \text{否则} : T & \end{cases} \\
\text{splay } y (a, p, (b, x, c), g, d) &= \begin{cases} x = y : ((a, p, b), x, (c, g, d)) & \text{zig-zag} \\ \text{否则} : T & \end{cases} \\
\text{splay } y (a, g, ((b, x, c), p, d)) &= \begin{cases} x = y : ((a, g, b), x, (c, p, d)) & \text{zig-zag 对称} \\ \text{否则} : T & \end{cases} \\
\text{splay } y ((a, x, b), p, c) &= \begin{cases} x = y : (a, x, (b, p, c)) & \text{zig} \\ \text{否则} : T & \end{cases} \\
\text{splay } y (a, p, (b, x, c)) &= \begin{cases} x = y : ((a, p, b), x, c) & \text{zig 对称} \\ \text{否则} : T & \end{cases} \\
\text{splay } y T &= T & \text{其它}
\end{aligned} \tag{8.13}$$

其它情况下树保持不变。每次插入时,执行伸展操作调整平衡。如果树为空,结果为一个叶子节点;否则比较待插入元素  $y$  和根节点  $x$ ,如果  $y < x$  就递归插入左子树,然后伸展;否则插入右子树再伸展。

$$\begin{aligned}
\text{insert } y \ \emptyset &= (\emptyset, y, \emptyset) \\
\text{insert } y (L, x, R) &= \begin{cases} y < x : \text{splay } y (\text{insert } y L, x, R) \\ \text{否则} : \text{splay } y (L, x, \text{insert } y R) \end{cases} \tag{8.14}
\end{aligned}$$

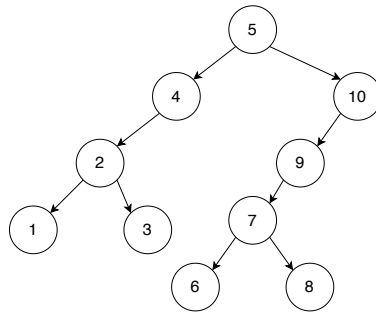


图 8.11: 从  $[1, 2, \dots, 10]$  产生的伸展树。

图 8.11 给出了逐一插入  $[1, 2, \dots, 10]$  的结果,产生的伸展树较平衡。冈崎发现了一条简单的伸展规则<sup>[3]</sup>:每次连续向左或向右访问两次就旋转节点。用当前访问的元素  $x$  将树  $T$  分割成两部分: $L$  和  $R$ ,使得  $L < x < R$ ( $L$  中元素都小于  $x$ , $R$  中元素都大于  $x$ )。然后以  $x$  为根, $L$ 、 $R$  为子树构建一棵新树。分割过程递归地对子树进行伸展。



$$\begin{aligned}
\text{partition } y \ \emptyset &= (\emptyset, \emptyset) \\
\text{partition } y \ (L, x, R) &= \left\{ \begin{array}{l} x < y \left\{ \begin{array}{l} R = \emptyset \quad (\emptyset, T) \\ R = (L', x', R') \left\{ \begin{array}{l} x' < y \quad (((L, x, L'), x', A), B) \\ \text{其中: } (A, B) = \text{partition } y \ R' \\ \text{否则} \quad ((L, x, A), (B, x', R')) \\ \text{其中: } (A, B) = \text{partition } y \ L' \end{array} \right. \\ \text{否则} \left\{ \begin{array}{l} L = \emptyset \quad (\emptyset, T) \\ L = (L', x', R') \left\{ \begin{array}{l} y < x' \quad (A, (L', x', (R', x, R))) \\ \text{其中: } (A, B) = \text{partition } y \ L' \\ \text{否则} \quad ((L', x', A), (B, x, R)) \\ \text{其中: } (A, B) = \text{partition } y \ R' \end{array} \right. \end{array} \right. \end{array} \right. \quad (8.15)
\end{aligned}$$

*partition* 接受一个基准值  $y$  和一棵树  $T$ 。如果树为空, 结果为  $(\emptyset, \emptyset)$ 。否则, 记  $T = (L, x, R)$ , 若  $x < y$ , 分为两种子情况: (1)  $R = \emptyset$ , 根据二叉搜索树的性质, 所有元素都小于  $y$ , 结果为  $(T, \emptyset)$ 。(2) 否则, 令  $R = (L', x', R')$ , 若  $x' < y$ , 递归地用基准值分割  $R'$ , 将所有小于  $y$  的元素放入树  $A$ , 其余放入树  $B$ 。结果为一对树:  $((L, x, L'), x', A)$  和  $B$ 。若  $x' > y$ , 递归地用  $y$  分割  $L'$  得  $(A, B)$ 。结果为一对树:  $(L, x, A)$  和  $(B, x', R')$ 。当  $y < x$  时, 情况是对称的。

可以用 *partition* 实现插入算法。向伸展堆  $T$  插入元素  $k$  时, 先将堆分割为两棵子树  $L$  和  $R$ , 使得  $L < k < R$  ( $L$  含有所有小于  $k$  的节点,  $R$  含有剩余的部分)。然后构造出树  $(L, k, R)$ 。

$$\text{insert } k \ T = (L, k, R), \text{ 其中: } (L, R) = \text{partition } k \ T \quad (8.16)$$

### 8.4.2 弹出顶部

伸展堆本质上是二叉搜索树, 最小的元素位于最左侧。通过向左遍历可以获取顶部元素:

$$\begin{aligned}
\text{top} (\emptyset, k, R) &= k \\
\text{top} (L, k, R) &= \text{top } L
\end{aligned} \quad (8.17)$$

这实际上就是二叉搜索树的 *min* 函数(也可以定义:  $\text{top} = \text{min}$ )。弹出时, 每当连续向左两次, 就进行伸展。

$$\begin{aligned}
\text{pop} (\emptyset, k, R) &= R \\
\text{pop} ((\emptyset, k', R'), k, R) &= (R', k, R) \\
\text{pop} ((L', k', R'), k, R) &= (\text{pop } L', k', (R', k, R))
\end{aligned} \quad (8.18)$$

第三条子式进行了伸展。它使用了二叉搜索树的性质,并未调用 *partition* 函数。伸展树平衡时,弹出操作的复杂度是  $O(\lg n)$ 。

### 8.4.3 合并

通过 *partition*,我们可以实现  $O(\lg n)$  时间的合并算法。合并两棵伸展树时,如果它们都不为空,可以将第一棵树的根节点作为新的根,然后将其作为基准值分割第二棵树。再递归地将第一棵树的子树合并。

$$\begin{aligned} \text{merge } T \ \emptyset &= T \\ \text{merge } \emptyset \ T &= T \\ \text{merge } (L, x, R) \ T &= ((\text{merge } L \ L') \ x \ (\text{merge } R \ R')) \end{aligned} \quad (8.19)$$

其中:

$$(L', R') = \text{partition } x \ T$$

## 8.5 小结

本章中,我们介绍了二叉堆概念。只要满足堆性质,可以使用各种形式的二叉树实现。用数组表示完全二叉树易于随机访问,适于命令式实现。直接用二叉树实现的堆,适于函数式实现。大部分操作的时间复杂度可以达到  $O(\lg n)$ ,有些操作的分摊复杂度是常数时间的。冈崎在<sup>[3]</sup>中给出了详细分析。如果将二叉树扩展成多叉树,就会得到更丰富的堆结构,如二项式堆、斐波那契堆、配对堆(见第十章)。

### 练习 8.2

8.2.1. 用命令式的方式实现左偏堆、斜堆。

8.2.2. 定义堆的叠加操作

答案: 375 页

## 8.6 附录:例子程序

在数组表示的完全二叉树中,利用位运算访问父节点和子树,索引从 0 开始:

```
Int parent(Int i) = ((i + 1) >> 1) - 1
Int left(Int i) = (i << 1) + 1
Int right(Int i) = (i + 1) << 1
```

堆调整,将元素间的比较运算抽象为参数:

```

void heapify([K] a, Int i, Less<K> lt) {
    Int l, r, m
    Int n = length(a)
    loop {
        m = i
        l = left(i)
        r = right(i)
        if l < n and lt(a[l], a[i]) then m = l
        if r < n and lt(a[r], a[m]) then m = r
        if m ≠ i {
            swap(a, i, m);
            i = m
        } else {
            break
        }
    }
}

```

从数组构造堆:

```

void buildHeap([K] a, Less<K> lt) {
    Int n = length(a)
    for Int i = (n-1) / 2 downto 0 {
        heapify(a, i, lt)
    }
}

```

弹出堆顶:

```

K pop([K] a, Less<K> lt) {
    var n = length(a)
    t = a[n]
    swap(a, 0, n - 1)
    remove(a, n - 1)
    if a ≠ [] then heapify(a, 0, lt)
    return t
}

```

寻找前  $k$  个元素:

```

[K] topk([K] a, Int k, Less<K> lt) {
    buildHeap(a, lt)
    [K] r = []
    loop min(k, length(a)) {
        append(r, pop(a, lt))
    }
    return r
}

```

减小堆中某元素的值:

```

void decreaseKey([K] a, Int i, K k, Less<K> lt) {
    if lt(k, a[i]) {
        a[i] = k
    }
}

```

```

        heapFix(a, i, lt)
    }
}

void heapFix([K] a, Int i, Less<K> lt) {
    while i > 0 and lt(a[i], a[parent(i)]) {
        swap(a, i, parent(i))
        i = parent(i)
    }
}

```

堆插入:

```

void push([K] a, K k, less<K> lt) {
    append(a, k)
    heapFix(a, length(a) - 1, lt)
}

```

堆排序:

```

void heapSort([K] a, less<K> lt) {
    buildHeap(a, not ◦ lt)
    n = length(a)
    while n > 1 {
        swap(a, 0, n - 1)
        n = n - 1
        heapify(a[0 .. (n - 1)], 0, not ◦ lt)
    }
}

```

左偏堆合并:

```

merge Empty h = h
merge h Empty = h
merge h1@(Node _ x l r) h2@(Node _ y l' r') =
    if x < y then makeNode x l (merge r h2)
    else makeNode y l' (merge h1 r')

makeNode x a b = if rank a < rank b then Node (rank a + 1) x b a
                else Node (rank b + 1) x a b

```

斜堆合并:

```

merge Empty h = h
merge h Empty = h
merge h1@(Node x l r) h2@(Node y l' r') =
    if x < y then Node x (merge r h2) l
    else Node y (merge h1 r') l'

```

伸展操作:

```

— zig-zig
splay t@(Node (Node (Node a x b) p c) g d) y =
    if x == y then Node a x (Node b p (Node c g d)) else t

```

```

splay t@(Node a g (Node b p (Node c x d))) y =
  if x == y then Node (Node (Node a g b) p c) x d else t
— zig-zag
splay t@(Node (Node a p (Node b x c)) g d) y =
  if x == y then Node (Node a p b) x (Node c g d) else t
splay t@(Node a g (Node (Node b x c) p d)) y =
  if x == y then Node (Node a g b) x (Node c p d) else t
— zig
splay t@(Node (Node a x b) p c) y = if x == y then Node a x (Node b p c) else t
splay t@(Node a p (Node b x c)) y = if x == y then Node (Node a p b) x c else t
— 否则
splay t _ = t

```

伸展堆的插入:

```

insert Empty y = Node Empty y Empty
insert (Node l x r) y
  | x > y    = splay (Node (insert l y) x r) y
  | otherwise = splay (Node l x (insert r y)) y

```

伸展树的分割:

```

partition Empty _ = (Empty, Empty)
partition t@(Node l x r) y
  | x < y =
    case r of
      Empty → (t, Empty)
      Node l' x' r' →
        if x' < y then
          let (small, big) = partition r' y in
            (Node (Node l x l') x' small, big)
        else
          let (small, big) = partition l' y in
            (Node l x small, Node big x' r')
  | otherwise =
    case l of
      Empty → (Empty, t)
      Node l' x' r' →
        if y < x' then
          let (small, big) = partition l' y in
            (small, Node l' x' (Node r' x r))
        else
          let (small, big) = partition r' y in
            (Node l' x' small, Node big x r)

```

伸展树合并:

```

merge t Empty = t
merge Empty t = t
merge (Node l x r) t = Node (merge l l') x (merge r r')
  where (l', r') = partition t x

```



## 第九章 选择排序

本章介绍另一种直观的排序方法——选择排序。它在性能上不如快速排序和归并排序等分治算法。我们给出选择排序性能的简要分析,并且从不同的角度加以改进,最终演进到堆排序,从而达到基于比较的排序算法性能上限  $O(n \lg n)$ 。选择排序的思想可以在日常生活中找到。观察孩子们吃葡萄时,会发现两种类型的吃法:一种属于“乐观型”,每次吃掉最大的一颗;另一种属于“悲观型”,每次总吃掉最小的一颗。第一种孩子实际上按照由大到小的顺序吃葡萄;第二种按照由小到大的顺序吃葡萄。实际上,孩子们把葡萄按照大小进行了选择排序。选择排序的算法描述为:

1. 如果序列为空,排序结果也为空;
2. 否则,找到最小的元素,将其附加到结果的后面。

这一算法产生升序结果。如果每次选择最大的元素,则结果是降序的。我们可以用抽象的比较操作实现排序。

$$\begin{aligned} \text{sort } [] &= [] \\ \text{sort } A &= m : \text{sort } (A - [m]) \quad \text{其中 } m = \min A \end{aligned} \tag{9.1}$$

$A - [m]$  从序列  $A$  中去除元素  $m$ 。对应的命令式描述为:

```
1: function SORT( $A$ )
2:    $X \leftarrow []$ 
3:   while  $A \neq []$  do
4:      $x \leftarrow \text{MIN}(A)$ 
5:      $\text{DEL}(A, x)$ 
6:      $\text{APPEND}(X, x)$ 
7:   return  $X$ 
```

图 9.1 描述了选择排序的过程。作为改进,我们可以在  $A$  中进行原地排序,去掉列表  $X$ 。将最小的元素保存在  $A[1]$ ,将次小的元素保存在  $A[2]$ ……。可以通过交换位置实现:当找到第  $i$  小的元素后,将它和  $A[i]$  交换。

```
1: function SORT( $A$ )
2:   for  $i \leftarrow 1$  to  $|A|$  do
3:      $m \leftarrow \text{MIN-AT}(A, i)$ 
```

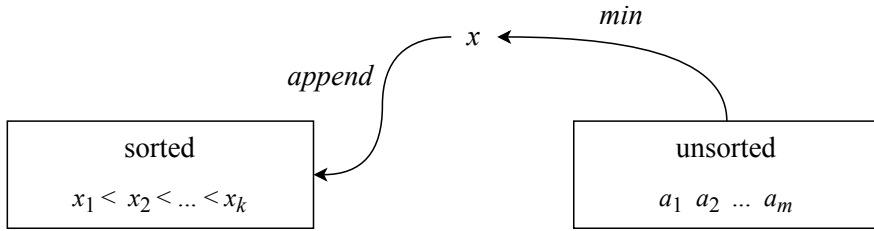


图 9.1: 左侧部分为已序元素,不断从剩余部分选择最小元素附加在左侧尾部

4:           EXCHANGE  $A[i] \leftrightarrow A[m]$

令  $A = [a_1, a_2, \dots, a_n]$ , 当处理第  $i$  个元素时,  $[a_1, a_2, \dots, a_{i-1}]$  都已排序。我们找到  $[a_i, a_{i+1}, \dots, a_n]$  中的最小元素, 将其和  $a_i$  交换, 这样第  $i$  个位置就保存了正确的元素。重复这一过程直到最后一个元素。图 9.2 描述了这一思路。

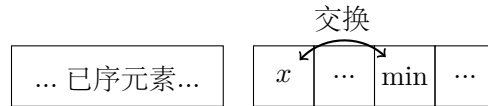


图 9.2: 左侧部分为已序元素,不断从剩余部分找到最小的交换到正确位置

## 9.1 查找最小元素

我们可以用比较——交换方法在一组元素中寻找最小值。将元素编号为  $1, 2, \dots, n$ 。比较编号为  $1, 2$  的两个元素, 选择较小的留下与编号为  $3$  的元素比较……重复这一步骤直到第  $n$  号元素。这一方法适合处理数组。

```

1: function MIN-AT( $A, i$ )
2:    $m \leftarrow i$ 
3:   for  $i \leftarrow m + 1$  to  $|A|$  do
4:     if  $A[i] < A[m]$  then
5:        $m \leftarrow i$ 
6:   return  $m$ 

```

MIN-AT 查找片断  $A[i..n]$  中最小元素的位置  $m$ 。令  $m$  指向第一个元素  $A[i]$ , 并逐一检查元素  $A[i+1], A[i+2], \dots$ 。也可以用递归的方法在在一组元素  $L$  中查找最小值。如果  $L$  只有一个元素, 它就是最小值。否则从  $L$  中取出一个元素  $x$ , 然后在剩余部分中递归找到最小值  $y$ 。  $x, y$  中较小的一个就是最终的最小值。

$$\min [x] = (x, [])$$

$$\min (x:xs) = \begin{cases} x < y : (x, xs), & \text{其中: } (y, ys) = \min xs \\ \text{否则: } (y, x:ys) \end{cases} \quad (9.2)$$



可以进一步用尾递归优化。将全部元素分成两组： $A$ 、 $B$ 。开始时  $A$  为空 ( $[]$ )， $B$  包含全部元素。我们从  $B$  中任选两个元素比较，将较大的放入  $A$ ，而留下较小的记为  $m$ 。此后不断从  $B$  中任取一个元素，和  $m$  对比直到  $B$  变成空。这时， $m$  就是最小值。在任何时候有不变关系： $L = A \# [m] \# B$ ，对任意  $a \in A, b \in B$  有  $a \leq m \leq b$ 。

$$\min (x:xs) = \min' [] x xs \quad (9.3)$$

其中：

$$\begin{aligned} \min' as m [] &= (m, as) \\ \min' as m (b:bs) &= \begin{cases} b < m : \min' (m:as) b bs \\ \text{否则} : \min' (b:as) m bs \end{cases} \end{aligned} \quad (9.4)$$

函数  $\min$  返回一对值：最小元素和剩余元素列表。这样选择排序就可以实现为：

$$\begin{aligned} \text{sort} [] &= [] \\ \text{sort} xs &= m : (\text{sort} xs'), \text{ 其中} : (m, xs') = \min xs \end{aligned} \quad (9.5)$$

### 9.1.1 选择排序的性能

选择排序在每轮中检查所有未排好的元素以挑选出最小值。总共进行了  $n$  次挑选， $n + (n - 1) + (n - 2) + \dots + 1$  次比较，时间复杂度为  $O(\frac{n(n+1)}{2}) = O(n^2)$ 。和插入排序相比，选择排序在最好、最差和平均情况下的性能是相同的，而插入排序在最好情况下性能为线性时间  $O(n)$ （元素逆序存储在一个链表中），最差情况下性能为平方时间  $O(n^2)$ 。

## 练习 9.1

9.1.1. 下面的尾递归查找最小值实现有何问题？

$$\begin{aligned} \min' as m [] &= (m, as) \\ \min' as m (b:bs) &= \begin{cases} b < m : \min' (as \# [m]) b bs \\ \text{否则} : \min' (as \# [b]) m bs \end{cases} \end{aligned}$$

9.1.2. 实现原地选择排序程序。

答案：[377 页](#)

## 9.2 改进

为了支持升序、降序、和不同的比较，我们可以将比较操作抽出作为参数  $\triangleleft$ 。

$$\begin{aligned} \text{sortBy} \triangleleft [] &= [] \\ \text{sortBy} \triangleleft xs &= m : \text{sortBy} \triangleleft xs', \text{ 其中} : (m, xs') = \minBy \triangleleft xs \end{aligned} \quad (9.6)$$

“最小值”也使用  $\triangleleft$  进行比较:

$$\begin{aligned} \text{minBy } \triangleleft [x] &= (x, []) \\ \text{minBy } \triangleleft (x:xs) &= \begin{cases} x \triangleleft y : (x, xs), \text{ 其中: } (y, ys) = \text{minBy } \triangleleft xs \\ \text{否则: } (y, x:ys) \end{cases} \quad (9.7) \end{aligned}$$

对于一组整数,传入小于号得到升序结果: $\text{sortBy } (<) [3, 1, 4, \dots]$ 。这里要求  $\triangleleft$  满足严格弱序<sup>[52]</sup>条件:

- 非自反性:对任何  $x, x \not\triangleleft x$ ;
- 非对称性:对任何  $x, y$ ,若  $x \triangleleft y$ ,则  $y \not\triangleleft x$ ;
- 传递性:对任何  $x, y, z$ ,若  $x \triangleleft y$  且  $y \triangleleft z$ ,则  $x \triangleleft z$ 。

命令式原地选择排序遍历所有的元素,并在内冲循环中查找最小值:

```
1: procedure SORT(A)
2:   for  $i \leftarrow 1$  to  $|A|$  do
3:      $m \leftarrow i$ 
4:     for  $j \leftarrow i + 1$  to  $|A|$  do
5:       if  $A[i] < A[m]$  then
6:          $m \leftarrow i$ 
7:     EXCHANGE  $A[i] \leftrightarrow A[m]$ 
```

前  $n - 1$  个元素排好后,最后一个元素必然是第  $n$  大的。无需再进行一次最小值查找。这样可以减少一次外重循环。另外,如果第  $i$  大的元素恰好是  $A[i]$ ,则无需交换:

```
1: procedure SORT(A)
2:   for  $i \leftarrow 1$  to  $|A| - 1$  do
3:      $m \leftarrow i$ 
4:     for  $j \leftarrow i + 1$  to  $|A|$  do
5:       if  $A[i] < A[m]$  then
6:          $m \leftarrow i$ 
7:     if  $m \neq i$  then
8:       EXCHANGE  $A[i] \leftrightarrow A[m]$ 
```

### 9.2.1 鸡尾酒排序

高德纳给出了另一种实现<sup>[51]</sup>:每次不是查找最小元素,而是最大元素,将其放在末尾位置。如图 9.3 所示,任何时候,最右侧的元素都是已序的。算法扫描未排序元素,定位到其中的最大值,然后交换到未排序部分的末尾。

```
1: procedure SORT'(A)
2:   for  $i \leftarrow |A|$  down-to 2 do
```

```

3:    $m \leftarrow i$ 
4:   for  $j \leftarrow 1$  to  $i - 1$  do
5:       if  $A[m] < A[j]$  then
6:            $m \leftarrow j$ 
7:   EXCHANGE  $A[i] \leftrightarrow A[m]$ 

```

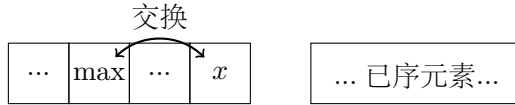


图 9.3: 每次选择最大的元素放到末尾

进一步, 每次扫描可以同时查找最小、最大值, 将最小值放到开头, 最大值放到末尾。这样可以将外重循环次数减半。这一算法称为“鸡尾酒排序”。如图 9.4 所示。

```

1: procedure SORT( $A$ )
2:   for  $i \leftarrow 1$  to  $\lfloor \frac{|A|}{2} \rfloor$  do
3:        $min \leftarrow i$ 
4:        $max \leftarrow |A| + 1 - i$ 
5:       if  $A[max] < A[min]$  then
6:           EXCHANGE  $A[min] \leftrightarrow A[max]$ 
7:       for  $j \leftarrow i + 1$  to  $|A| - i$  do
8:           if  $A[j] < A[min]$  then
9:                $min \leftarrow j$ 
10:          if  $A[max] < A[j]$  then
11:               $max \leftarrow j$ 
12:          EXCHANGE  $A[i] \leftrightarrow A[min]$ 
13:          EXCHANGE  $A[|A| + 1 - i] \leftrightarrow A[max]$ 

```

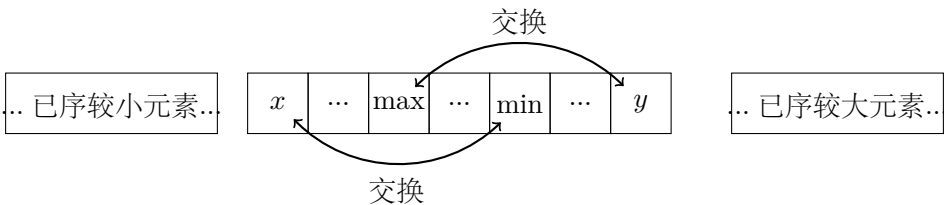


图 9.4: 一次扫描同时定位出最小和最大元素, 然后将它们放到正确的位置

在内重循环开始前, 如果最右侧小于最左侧的元素, 需要进行交换。这是因为扫描范围不包括两端。也可以用递归的方式实现鸡尾酒排序:

1. 若待排序的序列为空或者仅含有一个元素, 排序结果为原序列;

2. 否则, 找到最小和最大值, 分别放到开头和结尾位置, 然后递归地将剩余元素排序。

$$\begin{aligned} \text{sort } [] &= [] \\ \text{sort } [x] &= [x] \\ \text{sort } xs &= a : (\text{sort } xs') \# [b], \text{ 其中 } : (a, b, xs') = \text{min-max } xs \end{aligned} \quad (9.8)$$

其中, 函数 *min-max* 从序列中抽取出最小值和最大值:

$$\text{min-max } (x:y:xs) = \text{foldr sel } (\text{min } x y, \text{max } x y, []) xs \quad (9.9)$$

我们选取前两个元素作为已找到的最小值  $x_0$ , 最大值  $x_1$ , 用 *foldr* 扫描序列, 其中 *sel* 定义为:

$$\text{sel } x (x_0, x_1, xs) = \begin{cases} x < x_0 : & (x, x_1, x_0:xs) \\ x_1 < x : & (x_0, x, x_1:xs) \\ \text{否则} : & (x_0, x_1, x:xs) \end{cases}$$

尽管 *min-max* 只需要线性时间  $O(n)$ , 但  $\#[b]$  的代价较大。如图 9.4 所示, 令左侧  $A$  包含较小的已序元素; 右侧  $B$  包含较大的已序元素。用  $A$  和  $B$  作为累积器, 可将排序转换为尾递归:

$$\begin{aligned} \text{sort}' A B [] &= A \# B \\ \text{sort}' A B [x] &= A \# (x:B) \\ \text{sort}' A B (x:xs) &= \text{sort}' (A \# [x_0]) xs' (x_1:B), \text{ 其中 } : (x_0, x_1, xs') = \text{min-max } xs \end{aligned} \quad (9.10)$$

传入空的  $A, B$  启动排序:  $\text{sort} = \text{sort}' [] []$ 。追加操作仅仅发生在  $A \# [x_0]$ ; 而  $x_1$  则被链结到  $B$  的前面。为了消除追加操作, 我们将  $A$  保存为逆序  $\overleftarrow{A}$ , 这样就可以将  $x_0$  链结到前面。我们有如下等价关系:

$$\begin{aligned} A' &= A \# [x] \\ &= \text{reverse } (x : \text{reverse } A) \\ &= \text{reverse } (x : \overleftarrow{A}) \\ &= \overleftarrow{\overleftarrow{x : \overleftarrow{A}}} \\ &= x : \overleftarrow{A} \end{aligned} \quad (9.11)$$

最后执行一次反转操作将  $\overleftarrow{A'}$  转换回  $A'$ 。根据这一思路, 可改进如下:

$$\begin{aligned} \text{sort}' A B [] &= (\text{reverse } A) \# B \\ \text{sort}' A B [x] &= (\text{reverse } x:A) \# B \\ \text{sort}' A B (x:xs) &= \text{sort}' (x_0:A) xs' (x_1:B) \end{aligned} \quad (9.12)$$

## 9.3 继续改进

虽然鸡尾酒排序将循环次数减半,但时间复杂度仍然是  $O(n^2)$ 。通过比较进行排序,需要检查元素间的大小顺序,外重循环是必须的。必需每次扫描全部元素以查找最小值么? 在第一次查找最小元素时,我们遍历了序列,知道哪些元素相对较小,哪些相对较大。但在后继查找中,我们没有复用这些相对大小的信息,而是从头开始再次遍历。进一步改进的关键在于重用已有的结果。其中一种思路来自体育竞赛。

### 9.3.1 锦标赛淘汰法

足球世界杯每四年举办一次。来自各个大洲的 32 支球队最终进入决赛。1982 年前,决赛阶段只有 16 支球队。我们回到 1978 年,并且想像一种特殊的方法来决定冠军:在第一轮比赛中,所有参赛球队被分为 8 组进行比赛;比赛产生 8 支获胜球队,其余 8 支被淘汰。接下来,在第二轮比赛中,8 支球队被分成 4 组。比赛产生 4 支获胜球队;然后这 4 支球队分成两对,比赛产生最终的两支球队争夺冠军。经过 4 轮比赛,冠军就可产生。总共的比赛场次为:  $8 + 4 + 2 + 1 = 15$ 。但是我们并不满足仅仅知道谁是冠军,我们还想知道哪支球队是亚军。有人会问最后一场比赛中,被冠军击败的队伍不是亚军么? 在真实的世界杯中,的确如此。但是这个规则在某种程度上并不公平。我们常常听说过“死亡之组”,假设巴西队一开始就和德国队进行比赛。虽然它们两个都是强队,但是必须有一支在一上来就被淘汰。这支被淘汰的球队,很可能会打败除冠军外的其他所有球队。图 9.5 描述了这一情况。

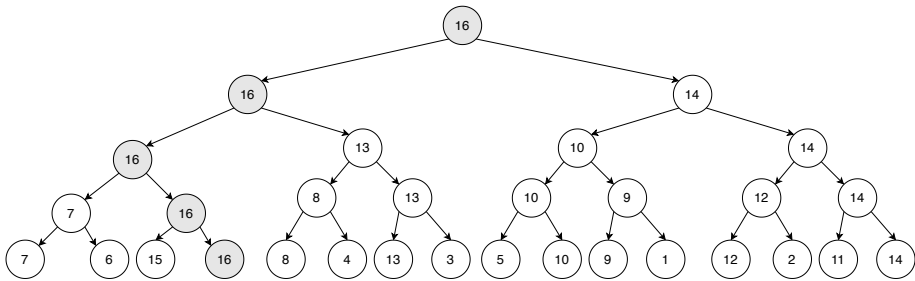
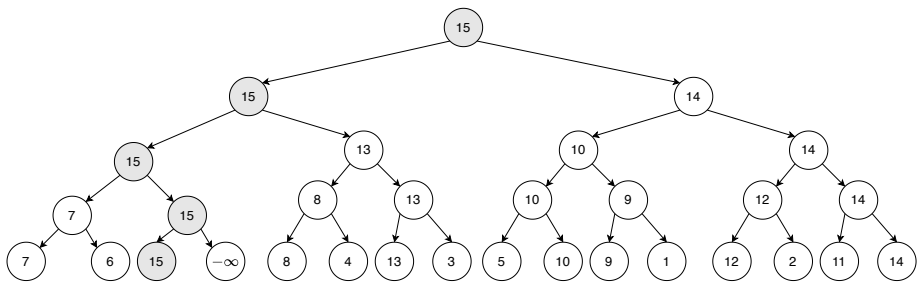


图 9.5: 元素 15 在第一轮就被淘汰

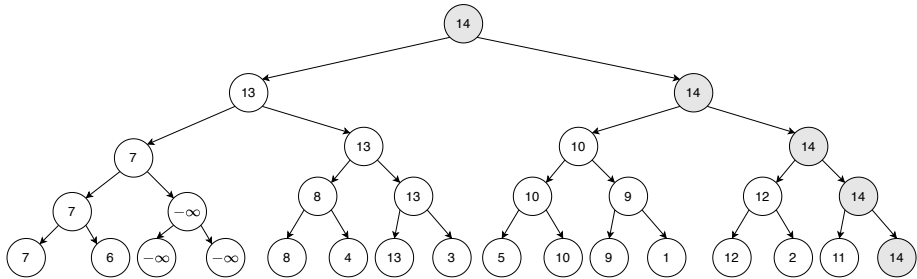
每支队伍有一个代表其实力的数字。数字越大,实力越强。假设数字较大的队永远会战胜数字较小的队(虽然现实中不会这样)。代表冠军的数字为 16,根据假设的规则,数字 14 不是亚军,而是在第一轮就被淘汰的 15。我们需要一种快速的方法在锦标赛树中找到次大值。此后,不断重复这一方法,找出第三大,第四大……就可以完成基于选择的排序。把冠军的数字修改成一个很小的值(例如  $-\infty$ ),这样以后它不会被选中,第二名就会成为新的冠军。假设有  $2^m$  支球队,  $m$  是自然数,仍然需要  $2^{m-1} + 2^{m-2} + \dots + 2 + 1 = 2^m - 1$  次比较才能产生新的冠军,这和第一次寻找冠军花费的代价相同。实际上,我们无需再进行自底向上的比较。锦标赛树中保存了足够的

顺序信息。实力第二强的队，一定在某个时刻被冠军击败，否则它就会是最终的冠军。因此我们可以从锦标赛树的根节点出发，沿着产生冠军的路径向叶子方向遍历，在这条路径上寻找第二强的队。图 9.5 中，这条路径被标记为灰色，需要检查的元素包括 [14, 13, 7, 15]，这一思路可以描述如下：

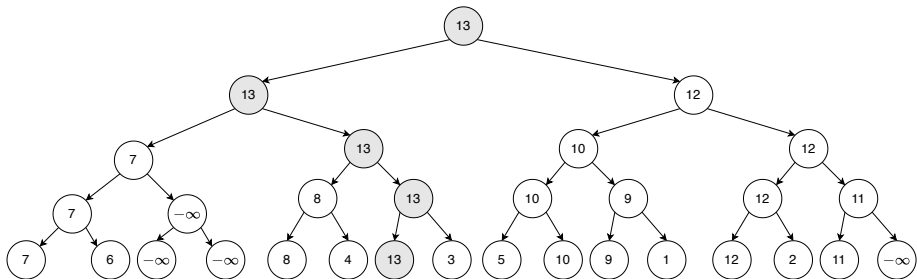
1. 从待排序元素构建一棵锦标赛树，冠军(最大值)位于树根；
2. 取出树根，自顶向下沿着冠军路径将最大值替换为  $-\infty$ ；
3. 自底向上沿着刚才的路径回溯，找出新的冠军，并将其置于树根；
4. 重复步骤 2，直到所有的元素都被取出。



取出 16，将其替换为  $-\infty$ ，15 上升为新的根



取出 15，将其替换为  $-\infty$ ，14 上升为新的根



取出 14，将其替换为  $-\infty$ ，13 上升为新的根

图 9.6: 锦标赛树排序的前几步

为了对一组元素排序，我们先从它们构造一棵锦标赛树，然后不断取出冠军，图 9.6 给出了这一排序的前几个步骤。复用二叉树的定义，为了方便回溯，每个节点需

要指向它的父节点。如果元素数目  $n$  不是  $2^m$ , 两两比较后会剩余元素没有“对手”, “轮空”进入下一轮。为了构造锦标赛树, 从每个元素构造一棵叶子得到  $n$  棵二叉树。然后每次取出两棵树  $t_1, t_2$ , 构造一棵更大的二叉树  $t$ 。其中  $t$  的根为  $\max(\text{key}(t_1), \text{key}(t_2))$ , 左右子树为  $t_1, t_2$ 。重复这一步骤得到一组新树, 每棵的高度增加 1, 如有剩余则进入下一轮(奇数棵)。这样每轮过后, 树减半为  $\lfloor \frac{n}{2} \rfloor, \lfloor \frac{n}{4} \rfloor, \dots$  持续同样的操作最终得到一棵锦标赛树。总时间复杂度为  $O(n + \frac{n}{2} + \frac{n}{4} + \dots) = O(2n) = O(n)$ 。

```

1: function BUILD-TREE( $A$ )
2:    $T \leftarrow []$ 
3:   for each  $x \in A$  do
4:     APPEND( $T$ , NODE(NIL,  $x$ , NIL))
5:   while  $|T| > 1$  do
6:      $T' \leftarrow []$ 
7:     for every  $t_1, t_2 \in T$  do
8:        $k \leftarrow \text{MAX}(\text{KEY}(t_1), \text{KEY}(t_2))$ 
9:       APPEND( $T'$ , NODE( $t_1$ ,  $k$ ,  $t_2$ ))
10:    if  $|T|$  is odd then
11:      APPEND( $T'$ , LAST( $T$ ))
12:     $T \leftarrow T'$ 
13:  return  $T[1]$ 

```

每次取出锦标赛树的根节点后, 我们自顶向下将其替换为  $-\infty$ , 然后在通过父节点向上回溯, 找出新的最大值。

```

1: function POP( $T$ )
2:    $m \leftarrow \text{KEY}(T)$ 
3:    $\text{KEY}(T) \leftarrow -\infty$ 
4:   while  $T$  is not leaf do ▷ 自顶向下将  $m$  替换为  $-\infty$ 
5:     if  $\text{KEY}(\text{LEFT}(T)) = m$  then
6:        $T \leftarrow \text{LEFT}(T)$ 
7:     else
8:        $T \leftarrow \text{RIGHT}(T)$ 
9:    $\text{KEY}(T) \leftarrow -\infty$ 
10:  while  $\text{PARENT}(T) \neq \text{NIL}$  do ▷ 自底向上决出新冠军
11:     $T \leftarrow \text{PARENT}(T)$ 
12:     $\text{KEY}(T) \leftarrow \text{MAX}(\text{KEY}(\text{LEFT}(T)), \text{KEY}(\text{RIGHT}(T)))$ 
13:  return ( $m, T$ ) ▷ 返回最大元素和更新的树

```

POP 上下处理两遍, 自顶向下一遍, 接着自底向上沿着“冠军之路”一遍。由于锦标赛树是平衡的, 路径的长度, 也就是树的高度为  $O(\lg n)$ 。时间复杂度为  $O(\lg n)$ 。下

面是锦标赛排序的实现。算法首先用  $O(n)$  时间构建一棵锦标赛树, 然后执行  $n$  次弹出操作, 逐一从树中取出最大值。每次弹出操作的性能为  $O(\lg n)$ , 锦标赛排序的总时间复杂度为  $O(n \lg n)$ 。

```

1: procedure SORT( $A$ )
2:    $T \leftarrow$  BUILD-TREE( $A$ )
3:   for  $i \leftarrow |A|$  down to 1 do
4:      $(A[i], T) \leftarrow$  POP( $T$ )

```

也可以递归实现锦标赛排序。复用二叉树的定义。令一棵非空的树为  $(l, k, r)$ , 其中  $k$  为元素,  $l, r$  是左右子树。定义  $wrap\ x = (\emptyset, x, \emptyset)$  构造一个叶子节点。这样就可以从列表  $xs$  构造出  $n$  棵叶子:  $ts = map\ wrap\ xs$ 。构造锦标赛树时, 比较两棵树  $t_1, t_2$ , 选择较大的作为根,  $t_1, t_2$  作为左右子树。

$$merge\ t_1\ t_2 = (t_1, \max\ k_1\ k_2, t_2) \quad (9.13)$$

其中  $k_1 = key\ t_1, k_2 = key\ t_2$ 。函数  $build\ ts$  不断取出两棵树进行合并, 最终构造出锦标赛树。

$$\begin{aligned} build\ [] &= \emptyset \\ build\ [t] &= t \\ build\ ts &= build\ (pairs\ ts) \end{aligned} \quad (9.14)$$

其中:

$$\begin{aligned} pairs\ (t_1:t_2:ts) &= (merge\ t_1\ t_2) : pairs\ ts \\ pairs\ ts &= ts \end{aligned} \quad (9.15)$$

为了取得冠军, 我们检查左右子树, 看哪一棵和根节点的元素相等。然后递归地从子树中取出冠军直到叶子节点。最后将叶子中的元素替换为  $-\infty$ 。

$$\begin{aligned} pop\ (\emptyset, k, \emptyset) &= (\emptyset, -\infty, \emptyset) \\ pop\ (l, k, r) &= \begin{cases} k = key\ l : (l', \max\ (key\ l')\ (key\ r), r), \text{ 其中 } l' = pop\ l \\ k = key\ r : (l, \max\ (key\ l)\ (key\ r'), r'), \text{ 其中 } r' = pop\ r \end{cases} \end{aligned} \quad (9.16)$$

排序的过程不断从一棵锦标赛树弹出冠军(降序):

$$\begin{aligned} sort\ \emptyset &= [] \\ sort\ (l, -\infty, r) &= [] \\ sort\ t &= (key\ t) : sort\ (pop\ t) \end{aligned} \quad (9.17)$$

## 练习 9.2

9.2.1. 将递归的锦标赛排序实现为升序。



9.2.2. 锦标赛树排序可以处理相等元素么? 它是稳定排序么?

9.2.3. 比较锦标赛树排序和二叉搜索树排序,它们的时间和空间效率如何。

9.2.4. 比较堆排序和锦标赛树排序,它们的时间和空间效率如何。

答案: 377 页

### 9.3.2 改进为堆排序

锦标赛树将时间复杂度提高到  $O(n \lg n)$ , 达到了基于比较的排序算法上限<sup>[51]</sup>。这里仍有改进的空间。排序完成后, 锦标赛树的所有节点都变成了负无穷, 这棵二叉树不再含有任何有效信息, 但却占据了空间。有没有办法在弹出后释放节点呢? 如果待排序的元素有  $n$  个, 锦标赛树实际上占用了  $2n$  个节点。其中有  $n$  个叶子和  $n$  个分支。有没有办法能节约一半空间呢? 如果当根节点为负无穷时, 认为树为空, 并将 *key* 重命名为 *top*, 那么式 (9.17) 就可以进一步转化为更通用的形式:

$$\begin{aligned} \text{sort } \emptyset &= [] \\ \text{sort } t &= (\text{top } t) : \text{sort } (\text{pop } t) \end{aligned} \quad (9.18)$$

这和堆排序定义完全一样。堆总在顶部保存最小(或最大)值, 并且提供了快速的弹出操作。使用数组的二叉堆实际上将树结构“编码”成数组索引, 因此除了  $n$  个单元外, 无需任何额外的空间。函数式的堆, 如左偏堆和伸展堆也只需要  $n$  个节点。我们将在下一章介绍更多种类的堆, 它们在许多情况下都有很好的性能。

## 9.4 附录:例子程序

尾递归实现的选择排序:

```

sort [] = []
sort xs = x : sort xs'
  where
    (x, xs') = extractMin xs

extractMin (x:xs) = min' [] x xs
  where
    min' ys m [] = (m, ys)
    min' ys m (x:xs) = if m < x then min' (x:ys) m xs
                      else min' (m:ys) x xs

```

鸡尾酒排序:

```

[A] cocktailSort([A] xs) {
  Int n = length(xs)
  for Int i = 0 to n / 2 {
    var (mi, ma) = (i, n - 1 - i)
    if xs[ma] < xs[mi] then swap(xs[mi], xs[ma])
    for Int j = i + 1 to n - 1 - i {
      if xs[j] < xs[mi] then mi = j
    }
  }
}

```

```

        if xs[ma] < xs[j] then ma = j
    }
    swap(xs[i], xs[mi])
    swap(xs[n - 1 - i], xs[ma])
}
return xs
}

```

尾递归的鸡尾酒排序:

```

csort xs = cocktail [] [] xs
where
    cocktail as bs [] = reverse as # bs
    cocktail as bs [x] = reverse (x:as) # bs
    cocktail as bs xs = let (mi, ma, xs') = minMax xs
                        in cocktail (mi:as) (ma:bs) xs'

minMax (x:y:xs) = foldr sel (min x y, max x y, []) xs
where
    sel x (mi, ma, ys) | x < mi = (x, ma, mi:ys)
                      | ma < x = (mi, x, ma:ys)
                      | otherwise = (mi, ma, x:ys)

```

复用二叉树构造锦标赛树:

```

Node<T> build([T] xs) {
    [T] ts = []
    for x in xs {
        append(ts, Node(null, x, null))
    }
    while length(ts) > 1 {
        [T] ts' = []
        for l, r in ts {
            append(ts', Node(l, max(l.key, r.key), r))
        }
        if odd(length(ts)) then append(ts', last(ts))
        ts = ts'
    }
    return ts[0];
}

```

从锦标赛树取出冠军:

```

T pop(Node<T> t) {
    T m = t.key
    t.key = -INF
    while not isLeaf(t) {
        t = if t.left.key == m then t.left else t.right
        t.key = -INF
    }
    while (t.parent ≠ null) {
        t = t.parent
        t.key = max(t.left.key, t.right.key)
    }
}

```

```

return (m, t);
}

```

锦标赛树排序:

```

void sort([A] xs) {
  Node<T> t = build(xs)
  for Int n = length(xs) - 1 downto 0 {
    (xs[n], t) = pop(t)
  }
}

```

递归的锦标赛排序(降序):

```

data Tr a = Empty | Br (Tr a) a (Tr a)

data Infinite a = NegInf | Only a | Inf deriving (Eq, Ord)

key (Br _ k _) = k

wrap x = Br Empty (Only x) Empty

merge t1@(Br _ k1 _) t2@(Br _ k2 _) = Br t1 (max k1 k2) t2

fromList = build ◦ (map wrap) where
  build [] = Empty
  build [t] = t
  build ts = build (pairs ts)
  pairs (t1:t2:ts) = (merge t1 t2) : pair ts
  pairs ts = ts

pop (Br Empty _ Empty) = Br Empty NegInf Empty
pop (Br l k r) | k == key l = let l' = pop l in Br l' (max (key l') (key r)) r
               | k == key r = let r' = pop r in Br l (max (key l) (key r')) r'

toList Empty = []
toList (Br _ Inf _) = []
toList t@(Br _ Only k _) = k : toList (pop t)

sort = toList ◦ fromList

```



# 第十章 二项式堆、斐波那契堆、配对堆

## 10.1 简介

二叉堆使用二叉树存储元素,将二叉树扩展成  $k$  叉树<sup>[54]</sup> ( $k > 2$ ) 甚至多棵树(称为森林)可得到更丰富的堆结构。本章介绍二项式堆,它由多棵  $k$  叉树的森林组成。如果延迟执行二项式堆的某些操作,可以得到斐波那契堆。它将堆合并的性能从对数时间复杂度提升到常数时间,这对于图算法很重要。本章还介绍配对堆。它实现简单并且性能良好。

## 10.2 二项式堆

二项式堆得名于牛顿二项式。它由一组  $k$  叉树的森林组成,每棵树的大小为二项式展开中的各项系数。牛顿证明了形如  $(a + b)^n$  的二项式展开后,各项系数可以表示为:

$$(a + b)^n = a^n + \binom{n}{1} a^{n-1} b + \dots + \binom{n}{n-1} a b^{n-1} + b \quad (10.1)$$

当  $n$  为自然数时,各项系数就呈现出帕斯卡三角形中的一行,如下图所示<sup>1[55]</sup>。

$$\begin{array}{c} 1 \\ 1 \quad 1 \\ 1 \quad 2 \quad 1 \\ 1 \quad 3 \quad 3 \quad 1 \\ 1 \quad 4 \quad 6 \quad 4 \quad 1 \\ \dots \end{array}$$

有多种方法可以产生一系列二项式系数,其中一种是使用递归。帕斯卡三角形中第一行为 1,任何一行的两端为 1,其它数字是上一行中左上和右上数字之和。

### 10.2.1 二项式树

二项式树是定义了秩(rank)的多叉树。秩是  $\geq 0$  的整数。秩为 0 的二项式树为  $B_0$ ,秩为  $n$  的二项式树为  $B_n$ 。

<sup>1</sup>中国称“贾宪”三角形。贾宪(1010-1070),牛顿在 1665 年证明了  $n$  为有理数时的情形,欧拉后来将  $n$  推广到实数。

1.  $B_0$  树只包含一个节点;
2.  $B_n$  树由两棵  $B_{n-1}$  树组成, 其中根节点元素较大的一棵是另一棵的最左侧子树。  
如图 10.1 所示。

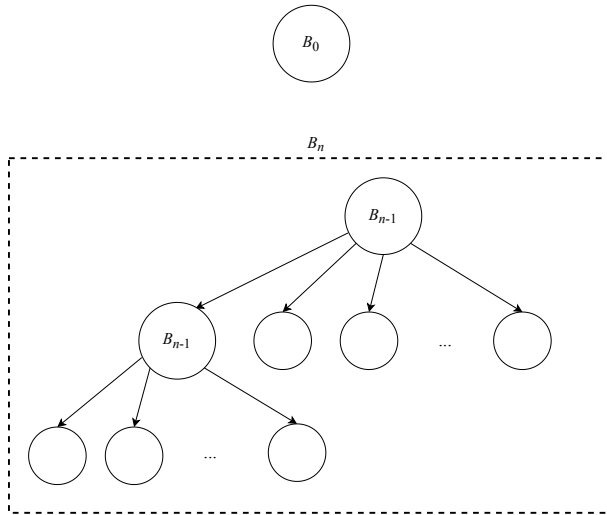


图 10.1: 二项式树

图 10.2 给出了  $B_0$ 、 $B_1$ 、 $B_2$ 、 $B_3$ 、 $B_4$  的例子。观察它们, 可以发现  $B_n$  中每行的节点数目恰好是二项式系数。例如  $B_4$  第 0 层有 1 个节点(根节点), 第 1 层有 4 个节点, 第 2 层有 6 个节点, 第 3 层有 4 个节点, 第 4 层有 1 个节点。它们恰好是帕斯卡三角形的第 4 行(从第 0 行开始): 1、4、6、4、1。这就是二项式树名字的由来。进一步我们可以得知二项式树  $B_n$  中含有  $2^n$  个元素。

二项式堆包含一组二项式树(森林), 并满足:

1. 每棵树都满足**堆性质**, 对于小顶堆, 任意节点元素都  $\geq$  父节点元素;
2. 堆中任何两棵二项式树的秩都不同。

从性质 2 可知: 含有  $n$  个元素的二项式堆, 如果将  $n$  转换为二进制数  $(a_m \dots a_1 a_0)_2$ , 其中  $a_0$  是最低位(LSB),  $a_m$  是最高位(MSB), 当且仅当  $a_i = 1$  时, 堆中存在秩为  $i$  的二项式树。例如, 若二项式堆有 5 个元素, 5 的二进制为  $(101)_2$ , 堆中含有两棵二项式树:  $B_0$  和  $B_2$ 。图 10.3 中的二项式堆含有 19 个元素,  $19 = (10011)_2$ , 它包含三棵树:  $B_0$ 、 $B_1$ 、 $B_4$ 。

定义二项式树为多叉树  $(r, k, ts)$ , 包含根节点元素  $k$ 、秩  $r$  和若干子树  $ts$ 。定义二项式堆为按秩递增的二项式树列表:

```
data BiTree a = Node Int a [BiTree a]
type BiHeap a = [BiTree a]
```

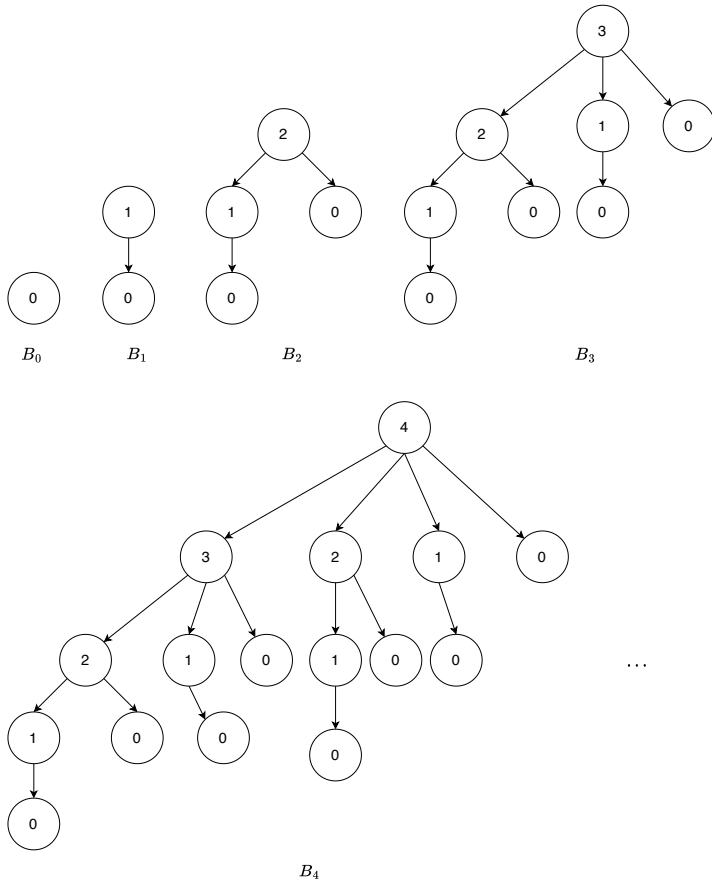


图 10.2: 秩为 0、1、2、3、4……的二项式树

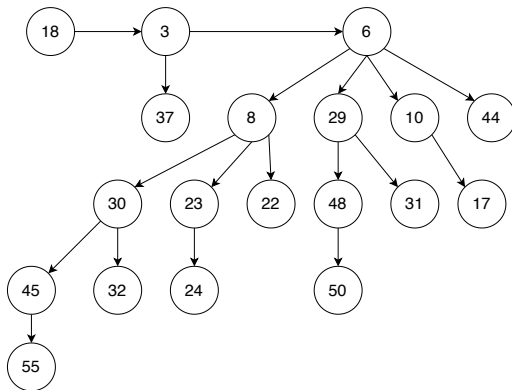


图 10.3: 含有 19 个元素的二项式堆

有一种叫做“左侧孩子，右侧兄弟”<sup>[4]</sup>的方法，可以复用二叉树的结构来定义多叉树。每个节点包含左侧和右侧部分：左侧部分指向节点的第一棵子树，右侧部分指向兄弟节点。所有兄弟节点组成一个链表，如图 10.4 所示。也可以直接利用数组或列表来表示一个节点的子树。

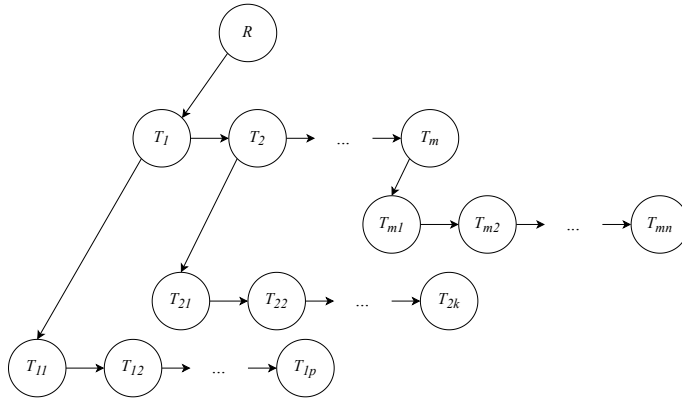


图 10.4:  $R$  为根节点,  $T_1, T_2, \dots, T_m$  为  $R$  的子树。  $R$  的左侧为  $T_1$ , 右侧为空。  $T_{11}, \dots, T_{1p}$  为  $T_1$  的子树。  $T_1$  的左侧是子树  $T_{11}$ , 右侧是兄弟节点  $T_2$ 。  $T_2$  的左侧是子树  $T_{21}$ , 右侧是兄弟节点。

### 10.2.2 树的链接

链接两棵树  $B_n$  树可以构造出  $B_{n+1}$ 。比较两棵树的根节点元素,选择较小的作为新的根,然后将另一棵置于其它子树前面,如图 10.5 所示。

$$link(r, x, ts) (r, y, ts') = \begin{cases} x < y : (r + 1, x, (r, t, ts'):ts) \\ \text{否则} : (r + 1, y, (r, x, ts):ts') \end{cases} \quad (10.2)$$

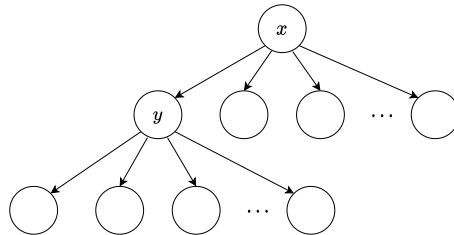


图 10.5: 如果  $x < y$ , 将  $y$  作为  $x$  的第一棵子树。

使用“左侧孩子，右侧兄弟”的实现如下。链接操作可以在常数时间内完成。

- 1: **function** LINK( $x, y$ )
- 2:     **if** KEY( $y$ ) < KEY( $x$ ) **then**
- 3:         Exchange  $x \leftrightarrow y$



```

4:  SIBLING( $y$ )  $\leftarrow$  SUB-TREES( $T_1$ )
5:  SUB-TREES( $x$ )  $\leftarrow y$ 
6:  PARENT( $y$ )  $\leftarrow x$ 
7:  RANK( $x$ )  $\leftarrow$  RANK( $y$ ) + 1
8:  return  $x$ 

```

### 练习 10.1

10.1.1. 编程产生帕斯卡三角形

10.1.2. 证明二项式树  $B_n$  中第  $i$  行的节点数为  $\binom{n}{i}$ 。

10.1.3. 证明二项式树  $B_n$  中含有  $2^n$  个节点。

10.1.4. 用容器保存子树,实现二项式树的链接。这种方式有何问题,怎样解决?

答案: 378 页

### 10.2.3 插入

二项式堆中的树按秩递增排列,需要在插入新树时保持有序:

$$\begin{aligned}
 \text{ins } t \ [ ] &= [t] \\
 \text{ins } t \ (t':ts) &= \begin{cases} \text{rank } t < \text{rank } t' : t:t':ts \\ \text{rank } t' < \text{rank } t : t' : \text{ins } t \ ts \\ \text{否则} : & \text{ins } (\text{link } t \ t') \ ts \end{cases} \quad (10.3)
 \end{aligned}$$

其中  $\text{rank}(r, k, ts) = r$ , 获取二项式树的秩。如果堆为空  $[ ]$ , 结果为  $[t]$ ; 否则比较  $t$  和堆中第一棵树  $t'$  的秩。如果  $t$  的秩较小, 则它成为第一棵树; 如果  $t'$  的秩较小, 递归地将  $t$  插入到剩余的树中; 如果秩相等, 就将  $t, t'$  链接成一棵更大的树, 然后递归插入到其余的树中。如果有  $n$  个元素, 堆中最多有  $O(\lg n)$  棵二项式树。 $\text{ins}$  最多执行  $O(\lg n)$  次常数时间的链接。时间复杂度为  $O(\lg n)^2$ 。使用  $\text{ins}$ , 我们可以定义二项式堆的插入算法。将待插入元素  $x$  放入一个叶子节点中, 然后再插入到堆中:

$$\text{insert } x = \text{ins } (0, x, [ ]) \quad (10.4)$$

这一定义是柯里化的。我们可以利用叠加操作将若干元素插入到堆中:

$$\text{fromList} = \text{foldr } \text{ins } [ ] \quad (10.5)$$

对应的“左侧孩子, 右侧兄弟”实现如下:

```

1: function INSERT-TREE( $T, H$ )
2:    $\perp \leftarrow p \leftarrow \text{NODE}(0, \text{NIL}, \text{NIL})$ 

```

<sup>2</sup>这一过程和两个二进制数的加法相似, 可以引出一类问题: 数值表示 (numeric representation) [3]。

```

3:  while  $H \neq \text{NIL}$  且  $\text{RANK}(H) \leq \text{RANK}(T)$  do
4:     $T_1 \leftarrow H$ 
5:     $H \leftarrow \text{SIBLING}(H)$ 
6:    if  $\text{RANK}(T) = \text{RANK}(T_1)$  then
7:       $T \leftarrow \text{LINK}(T, T_1)$ 
8:    else
9:       $\text{SIBLING}(p) \leftarrow T_1$ 
10:      $p \leftarrow T_1$ 
11:     $\text{SIBLING}(p) \leftarrow T$ 
12:     $\text{SIBLING}(T) \leftarrow H$ 
13:  return  $\text{REMOVE-FIRST}(\perp)$ 

14: function  $\text{REMOVE-FIRST}(H)$ 
15:    $n \leftarrow \text{SIBLING}(H)$ 
16:    $\text{SIBLING}(H) \leftarrow \text{NIL}$ 
17:   return  $n$ 

```

### 10.2.4 堆合并

合并两个二项式堆相当于合并两个二项式树森林。合并结果中没有秩相同的树，并且按照秩递增。合并过程和归并排序类似。每次从两个堆中各取出第一棵树，比较它们的秩，将较小的一棵放入结果中。如果两棵树的秩相等，我们将它们链接成为一棵较大的树，然后递归插入到合并结果中。

$$\begin{aligned}
 \text{merge } ts_1 \ [ ] &= ts_1 \\
 \text{merge } [ ] \ ts_2 &= ts_2 \\
 \text{merge } (t_1:ts_1) \ (t_2:ts_2) &= \begin{cases} \text{rank } t_1 < \text{rank } t_2 : & t_1 : (\text{merge } ts_1 \ (t_2:ts_2)) \\ \text{rank } t_2 < \text{rank } t_1 : & t_2 : (\text{merge } (t_1:ts_1) \ ts_2) \\ \text{否则} : & \text{ins } (\text{link } t_1 \ t_2) \ (\text{merge } ts_1 \ ts_2) \end{cases} \\
 & \hspace{15em} (10.6)
 \end{aligned}$$

当  $t_1, t_2$  秩相同时，我们也可以将链接后的树插入回任意一个堆，然后递归合并：

$$\text{merge } (\text{ins } (\text{link } t_1 \ t_2) \ ts_1) \ ts_2$$

用这种方式可以消除递归，实现迭代式的合并：

```

1: function  $\text{MERGE}(H_1, H_2)$ 
2:    $H \leftarrow p \leftarrow \text{NODE}(0, \text{NIL}, \text{NIL})$ 
3:   while  $H_1 \neq \text{NIL}$  且  $H_2 \neq \text{NIL}$  do
4:     if  $\text{RANK}(H_1) < \text{RANK}(H_2)$  then
5:        $\text{SIBLING}(p) \leftarrow H_1$ 

```

```

6:         p ← SIBLING(p)
7:         H1 ← SIBLING(H1)
8:     else if RANK(H2) < RANK(H1) then
9:         SIBLING(p) ← H2
10:        p ← SIBLING(p)
11:        H2 ← SIBLING(H2)
12:    else ▷ 秩相等
13:        T1 ← H1, T2 ← H2
14:        H1 ← SIBLING(H1), H2 ← SIBLING(H2)
15:        H1 ← INSERT-TREE(LINK(T1, T2), H1)
16:    if H1 ≠ NIL then
17:        SIBLING(p) ← H1
18:    if H2 ≠ NIL then
19:        SIBLING(p) ← H2
20:    return REMOVE-FIRST(H)

```

若堆  $H_1$  中有  $m_1$  棵树,  $H_2$  中有  $m_2$  棵树。合并后的结果中最多有  $m_1 + m_2$  棵树。如果没有秩相同的树, 则合并时间为  $O(m_1 + m_2)$ 。如果存在秩相同的树, 最多需要调用  $O(m_1 + m_2)$  次 *ins*。考虑  $m_1 = 1 + \lfloor \lg n_1 \rfloor$ ,  $m_2 = 1 + \lfloor \lg n_2 \rfloor$ , 其中  $n_1$  和  $n_2$  是两个堆各自的大小, 且  $\lfloor \lg n_1 \rfloor + \lfloor \lg n_2 \rfloor \leq 2\lfloor \lg n \rfloor$ , 其中  $n = n_1 + n_2$ 。最终合并的复杂度为  $O(\lg n)$ 。

### 10.2.5 弹出

二项式堆中每棵树的根节点保存了各自的最小值, 但这些最小值间的关系是不确定的。为了获取堆的最小元素(堆顶), 需要在全部根节点元素中查找。堆中有  $O(\lg n)$  棵树, 因此获取堆顶的复杂度为  $O(\lg n)$ 。

$$\text{top}(t:ts) = \text{foldr } f \text{ (key } t) \text{ } ts \quad \text{其中 } f(r, x, ts) \text{ } y = \min x \text{ } y \quad (10.7)$$

对应的迭代实现:

```

1: function TOP(H)
2:     m ← ∞
3:     while H ≠ NIL do
4:         m ← MIN(m, KEY(H))
5:         H ← SIBLING(H)
6:     return m

```

弹出堆顶后还需要保持堆性质。令堆中的二项式树为  $B_i, B_j, \dots, B_p, \dots, B_m$ , 其中  $B_p$  的根节点为堆中最小元素。将其删除后产生  $p$  棵二项式子树, 秩为  $p-1, p-2, \dots, 0$ 。将这  $p$  棵子树逆序, 形成一个新二项式堆  $H_p$ 。除去  $B_p$  的树也构成一个二项式堆

$H' = H - [B_p]$ 。将  $H_p$  和  $H'$  合并就可以得到最终结果,如图 10.6 所示。为了实现弹出,我们需要从堆中分离出堆顶所在的树:

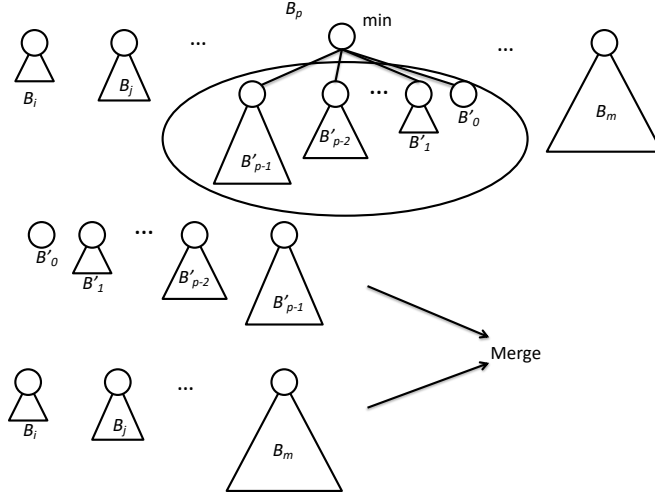


图 10.6: 二项式堆的弹出操作

$$\begin{aligned} \text{min}' [t] &= (t, []) \\ \text{min}' (t:ts) &= \begin{cases} \text{key } t < \text{key } t' : (t, ts), \text{ 其中 } : (t', ts') = \text{min}' ts \\ \text{否则} : (t', t:ts') \end{cases} \quad (10.8) \end{aligned}$$

其中  $\text{key}(r, k, ts) = k$  获取根节点元素。 $\text{min}'$  的结果为一对值: 堆顶所在的树, 和其它树。接下来定义弹出:

$$\text{pop } H = \text{merge}(\text{reverse } ts) H' \quad \text{其中 } ((r, k, ts), H') = \text{min}' H \quad (10.9)$$

对应的迭代实现为:

- 1: **function** POP( $H$ )
- 2:      $(T_m, H) \leftarrow \text{EXTRACT-MIN}(H)$
- 3:      $H \leftarrow \text{MERGE}(H, \text{REVERSE}(\text{SUB-TREES}(T_m)))$
- 4:     SUB-TREES( $T_m$ )
- 5:     **return** (KEY( $T_m$ ),  $H$ )

其中反转操作的实现见第一章, EXTRACT-MIN 的迭代实现如下:

- 1: **function** EXTRACT-MIN( $H$ )
- 2:      $H' \leftarrow H, p \leftarrow \text{NIL}$
- 3:      $T_m \leftarrow T_p \leftarrow \text{NIL}$
- 4:     **while**  $H \neq \text{NIL}$  **do**

```

5:      if  $T_m = \text{NIL}$  或  $\text{KEY}(H) < \text{KEY}(T_m)$  then
6:           $T_m \leftarrow H$ 
7:           $T_p \leftarrow p$ 
8:           $p \leftarrow H$ 
9:           $H \leftarrow \text{SIBLING}(H)$ 
10:     if  $T_p \neq \text{NIL}$  then
11:          $\text{SIBLING}(T_p) \leftarrow \text{SIBLING}(T_m)$ 
12:     else
13:          $H' \leftarrow \text{SIBLING}(T_m)$ 
14:      $\text{SIBLING}(T_m) \leftarrow \text{NIL}$ 
15:     return  $(T_m, H')$ 

```

弹出可以实现堆排序。首先从待排序元素构建一个二项式堆,然后不断弹出最小元素。

$$\text{sort} = \text{heapSort} \circ \text{fromList} \quad (10.10)$$

其中:

$$\begin{aligned} \text{heapSort } [] &= [] \\ \text{heapSort } H &= (\text{top } H) : \text{heapSort } (\text{pop } H) \end{aligned} \quad (10.11)$$

二项式堆的插入、合并的时间复杂度在最坏情况下是  $O(\lg n)$ ,分摊复杂度为常数时间,我们这里略去了证明。

## 10.3 斐波那契堆

二项式堆的名字来自二项式定理,斐波那契堆的名字来自斐波那契数列<sup>3</sup>。斐波那契堆本质上是惰性二项式堆。但这并不意味着在惰性求值环境下二项式堆会自动成为斐波那契堆。惰性环境仅提供了实现上的便利<sup>[56]</sup>。除弹出外,斐波那契堆所有操作的分摊复杂度都是常数时间<sup>[57]</sup>。

向二项式堆插入元素  $x$  时,我们将  $x$  放入叶子节点,再插入到森林中。树按照秩递增的顺序插入,如果秩相等,则进行链接,然后递归插入。时间复杂度为  $O(\lg n)$ 。使用惰性策略,我们将按秩插入和链接推迟进行。将叶子  $x$  直接加入森林中。为了快速获取堆顶,我们需要记录哪棵树的根节点保存了最小元素。一个斐波那契堆或者为空  $\emptyset$ ,或者是若干树的森林,记为  $(n, t_m, ts)$ 。其中  $n$  是元素个数,  $t_m$  是堆顶所在的树,  $ts$  是其余二项式树的列表。下面的例子程序定义了斐波那契堆(复用了二项式树的定义):

<sup>3</sup>Michael L. Fredman 和 Robert E. Tarjan 在证明这种堆的时间复杂度时使用了斐波那契数列的性质,于是他们命名为“斐波那契堆”<sup>[4]</sup>。

操作	二项式堆	斐波那契堆
插入	$O(\lg n)$	$O(1)$
合并	$O(\lg n)$	$O(1)$
获取堆顶	$O(\lg n)$	$O(1)$
弹出	$O(\lg n)$	分摊 $O(\lg n)$

表 10.1: 斐波那契堆和二项式堆的(分摊)复杂度对比

```

data FibHeap a = Empty | FibHeap { size :: Int
    , minTree :: BiTree a
    , trees :: [BiTree a]}

```

这样就可以用常数时间获取堆顶(柯里化):  $top = key \circ minTree$ 。

### 10.3.1 插入

定义插入为一种特殊的合并, 其中一个堆仅含有一个叶节点:  $insert\ x\ H = merge\ (singleton\ x)\ H$ 。或写成柯里化形式:

$$insert\ x = merge \circ (singleton\ x) \quad (10.12)$$

其中:

$$singleton\ x = (1, (1, x, []), [])$$

下面是对应的命令式实现:

- 1: **function** INSERT( $k, H$ )
- 2:      $x \leftarrow SINGLETON(k)$  ▷ 将  $k$  装入一棵树
- 3:     ADD( $x, TREES(H)$ )
- 4:      $T_m \leftarrow MIN-TREE(H)$
- 5:     **if**  $T_m = NIL$  或  $k < KEY(T_m)$  **then**
- 6:         MIN-TREE( $H$ )  $\leftarrow x$
- 7:     SIZE( $H$ )  $\leftarrow SIZE(H) + 1$

其中 TREES( $H$ ) 获取堆  $H$  的森林(所有树的列表), MIN-TREE( $H$ ) 记录了  $H$  堆顶所在的树。

### 10.3.2 合并

合并时推迟链接, 仅将两个堆中的树放到一起, 然后找出新的堆顶。

$$\begin{aligned}
\text{merge } h \ \emptyset &= h \\
\text{merge } \emptyset \ h &= h \\
\text{merge } (n, t_m, ts) \ (n', t'_m, ts') &= \begin{cases} \text{key } t_m < \text{key } t'_m : & (n + n', t_m, t'_m : ts \# ts') \\ \text{否则} : & (n + n', t'_m, t_m : ts \# ts') \end{cases} \\
&\hspace{15em} (10.13)
\end{aligned}$$

当两个堆都不为空时,  $\#$  的耗时和其中一个堆中树的棵数成正比。使用双向链表可以把合并提高到常数时间。

```

data Node<K> {
    K key
    Int rank
    Node<K> next, prev, parent, subTrees
}

data FibHeap<K> {
    Int size
    Node<K> minTree, trees
}

```

```

1: function MERGE( $H_1, H_2$ )
2:    $H \leftarrow$  FIB-HEAP
3:   TREES( $H$ )  $\leftarrow$  CONCAT(TREES( $H_1$ ), TREES( $H_2$ ))
4:   if KEY(MIN-TREE( $H_1$ )) < KEY(MIN-TREE( $H_2$ )) then
5:     MIN-TREE( $H$ )  $\leftarrow$  MIN-TREE( $H_1$ )
6:   else
7:     MIN-TREE( $H$ )  $\leftarrow$  MIN-TREE( $H_2$ )
8:     SIZE( $H$ ) = SIZE( $H_1$ ) + SIZE( $H_2$ )
9:   return  $H$ 

```

```

9: function CONCAT( $s_1, s_2$ )
10:   $e_1 \leftarrow$  PREV( $s_1$ )
11:   $e_2 \leftarrow$  PREV( $s_2$ )
12:  NEXT( $e_1$ )  $\leftarrow$   $s_2$ 
13:  PREV( $s_2$ )  $\leftarrow$   $e_1$ 
14:  NEXT( $e_2$ )  $\leftarrow$   $s_1$ 
15:  PREV( $s_1$ )  $\leftarrow$   $e_2$ 
16:  return  $s_1$ 

```

### 10.3.3 弹出

合并时推迟了树的链接, 接下来需要在弹出时将其“补偿”回来。我们定义这一过程为树的归并。考虑另一个问题: 给定若干 2 的整数次幂, 如:  $L = [2, 1, 1, 4, 8, 1, 1, 2, 4]$ ,

不断将相等的两个数相加,直到没有任何相等的数。这个例子的最终结果为 [8, 16]。表 10.2 给出了归并步骤。第一列表示每次“扫描”到的数字;第二列是中间结果。被扫描的数字和结果列表中的第一个元素相比较。如果相等,就括起来;最后一列是归并结果,每个结果都用于下一步的处理。归并过程可以用叠加实现:

数字	比较、相加	结果
2	2	2
1	1, 2	1, 2
1	(1+1), 2	4
4	(4+4)	8
8	(8+8)	16
1	1, 16	1, 16
1	(1+1), 16	2, 16
2	(2+2), 16	4, 16
4	(4+4), 16	8, 16

表 10.2: 归并数字的步骤

$$\text{consolidate} = \text{foldr melt []} \quad (10.14)$$

其中:

$$\begin{aligned} \text{melt } x [] &= x \\ \text{melt } x (x':xs) &= \begin{cases} x = x' : \text{melt } 2x xs \\ x < x' : x:x':xs \\ x > x' : x' : \text{melt } x xs \end{cases} \end{aligned} \quad (10.15)$$

令  $n = \text{sum } L$ , 为所有数之和,  $\text{consolidate}$  相当于把  $n$  表示为二进制数, 当且仅当第  $i$  位上的数字是 1, 结果列表中包含  $2^i$  这个数 ( $i$  从 0 开始)。例如  $\text{sum } [2, 1, 1, 4, 8, 1, 1, 2, 4] = 24$ , 写成二进制是  $(11000)_2$ , 第 3 和第 4 位上是 1, 所以最终列表中包含  $2^3 = 8, 2^4 = 16$ 。用类似的方法可以实现树的归并。比较秩并把秩相同的树链接起来:

$$\begin{aligned} \text{melt } t [] &= [t] \\ \text{melt } t (t':ts) &= \begin{cases} \text{rank } t = \text{rank } t' : \text{melt } (\text{link } t t') ts \\ \text{rank } t < \text{rank } t' : t:t':ts \\ \text{rank } t > \text{rank } t' : t' : \text{melt } t ts \end{cases} \end{aligned} \quad (10.16)$$

图 10.7 给出了树归并的步骤, 和表 10.2 对比可以看出相似性。也可以使用辅助数组  $A$  进行归并。 $A[i]$  保存秩为  $i$  的树。遍历时, 如果遇到另一棵秩为  $i$  的树, 就将它们链接起来得到一棵秩为  $i+1$  的树。然后将  $A[i]$  清空, 接着检查  $A[i+1]$ , 若不为空就再次进行链接。遍历后,  $A$  中就保存了归并结果。



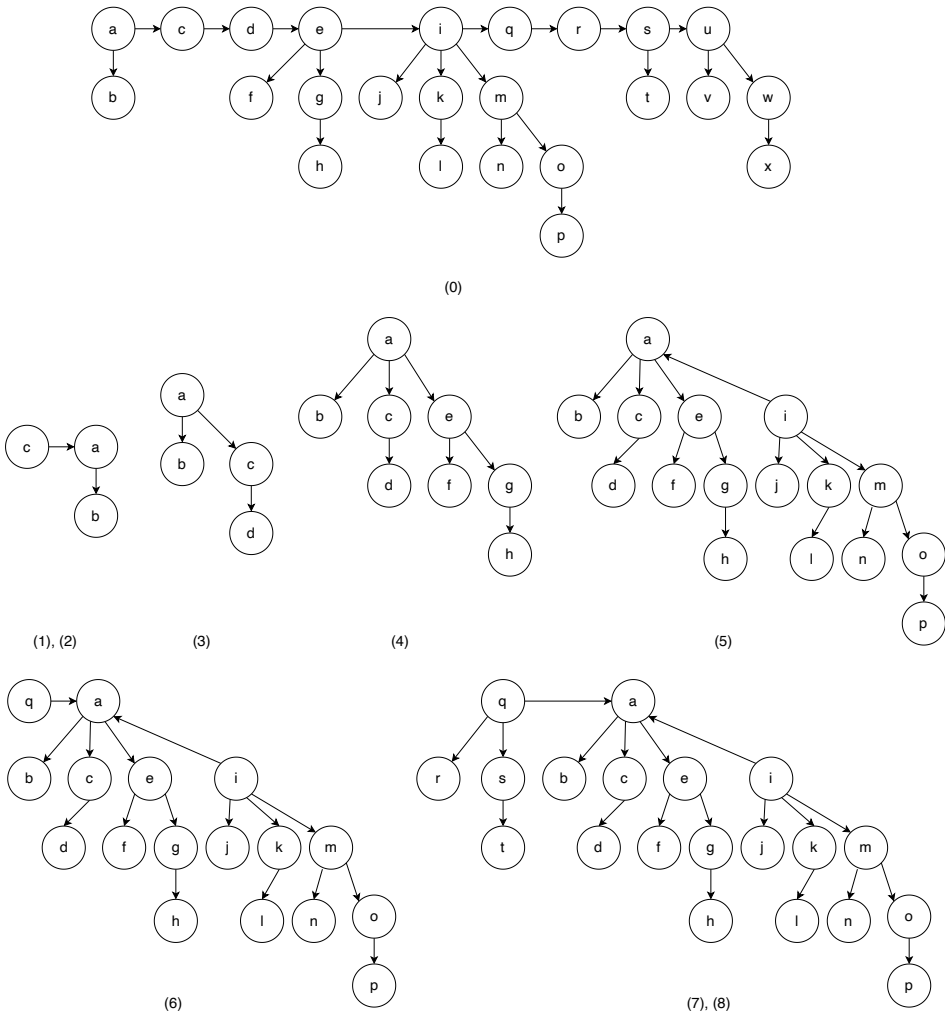


图 10.7: 树归并的步骤。第 3 步, d 先链接到 c, 然后链接到 a; 第 7、8 步, r 先链接到 q, 然后 s 链接到 q。

```

1: function CONSOLIDATE( $H$ )
2:    $R \leftarrow \text{MAX-RANK}(\text{SIZE}(H))$ 
3:    $A \leftarrow [\text{NIL}, \text{NIL}, \dots, \text{NIL}]$  ▷ 共  $R$  个
4:   for each  $T$  in TREES( $H$ ) do
5:      $r \leftarrow \text{RANK}(T)$ 
6:     while  $A[r] \neq \text{NIL}$  do
7:        $T' \leftarrow A[r]$ 
8:        $T \leftarrow \text{LINK}(T, T')$ 
9:        $A[r] \leftarrow \text{NIL}$ 
10:       $r \leftarrow r + 1$ 
11:      $A[r] \leftarrow T$ 
12:    $T_m \leftarrow \text{NIL}$ 
13:   TREES( $H$ )  $\leftarrow \text{NIL}$ 
14:   for each  $T$  in  $A$  do
15:     if  $T \neq \text{NIL}$  then
16:       append  $T$  to TREES( $H$ )
17:       if  $T_m = \text{NIL}$  or  $\text{KEY}(T) < \text{KEY}(T_m)$  then
18:          $T_m \leftarrow T$ 
19:   MIN-TREE( $H$ )  $\leftarrow T_m$ 

```

将所有二项式树归并后,斐波那契堆就变成了二项式堆,共有  $O(\lg n)$  棵树。若堆的大小为  $n$ ,  $\text{MAX-RANK}(n)$  返回最大可能的秩  $R$ 。由二项式堆的结论,秩最大的树  $B_R$  有  $2^R$  个元素。有  $2^R \leq n < 2^{R+1}$ 。据此推测它的上限为  $R \leq \log_2 n$ 。我们稍后给出  $R$  的更准确估计。接下来需要再扫描一遍树找出堆顶。可以复用式 (10.8) 定义的  $\text{min}'$  分离出堆顶所在的树。

$$\begin{aligned}
 \text{pop}(1, (0, x, [], []), []) &= (x, []) \\
 \text{pop}(n, (r, x, ts_m, ts), ts) &= (x, (n-1, t_m, ts'))
 \end{aligned} \tag{10.17}$$

其中  $(t_m, ts') = \text{min}'(\text{consolidate}(ts_m \uplus ts))$ 。注意到  $\uplus$  的复杂度是  $O(|ts_m|)$ , 和堆顶所在的树中的子树棵数成正比。对应的命令式实现如下:

```

1: function POP( $H$ )
2:    $T_m \leftarrow \text{MIN-TREE}(H)$ 
3:   for each  $T$  in SUB-TREES( $T_m$ ) do
4:     append  $T$  to TREES( $H$ )
5:     PARENT( $T$ )  $\leftarrow \text{NIL}$ 
6:   remove  $T_m$  from TREES( $H$ )
7:   SIZE( $H$ )  $\leftarrow \text{SIZE}(H) - 1$ 
8:   CONSOLIDATE( $H$ )
9:   return ( $\text{KEY}(T_m), H$ )

```

我们使用“势能方法”分析弹出的分摊性能。回忆物理学中重力势能的定义：

$$E = mgh \quad (10.18)$$

如图 10.8 所示,假设一个复杂的操作过程,将质量为  $m$  的物体上下移动,最终停在高为  $h'$  的位置。这一过程中的摩擦阻力做功  $W_f$ ,做功总和为:

$$W = mg(h' - h) + W_f$$

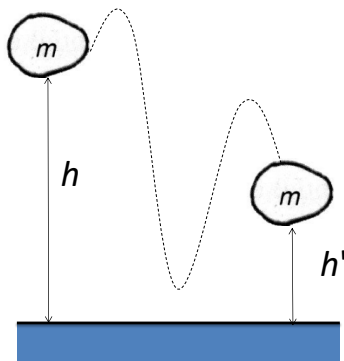


图 10.8: 重力势能

为了计算斐波那契堆弹出的总消耗,令弹出前的势能为  $\Phi(H)$ 。这个势能是由迄今为止的插入、合并操作累积的。经过树归并,形成了新堆  $H'$ ,由此计算新势能  $\Phi(H')$ 。 $\Phi(H')$  和  $\Phi(H)$  的差再加上树归并消耗的部分就是弹出的分摊复杂度。定义势能为:

$$\Phi(H) = t(H) \quad (10.19)$$

其中  $t(H)$  是堆中树的棵数。对于  $n$  个节点的斐波那契堆,令秩的上限为  $R(n)$ 。归并后的棵数最多为  $t(H') = R(n) + 1$ 。在归并前,我们还做了另一个重要操作:将堆顶所在的树根删除,将其子树添加到森林中。因此最多归并  $R(n) + t(H) - 1$  棵树。设弹出操作的时间复杂度为  $T$ ,树归并的时间复杂度为  $T_c$ ,推导分摊复杂度如下:

$$\begin{aligned} T &= T_c + \Phi(H') - \Phi(H) \\ &= O(R(n) + t(H) - 1) + (R(n) + 1) - t(H) \\ &= O(R(n)) \end{aligned} \quad (10.20)$$

插入、合并、弹出操作确保斐波那契堆中的所有树都为二项式树,此时  $R(n)$  的上限为  $O(\lg n)$ 。

### 10.3.4 提升优先级

用堆管理若干任务。若某个任务需要提前执行,我们将其优先级的值减小,使得任务更接近堆顶。某些图算法,例如最小生成树算法和 Dijkstra 算法都要求这一操作<sup>[4]</sup>的分摊复杂度达到常数时间。令  $x$  为堆  $H$  中的某个节点,需要将其的值减小为  $k$ 。如图 10.9,若  $x$  的值小于父节点  $y$  的值,将  $x$  切下并添加到堆(森林)中。尽管这样使得每棵树的父节点仍存有最小值,但切除节点后的树不再是二项式树。如果失去了很多子树,就无法保证合并的性能。为了解决这一问题,需要增加一个限制条件:

当节点第二次失去了某个子节点时,立即将它切下并添加到堆(森林)中。

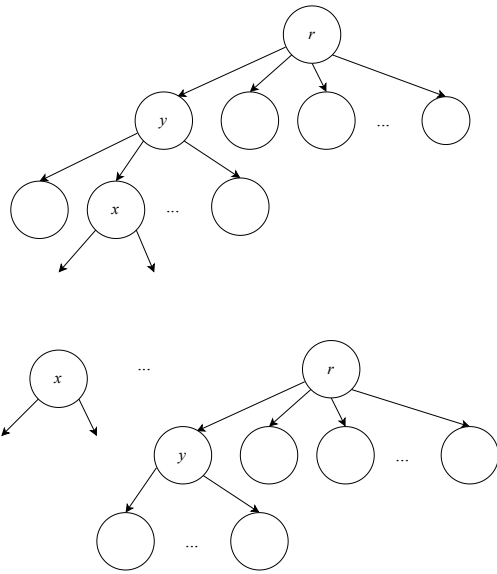


图 10.9: 若  $key\ x < key\ y$ , 将  $x$  切下, 然后添加到堆中。

```

1: function DECREASE( $H, x, k$ )
2:   KEY( $x$ )  $\leftarrow k$ 
3:    $p \leftarrow$  PARENT( $x$ )
4:   if  $p \neq$  NIL and  $k <$  KEY( $p$ ) then
5:     CUT( $H, x$ )
6:     CASCADE-CUT( $H, p$ )
7:   if  $k <$  TOP( $H$ ) then
8:     MIN-TREE( $H$ )  $\leftarrow x$ 

```

▷ 回溯切除

CASCADE-CUT 使用一个标记来记录它此前是否失去过子节点。并在 CUT 中清除标记。

```

1: function CUT( $H, x$ )
2:    $p \leftarrow$  PARENT( $x$ )

```

```

3:   remove  $x$  from  $p$ 
4:    $\text{RANK}(p) \leftarrow \text{RANK}(p) - 1$ 
5:   add  $x$  to  $\text{TREES}(H)$ 
6:    $\text{PARENT}(x) \leftarrow \text{NIL}$ 
7:    $\text{MARK}(x) \leftarrow \text{False}$ 

```

回溯时若节点  $x$  被标记了,说明它此前已经失去了子节点。我们需要继续回溯切除,直到根节点。

```

1: function CASCADE-CUT( $H, x$ )
2:    $p \leftarrow \text{PARENT}(x)$ 
3:   if  $p \neq \text{NIL}$  then
4:     if  $\text{MARK}(x) = \text{False}$  then
5:        $\text{MARK}(x) \leftarrow \text{True}$ 
6:     else
7:       CUT( $H, x$ )
8:       CASCADE-CUT( $H, p$ )

```

## 练习 10.2

试说明 DECREASE 的分摊复杂度为常数时间  $O(1)$ 。

答案: 379 页

### 10.3.5 斐波那契堆的命名

我们尚未给出 MAX-RANK( $n$ ) 的实现。它定义了  $n$  个元素的斐波那契堆中秩的上限。

**引理 10.3.1.** 对于斐波那契堆中的任何树  $x$ , 若其秩为  $k$ , 即  $k = \text{rank}(x)$ ,  $|x|$  为树中元素个数, 则:

$$|x| \geq F_{k+2} \quad (10.21)$$

其中  $F_k$  为斐波那契数列中的第  $k$  项:

$$\begin{aligned} F_0 &= 0 \\ F_1 &= 1 \\ F_k &= F_{k-1} + F_{k-2} \end{aligned}$$

**证明.** 考虑节点  $x$  的全部  $k$  棵子树:  $y_1, y_2, \dots, y_k$ 。顺序是被链接到  $x$  的时间先后。其中  $y_1$  最早加入,  $y_k$  最新加入。显然有  $|y_i| \geq 0$ 。当  $y_i$  链接到  $x$  的时候, 子树  $y_1, y_2, \dots, y_{i-1}$  已经存在了。因为我们只把秩相同的树链接起来, 所以在这一时刻, 有:

$$\text{rank}(y_i) = \text{rank}(x) = i - 1$$

此后  $y_i$  最多只能失去一个子节点(通过 DECREASE),一旦失去第二个子节点,它会被立即切除并加入到森林中。因此可以推断,对任何  $i = 2, 3, \dots, k$ , 有:

$$\text{rank}(y_i) \geq i - 2$$

令  $s_k$  为  $x$  的子节点个数可能的最小值,其中  $k = \text{rank}(x)$ 。对于边界情况,有  $s_0 = 1, s_1 = 2$ 。也就是说:秩为 0 的树,最少有 1 个节点,秩为 1 的树最少有 2 个节点,秩为  $k$  的树,最少有  $s_k$  个节点:

$$\begin{aligned} |x| &\geq s_k \\ &= 2 + s_{\text{rank}(y_2)} + s_{\text{rank}(y_3)} + \dots + s_{\text{rank}(y_k)} \\ &\geq 2 + s_0 + s_1 + \dots + s_{k-2} \end{aligned}$$

其中最后一行成立是因为  $\text{rank}(y_i) \geq i - 2$  并且  $s_k$  单调增,所以  $s_{\text{rank}(y_i)} \geq s_{i-2}$ 。接下来要证明  $s_k > F_{k+2}$ 。使用数学归纳法。对于边界情况,有  $s_0 = 1 \geq F_2 = 1$ , 以及  $s_1 = 2 \geq F_3 = 2$ 。对于  $k \geq 2$  的情况,有:

$$\begin{aligned} |x| &\geq s_k \\ &\geq 2 + s_0 + s_1 + \dots + s_{k-2} \\ &\geq 2 + F_2 + F_3 + \dots + F_k && \text{归纳假设} \\ &= 1 + F_0 + F_1 + F_2 + \dots + F_k && \text{利用 } F_0 = 0, F_1 = 1 \end{aligned}$$

接下来需要证明

$$F_{k+2} = 1 + \sum_{i=0}^k F_i \tag{10.22}$$

再次使用数学归纳法:

- 边界情况:  $F_2 = 1 + F_0 = 2$
- 递归情况,假设  $k + 1$  时成立。

$$\begin{aligned} F_{k+2} &= F_{k+1} + F_k \\ &= \left(1 + \sum_{i=0}^{k-1} F_i\right) + F_k && \text{归纳假设} \\ &= 1 + \sum_{i=0}^k F_i \end{aligned}$$

综上所述得到结论:

$$n \geq |x| \geq F_{k+2} \tag{10.23}$$

□

根据斐波那契数列的性质:  $F_k \geq \phi^k$ , 其中  $\phi = \frac{1 + \sqrt{5}}{2}$  为黄金分割比。我们同时证明了弹出操作的分摊复杂度为  $O(\lg n)$ 。根据这一结果, 可以定义  $maxRank$  为:

$$maxRank(n) = 1 + \lfloor \log_{\phi} n \rfloor \quad (10.24)$$

或者利用斐波那契数列的递归定义实现 MAX-RANK:

```

1: function MAX-RANK( $n$ )
2:    $F_0 \leftarrow 0, F_1 \leftarrow 1$ 
3:    $k \leftarrow 2$ 
4:   repeat
5:      $F_k \leftarrow F_{k-1} + F_{k-2}$ 
6:      $k \leftarrow k + 1$ 
7:   until  $F_k < n$ 
8:   return  $k - 2$ 

```

## 10.4 配对堆

斐波那契堆的实现较为复杂。配对堆实现简单、性能优异。大部分操作, 包括插入、获取堆顶、合并都是常数时间复杂度, 人们猜测它的弹出操作的分摊复杂度为  $O(\lg n)$  [58] [3]。

### 10.4.1 定义

配对堆实现为一棵多叉树。最小元素保存于树根。一个配对堆要么为空  $\emptyset$ , 要么是一棵  $k$  叉树, 包含一个根节点和一组子树, 记为  $(x, ts)$ 。多叉树也可用“左侧孩子, 右侧兄弟”方法进行定义。

**data** PHeap  $a = \text{Empty}$  | Node  $a$  [PHeap  $a$ ]

### 10.4.2 合并、插入、获取堆顶

合并两个配对堆时, 存在两种情况:

1. 任何一个堆为  $\emptyset$ , 结果为另一个堆;
2. 否则, 比较两个堆的根节点, 把较大的一个作为另一个的新子树。

$$\begin{aligned}
 merge \ \emptyset \ h_2 &= h_2 \\
 merge \ h_1 \ \emptyset &= h_1 \\
 merge \ (x, ts_1) \ (y, ts_2) &= \begin{cases} x < y : (x, (y, ts_2):ts_1) \\ \text{否则} : (y, (x, ts_1):ts_2) \end{cases} \quad (10.25)
 \end{aligned}$$

合并的性能为常数时间。使用“左侧孩子,右侧兄弟”方法,我们把根节点较大的堆链接到另一个堆的子树前:

```

1: function MERGE( $H_1, H_2$ )
2:   if  $H_1 = \text{NIL}$  then
3:     return  $H_2$ 
4:   if  $H_2 = \text{NIL}$  then
5:     return  $H_1$ 
6:   if  $\text{KEY}(H_2) < \text{KEY}(H_1)$  then
7:     EXCHANGE( $H_1 \leftrightarrow H_2$ )
8:   SUB-TREES( $H_1$ )  $\leftarrow$  LINK( $H_2, \text{SUB-TREES}(H_1)$ )
9:   PARENT( $H_2$ )  $\leftarrow H_1$ 
10:  return  $H_1$ 

```

使用合并函数,可以像斐波那契堆一样实现插入操作,如式 (10.12)。堆顶元素可以从根节点获取: $\text{top}(x, ts) = x$ 。插入和获取堆顶的复杂度都是常数时间。

### 10.4.3 提升优先级

节点的值减小后,我们将以它为根的子树切下,然后合并到堆中。如果是根节点,我们可以直接减小元素的值。

```

1: function DECREASE( $H, x, k$ )
2:   KEY( $x$ )  $\leftarrow k$ 
3:    $p \leftarrow$  PARENT( $x$ )
4:   if  $p \neq \text{NIL}$  then
5:     Remove  $x$  from SUB-TREES( $p$ )
6:     PARENT( $x$ )  $\leftarrow$  NIL
7:   return MERGE( $H, x$ )
8:  return  $H$ 

```

### 10.4.4 弹出

弹出堆顶的根节点后,我们将剩下的子树归并成一棵树:

$$\text{pop}(x, ts) = \text{consolidate } ts \quad (10.26)$$

我们先从左向右,两两成对地将子树合并。然后再从右向左合并成一棵树。配对堆的名字就来自这一合并过程。如图 10.10 和图 10.11 所示。合并过程和自底向上的归并排序<sup>[3]</sup>类似。



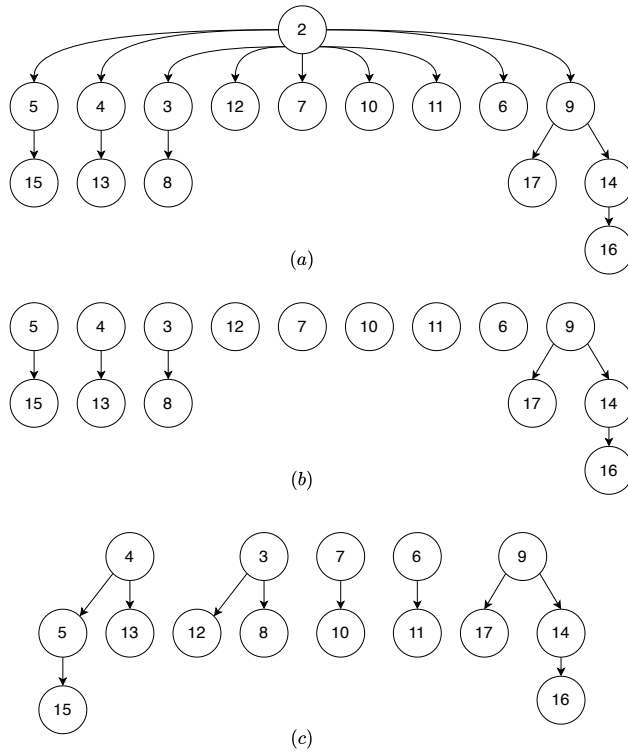


图 10.10: 删除根节点, 将子树成对合并。剩下 9 棵子树中的最后一棵。

$$\begin{aligned}
 \text{consolidate} [ ] &= \emptyset \\
 \text{consolidate} [t] &= t \\
 \text{consolidate} (t_1:t_2:ts) &= \text{merge} (\text{merge } t_1 \ t_2) (\text{consolidate } ts)
 \end{aligned}
 \tag{10.27}$$

对应“左侧孩子, 右侧兄弟”的实现如下:

```

1: function POP( $H$ )
2:    $L \leftarrow \text{NIL}$ 
3:   for every  $T_x, T_y$  in SUB-TREES( $H$ ) do
4:      $T \leftarrow \text{MERGE}(T_x, T_y)$ 
5:      $L \leftarrow \text{LINK}(T, L)$ 
6:    $H \leftarrow \text{NIL}$ 
7:   for  $T$  in  $L$  do
8:      $H \leftarrow \text{MERGE}(H, T)$ 
9:   return  $H$ 

```

从左向右每次成对迭代两棵子树  $T_x, T_y$  合并成  $T$ , 然后链接到  $L$  的前面。这样再次遍历  $L$  时, 实际是按照从右向左的顺序。堆中可能含有奇数棵子树, 这种情况下, 最后一次  $T_y = \text{NIL}$ , 成对合并后  $T = T_x$ 。

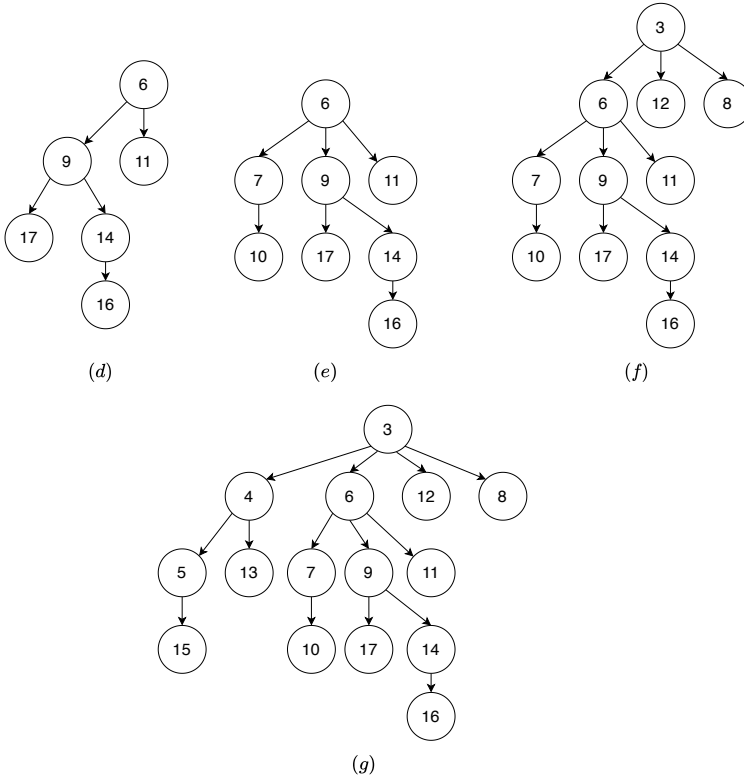


图 10.11: 从右向左合并的步骤:(d) 合并子树 9 和 6;(e) 合并子树 7;(f) 合并子树 3;(g) 合并子树 4。

### 10.4.5 删除

删除某个节点  $x$  时,可先将节点的值减小为  $-\infty$ ,然后再执行一次弹出操作。本节介绍另外一种删除方法。若  $x$  为根节点,我们只需要执行一次弹出操作。否则,我们将  $x$  从堆  $H$  中切下,然后对  $x$  执行一次弹出操作,再将结果合并回  $H$ :

```

1: function DELETE( $H, x$ )
2:   if  $H = x$  then
3:     POP( $H$ )
4:   else
5:      $H \leftarrow$  CUT( $H, x$ )
6:      $x \leftarrow$  POP( $x$ )
7:     MERGE( $H, x$ )

```

因为删除算法调用弹出操作,我们猜想它的分摊性能也是对数时间  $O(\lg n)$ 。

### 练习 10.3

- 10.3.1. 如果连续向配对堆中插入  $n$  个元素,然后执行弹出。当  $n$  很大时,弹出的单次性能变得很差(尽管分摊性能是  $O(\lg n)$ )。如何避免这一最差情况?
- 10.3.2. 实现配对堆的删除。
- 10.3.3. 实现配对堆的优先级调整 DECREASE-KEY

答案: 380 页

## 10.5 附录:例子程序

多叉树的定义(左侧孩子,右侧兄弟):

```

data Node<K> {
  Int rank
  K key
  Node<K> parent, subTrees, sibling,
  Bool mark

  Node(K x) {
    key = x
    rank = 0
    parent = subTrees = sibling = null
    mark = false
  }
}

```

二项式树链接:

```

Node<K> link(Node<K> t1, Node<K> t2) {
  if t2.key < t1.key then (t1, t2) = (t2, t1)
  t2.sibling = t1.subTrees
}

```

```

    t1.subTrees = t2
    t2.parent = t1
    t1.rank = t1.rank + 1
    return t1
}

```

二项式堆插入:

```

Node<K> insert(K x, Node<K> h) = insertTree(Node(x), h)

Node<K> insertTree(Node<K> t, Node<K> h) {
    var h1 = Node()
    var prev = h1
    while h ≠ null and h.rank ≤ t.rank {
        var t1 = h
        h = h.sibling
        if t.rank == t1.rank {
            t = link(t, t1)
        } else {
            prev.sibling = t1
            prev = t1
        }
    }
    prev.sibling = t
    t.sibling = h
    return removeFirst(h1)
}

Node<K> removeFirst(Node<K> h) {
    var next = h.sibling
    h.sibling = null
    return next
}

```

二项式堆插入的递归实现:

```

data BiTree a = Node { rank :: Int
                      , key :: a
                      , subTrees :: [BiTree a]}

type BiHeap a = [BiTree a]

link t1@(Node r x c1) t2@(Node _ y c2) =
    if x < y then Node (r + 1) x (t2:c1)
    else Node (r + 1) y (t1:c2)

insertTree t [] = [t]
insertTree t ts@(t':ts') | rank t < rank t' = t:ts
                        | rank t > rank t' = t' : insertTree t ts'
                        | otherwise = insertTree (link t t') ts'

insert x = insertTree (Node 0 x [])

```

二项式堆的合并:

```

Node<K> merge(h1, h2) {
    var h = Node()
    var prev = h
    while h1 ≠ null and h2 ≠ null {
        if h1.rank < h2.rank {
            prev.sibling = h1
            prev = prev.sibling
            h1 = h1.sibling
        } else if h2.rank < h1.rank {
            prev.sibling = h2
            prev = prev.sibling
            h2 = h2.sibling
        } else {
            var (t1, t2) = (h1, h2)
            (h1, h2) = (h1.sibling, h2.sibling)
            h1 = insertTree(link(t1, t2), h1)
        }
    }
    if h1 ≠ null then prev.sibling = h1
    if h2 ≠ null then prev.sibling = h2
    return removeFirst(h)
}

```

递归合并两个二项式堆:

```

merge ts1 [] = ts1
merge [] ts2 = ts2
merge ts1@(t1:ts1') ts2@(t2:ts2')
    | rank t1 < rank t2 = t1:(merge ts1' ts2)
    | rank t1 > rank t2 = t2:(merge ts1 ts2')
    | otherwise = insertTree (link t1 t2) (merge ts1' ts2')

```

二项式堆的弹出

```

Node<K> reverse(Node<K> h) {
    Node<K> prev = null
    while h ≠ null {
        var x = h
        h = h.sibling
        x.sibling = prev
        prev = x
    }
    return prev
}

(Node<K>, Node<K>) extractMin(Node<K> h) {
    var head = h
    Node<K> tp = null
    Node<K> tm = null
    Node<K> prev = null
    while h ≠ null {
        if tm == null or h.key < tm.key {
            tm = h
            tp = prev
        }
    }
}

```

```

    }
    prev = h
    h = h.sibling
  }
  if tp ≠ null {
    tp.sibling = tm.sibling
  } else {
    head = tm.sibling
  }
  tm.sibling = null
  return (tm, head)
}

(K, Node<K>) pop(Node<K> h) {
  var (tm, h) = extractMin(h)
  h = merge(h, reverse(tm.subtrees))
  tm.subtrees = null
  return (tm.key, h)
}

```

二项式堆弹出的递归实现：

```

pop h = merge (reverse $ subTrees t) ts where
  (t, ts) = extractMin h

extractMin [t] = (t, [])
extractMin (t:ts) = if key t < key t' then (t, ts)
                  else (t', t:ts') where
  (t', ts') = extractMin ts

```

使用双向链表合并斐波那契堆：

```

FibHeap<K> merge(FibHeap<K> h1, FibHeap<K> h2) {
  if isEmpty(h1) then return h2
  if isEmpty(h2) then return h1
  FibHeap<K> h = FibHeap<K>()
  h.trees = concat(h1.trees, h2.trees)
  h.minTree = if h1.minTree.key < h2.minTree.key
              then h1.minTree else h2.minTree
  h.size = h1.size + h2.size
  return h
}

bool isEmpty(FibHeap<K> h) = (h == null or h.trees == null)

Node<K> concat(Node<K> first1, Node<K> first2) {
  var last1 = first1.prev
  var last2 = first2.prev
  last1.next = first2
  first2.prev = last1
  last2.next = first1
  first1.prev = last2
  return first1
}

```

斐波那契树归并:

```
consolidate = foldr melt [] where
  melt t [] = [t]
  meld t (t':ts) | rank t == rank t' = meld (link t t') ts
                 | rank t < rank t' = t : t' : ts
                 | otherwise = t' : meld t ts
```

使用辅助数组进行归并:

```
void consolidate(FibHeap<K> h) {
  Int R = maxRank(h.size) + 1
  Node<K>[R] a = [null, ...]
  while h.trees ≠ null {
    var x = h.trees
    h.trees = remove(h.trees, x)
    Int r = x.rank
    while a[r] ≠ null {
      var y = a[r]
      x = link(x, y)
      a[r] = null
      r = r + 1
    }
    a[r] = x
  }
  h.minTr = null
  h.trees = null
  for var t in a if t ≠ null {
    h.trees = append(h.trees, t)
    if h.minTr == null or t.key < h.minTr.key then h.minTr = t
  }
}
```

斐波那契堆的弹出:

```
pop (FibHeap _ (Node _ x []) []) = (x, Empty)
pop (FibHeap sz (Node _ x tsm) ts) = (x, FibHeap (sz - 1) tm ts') where
  (tm, ts') = extractMin $ consolidate (tsm # ts)
```

提升优先级:

```
void decrease(FibHeap<K> h, Node<K> x, K k) {
  var p = x.parent
  x.key = k
  if p ≠ null and k < p.key {
    cut(h, x)
    cascadeCut(h, p)
  }
  if k < h.minTr.key then h.minTr = x
}

void cut(FibHeap<K> h, Node<K> x) {
  var p = x.parent
  p.subTrees = remove(p.subTrees, x)
}
```

```
p.rank = p.rank - 1
h.trees = append(h.trees, x)
x.parent = null
x.mark = false
}

void cascadeCut(FibHeap<K> h, Node<K> x) {
    var p = x.parent
    if p == null then return
    if x.mark {
        cut(h, x)
        cascadeCut(h, p)
    } else {
        x.mark = true
    }
}
```



# 第十一章 队列

队列提供了先进先出(FIFO)机制。可以用多种方法实现队列,如单向、双向链表,循环缓冲区等,冈崎给出了 16 种不同的实现方法<sup>[3]</sup>。队列需要满足两条基本要求:

1. 可以在常数时间内向末尾添加元素;
2. 可以在常数时间内从头部获取或删除元素。

我们略去双向链表的直观实现,而关注如何用其它基本数据结构(如列表、数组)实现队列。

## 11.1 列表实现

可以用常数时间在列表头部插入、删除元素。但先进先出要求只能在头部执行一种操作,而在尾部执行另一种操作。需要  $O(n)$  时间遍历列表到达尾部,其中  $n$  是列表长度,无法满足常数时间的性能要求。为了解决这个问题,我们增加一个变量记录尾部位置,并用一个额外节点  $S$  简化空队列处理,如图 11.1 所示。

```
data Node<K> {
  K key
  Node<K> next
}

data Queue<K> {
  Node<K> head, tail
}
```

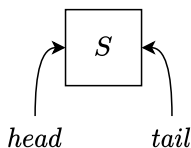


图 11.1: 空队列,头、尾都指向  $S$

最基本的两个操作是入队(Enqueue, 或 push、snoc、append、push back)和出队(Dequeue, 或 pop、pop front)。我们选择在列表头部加入元素、从列表尾部删除元素。

```

1: function ENQUEUE( $Q, x$ )
2:    $p \leftarrow \text{NODE}(x)$ 
3:    $\text{NEXT}(p) \leftarrow \text{NIL}$ 
4:    $\text{NEXT}(\text{TAIL}(Q)) \leftarrow p$ 
5:    $\text{TAIL}(Q) \leftarrow p$ 

```

队列至少有一个节点(空队列中有  $S$  节点),因此无需检查尾部是否为 NIL。

```

1: function DEQUEUE( $Q$ )
2:    $x \leftarrow \text{HEAD}(Q)$ 
3:    $\text{NEXT}(\text{HEAD}(Q)) \leftarrow \text{NEXT}(x)$ 
4:   if  $x = \text{TAIL}(Q)$  then ▷  $Q$  变为空
5:      $\text{TAIL}(Q) \leftarrow \text{HEAD}(Q)$ 
6:   return  $\text{KEY}(x)$ 

```

$S$  节点在所有其它节点的前面,HEAD 实际上返回  $S$  的下一个节点,如图 11.2 所示。这一实现可以扩展到并发环境。在头部和尾部各使用一把并发锁。 $S$  节点可以避免在队列空时死锁<sup>[59]、[60]</sup>。

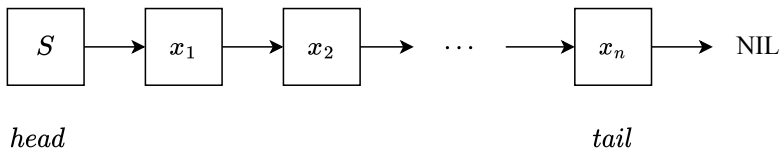


图 11.2: 带有  $S$  节点的列表

## 11.2 循环缓冲区

和列表对称,可以在常数时间向数组末尾添加元素,但需要线性时间  $O(n)$  从头部删除。这是因为要将剩余元素依次向前平移。为了达到队列的性能要求,可以把数组头尾连接起来,做成一个环,叫做循环缓冲区,如图 11.3、图 11.4 所示。用头部位置 head、队列长度 count、数组大小 size,就可以完全描述队列。count 等于 0 时队列为空、等于 size 时队列已满。我们还可以利用模运算简化入队、出队的实现。

```

1: function ENQUEUE( $Q, x$ )
2:   if not FULL( $Q$ ) then
3:     COUNT( $Q$ )  $\leftarrow$  COUNT( $Q$ ) + 1
4:     tail  $\leftarrow$  (HEAD( $Q$ ) + COUNT( $Q$ )) mod SIZE( $Q$ )
5:     BUF( $Q$ )[tail]  $\leftarrow$   $x$ 
1: function DEQUEUE( $Q$ )
2:    $x \leftarrow \text{NIL}$ 

```

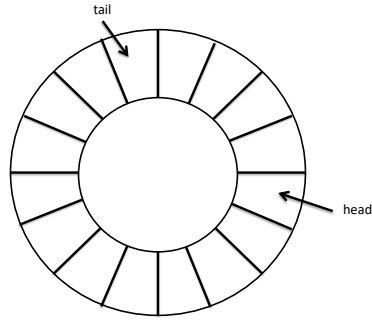


图 11.3: 循环缓冲区

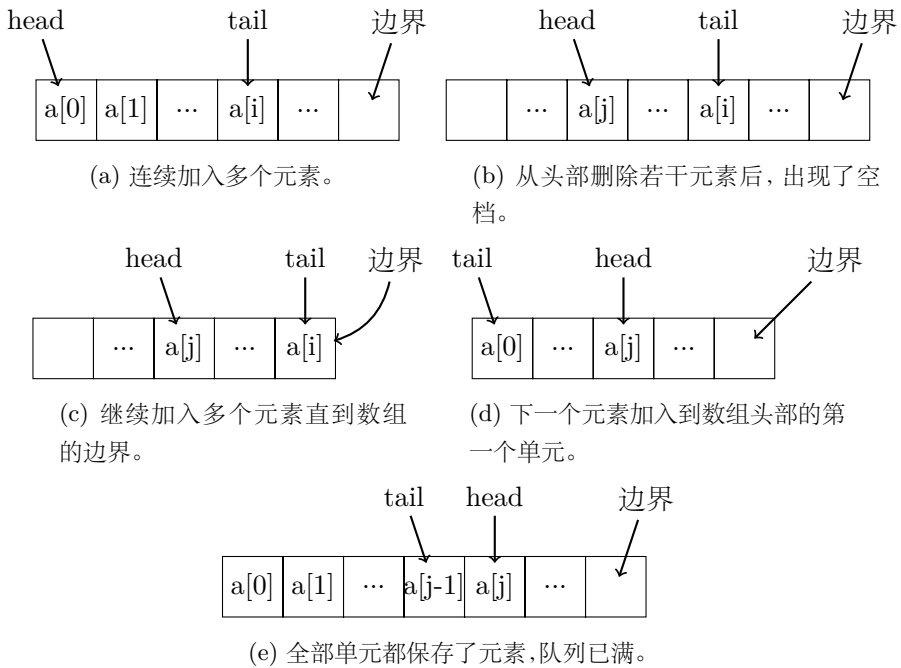


图 11.4: 使用循环缓冲区实现队列

```

3:  if not EMPTY( $Q$ ) then
4:      $h \leftarrow$  HEAD( $Q$ )
5:      $x \leftarrow$  BUF( $Q$ )[ $h$ ]
6:     HEAD( $Q$ )  $\leftarrow$  ( $h + 1$ ) mod SIZE( $Q$ )
7:     COUNT( $Q$ )  $\leftarrow$  COUNT( $Q$ ) - 1
8:  return  $x$ 

```

### 练习 11.1

循环缓冲区在初始化时规定了最大的容量, 如果使用头、尾两个指针, 而不用 count, 如何检测队列是否为空? 是否已满?

答案: 381 页

## 11.3 双列表队列

列表的头部操作为常数时间, 但尾部需要线性时间。可以把两个列表“尾对尾”连起来实现队列。形状类似一个马蹄形磁铁, 如图 11.5 所示。两个列表分别叫做前 (front)、后 (rear)。记队列为  $(f, r)$ , 空队列等于  $([], [])$ 。入队时把元素加入  $r$  的头部, 出队时把元素从  $f$  的头部取走, 性能都是常数时间。

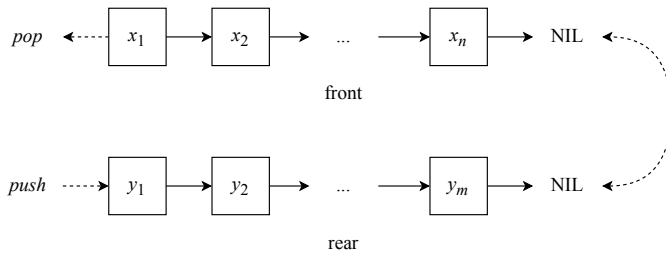


图 11.5: 双列表队列

$$\begin{cases} \text{push } x (f, r) = (f, x:r) \\ \text{pop } (x:f, r) = (f, r) \end{cases} \quad (11.1)$$

经过一系列出队操作后,  $f$  可能为空, 而  $r$  中还有元素。为了能继续出队, 我们将  $r$  反转后替换掉  $f$ , 即:  $([], r) \mapsto (\text{reverse } r, [])$ 。每次出入队后, 需要执行一次平衡检查和调整:

$$\begin{aligned} \text{balance } [] r &= (\text{reverse } r, []) \\ \text{balance } f r &= (f, r) \end{aligned} \quad (11.2)$$

一旦发生  $r$  的反转, 则这次操作的性能下降为线性时间。尽管如此, 整体的分摊复杂度是常数时间的。重新定义入队和出队为:

$$\begin{cases} \text{push } x(f, r) = \text{balance } f(x:r) \\ \text{pop}(x:f, r) = \text{balance } f r \end{cases} \quad (11.3)$$

可以用数组构造出双列表的对称实现。利用表 11.1 的对称性,将两个数组“头对头”连接起来形成队列,如图 11.6 所示。当  $R$  数组为空时,将  $F$  反转替换掉  $R$ 。

操作	数组	链表
在头部加入	$O(n)$	$O(1)$
在尾部加入	$O(1)$	$O(n)$
在头部删除	$O(n)$	$O(1)$
在尾部删除	$O(1)$	$O(n)$

表 11.1: 数组和链表各项操作的对比

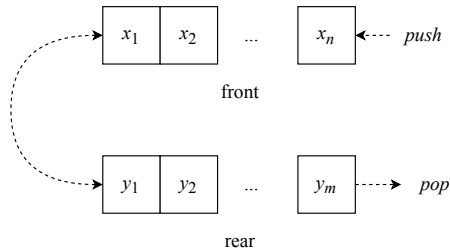


图 11.6: 双数组队列

## 练习 11.2

- 11.2.1. 为什么要在  $\text{push}$  时也要进行平衡检查和调整?
- 11.2.2. 试分析双列表队列的分摊复杂度。
- 11.2.3. 实现双数组队列。

答案: 382 页

## 11.4 平衡队列

虽然双列表队列的分摊复杂度为常数时间,但最坏情况下的性能是线性的。例如  $f$  中有一个元素,此后连续将  $n$  个元素加入队列,这时出队的复杂度为  $O(n)$ 。问题的原因是  $f$  和  $r$  不平衡。为了改进平衡性,我们加入一条规则,要求  $r$  的长度不大于  $f$  的长度,否则就反转列表。

$$|r| \leq |f| \quad (11.4)$$

为避免反复遍历获取列表长度,我们将长度记录下来,并在出入队时更新。将双列表队列表示为  $(f, n, r, m)$ , 其中  $n = |f|, m = |r|$ , 分别是两个列表的长度。根据平衡规则式 (11.4), 可以检查  $f$  的长度来判空:

$$Q = \phi \iff n = 0 \quad (11.5)$$

更新出、入队的定义为:

$$\begin{cases} \text{push } x (f, n, r, m) & = \text{balance } (f, n, x:r, m + 1) \\ \text{pop } (x:f, n, r, m) & = \text{balance } (f, n - 1, r, m) \end{cases} \quad (11.6)$$

其中:

$$\text{balance } (f, n, r, m) = \begin{cases} m \leq n : & (f, n, r, m) \\ \text{否则} : & (f \# \text{reverse } r, m + n, [], 0) \end{cases} \quad (11.7)$$

## 11.5 实时队列

平衡队列仍然需要线性时间连接、反转列表。在实时系统中需要进一步改进。性能瓶颈出现在  $f \# \text{reverse } r$  中。此时  $m > n$ , 违反了平衡规则。由于  $m, n$  都是整数, 此时:  $m = n + 1$ 。 $\#$  的复杂度是  $O(n)$ , 反转的复杂度是  $O(m)$ , 总复杂度是  $O(n + m)$ , 和队列中元素个数成正比。我们可以将它分派到各次出、入队中。首先分析一下尾递归的反转(柯里化):

$$\text{reverse} = \text{reverse}' [] \quad (11.8)$$

其中:

$$\begin{aligned} \text{reverse}' a [] &= a \\ \text{reverse}' a (x:xs) &= \text{reverse}' (x:a) xs \end{aligned} \quad (11.9)$$

可以将尾递归<sup>[61][62]</sup>定义转换为逐步计算。整体过程相当于一系列的状态转换。定义一个状态机, 包含两种状态: 反转状态  $S_r$  表示正在进行反转(未完成); 完成状态  $S_f$  表示反转已结束(完成)。接下来我们利用状态机调度(slow-down)反转计算:

$$\begin{aligned} \text{step } S_r a [] &= (S_f, a) \\ \text{step } S_r a (x:xs) &= (S_r, (x:a), xs) \end{aligned} \quad (11.10)$$

每一步先检查当前的状态, 如果为  $S_r$  (反转中), 但是列表中已没有剩余元素需要反转, 就将状态变为完成  $S_f$ ; 否则, 取出列表中的第一个元素  $x$ , 将其链结到  $a$  的前面。这一步计算到此结束, 不再进行递归。当前状态和反转的中间结果被保存下来, 供以后再次调用  $\text{step}$  时使用。例如:

$$\begin{aligned}
 \text{step } S_r \text{ "hello" } [ ] &= (S_r, \text{"ello"}, \text{"h"}) \\
 \text{step } S_r \text{ "ello" "h"} &= (S_r, \text{"llo"}, \text{"eh"}) \\
 &\dots \\
 \text{step } S_r \text{ "o" "lleh"} &= (S_r, [ ], \text{"olleh"}) \\
 \text{step } S_r [ ] \text{ "olleh"} &= (S_f, \text{"olleh"})
 \end{aligned}$$

但这仅解决了一半问题。我们还需要逐步分解、调度  $\#$ 。利用逐步反转的结果，并使用一个技巧：为了实现  $xs \# ys$ ，我们可以先将  $xs$  反转为  $\overleftarrow{xs}$ ，然后逐一将  $\overleftarrow{xs}$  中的元素取出，放到  $ys$  的前面。这和  $\text{reverse}'$  类似。

$$\begin{aligned}
 xs \# ys &= (\text{reverse reverse } xs) \# ys \\
 &= (\text{reverse}' [ ] \overleftarrow{xs}) \# ys \\
 &= \text{reverse}' ys \overleftarrow{xs}
 \end{aligned} \tag{11.11}$$

需要增加另一个状态来控制  $\text{step}$ ，在  $r$  反转后，逐步操作  $\overleftarrow{f}$  实现连接。三种状态为：反转  $S_r$ 、连接  $S_c$ 、完成  $S_f$ 。整个操作被分解为两个阶段：

1. 同时反转  $f$  和  $r$ ，逐步得到  $\overleftarrow{f}$  和  $\overleftarrow{r}$ ；
2. 逐步从  $\overleftarrow{f}$  取出元素，链接到  $\overleftarrow{r}$  前面。

$$\begin{aligned}
 \text{next } (S_r, f', x:f, r', y:r) &= (S_r, x:f', f, y:r', r) \quad \text{同时反转 } f, r \\
 \text{next } (S_r, f', [ ], r', [y]) &= \text{next } (S_c, f', y:r') \quad \text{反转结束、转入连接} \\
 \text{next } (S_c, a, [ ]) &= (S_f, a) \quad \text{连接结束} \\
 \text{next } (S_c, a, x:f') &= (S_c, x:a, f') \quad \text{逐步连接}
 \end{aligned} \tag{11.12}$$

接下来需要将这些递进的步骤分配到每个出、入队操作中以实现实时队列。根据平衡条件，当  $m = n + 1$  时，开始逐步计算  $f \# \text{reverse } r$ 。总共需要  $n + 1$  步来反转  $r$ ，同时在这些步骤内完成了对  $f$  的反转。此后，需要再用  $n + 1$  步来进行连接操作。因此总共花费了  $2n + 2$  步。最直接的想法是在每一个出、入队中分配一个递进步骤。但这里有一个关键问题：在完成  $2n + 2$  步操作之前，队列有没有可能由于接下来的一系列出、入队操作再次变得不平衡？

幸运的是，在完成  $f \# \text{reverse } r$  之前，连续的入队操作不可能再次破坏平衡。一旦开始恢复平衡的处理，经过  $2n + 2$  步后，我们就得到了一个新的  $f$  列表  $f' = f \# \text{reverse } r$ 。而再次不平衡时有：

$$\begin{aligned}
 |r'| &= |f'| + 1 \\
 &= |f| + |r| + 1 \\
 &= 2n + 2
 \end{aligned} \tag{11.13}$$

也就是说，从上次不平衡的时刻算起，即使不断持续入队新元素，以最快的速度再次使得队列不平衡时， $2n + 2$  步计算恰好已经完成了。此时新的  $f$  列表被计算出来。我们可以安全地继续计算  $f' \# \text{reverse } r'$ 。多亏了平衡规则，帮助我们保证了这一点。

但不幸的是,在  $2n + 2$  步计算完成前,出队操作可能随时发生。这会产生一个尴尬的情况:我们需要从  $f$  列表取出元素,但是新的  $f$  列表  $f' = f \# reverse r$  尚未计算好。此时没有一个可用的  $f$  列表。为了解决这个问题,我们在第一阶段并行计算  $reverse f$  时,另外保存一份  $f$  的副本。这样即使连续进行  $n$  次出队操作,仍然是安全的。表 11.2 给出了第一阶段逐步计算(同时反转  $f$  和  $r$ )的某个中间状态<sup>1</sup>。

保存的 $f$ 副本	进行中的计算	新的 $r$ 列表
$\{f_i, f_{i+1}, \dots, f_n\}$	$(S_r, \tilde{f}, \dots, \tilde{r}, \dots)$	$\{\dots\}$
前 $i - 1$ 个元素已出队	$\overleftarrow{f}$ 和 $\overleftarrow{r}$ 的中间结果	包含新入队的元素

表 11.2: 前  $n$  步完成之前的中间状态

经过  $n$  次出队,  $f$  的副本已经用光。此时刚刚开始逐步连接的阶段,如果继续出队会怎样?事实上,由于  $f$  的副本被用光,变成了  $[\ ]$ ,我们无需再进行连接了。因为  $f \# \overleftarrow{r} = [\ ] \# \overleftarrow{r} = \overleftarrow{r}$ 。事实上,只需要将  $f$  中尚未出队的部分连接起来。由于元素从  $f$  的头部逐一出队,可以使用一个计数器来记录  $f$  中剩余元素的个数。当开始计算  $f \# reverse r$  时,计数器为 0,每反转  $f$  中的一个元素时,就将计数器加一,表示将来需要连接这个元素;每次出队将计数器减一,表示将来可以少连接一个元素。显然在连接操作的每步中也需要递减计数器。当且仅当计数器为 0 的时候,无需继续进行连接操作。下面是增加了计数器的状态转换定义:

$$\begin{aligned}
 next(S_r, n, f', x:f, r', y:r) &= (S_r, n + 1, x:f', f, y:r', r) && \text{同时反转 } f, r \\
 next(S_r, n, f', [\ ], r', [y]) &= next(S_c, n, f', y:r') && \text{反转结束、转入连接} \\
 next(S_c, 0, a, f) &= (S_f, a) && \text{连接结束} \\
 next(S_c, n, a, x:f') &= (S_c, n - 1, x:a, f') && \text{逐步连接} \\
 next S_0 &= S_0 && \text{空闲状态}
 \end{aligned} \tag{11.14}$$

我们还定义了一个空闲状态  $S_0$  来简化控制逻辑。队列的数据结构分为三个部分:  $f$  列表及其长度  $n$ 、正在计算中的  $f \# reverse r$  的中间状态、 $r$  列表及其长度  $m$ 。记为  $(f, n, S, r, m)$ 。空闲队列记为  $([\ ], 0, S_0, [\ ], 0)$ 。根据平衡规则当  $n = 0$  时队列为空。修改出、入队定义为:

$$\begin{cases}
 push x(f, n, S, r, m) = balance f n S(x:r)(m + 1) \\
 pop(x:f, n, S, r, m) = balance f(n - 1)(abort S) r m
 \end{cases} \tag{11.15}$$

其中  $abort$  在出队时递减计数器,这样将来可以少连接一个元素。我们稍后定义这一撤销操作。 $balance$  检查平衡规则,若不满足则启动  $f \# reverse r$  逐步恢复平衡;否则执行一步尚未完成的递进计算:

<sup>1</sup>有人问:通常需要线性时间复制列表,这样整个方案就有问题了。实际上,线性时间的列表复制根本不会发生。我们复制的是  $f$  列表的引用。每个元素的复制被推迟到后续各个步骤中。



$$\text{balance } f \ n \ S \ r \ m = \begin{cases} m \leq n : & \text{step } f \ n \ S \ r \ m \\ \text{否则} : & \text{step } f \ (n + m) \ (\text{next } (S_r, 0, [], f, [], r)) \ [] \ 0 \end{cases} \quad (11.16)$$

其中 *step* 将状态机转换到下一个状态,全部计算结束后,转换到空闲状态  $S_0$ 。

$$\text{step } f \ n \ S \ r \ m = \text{queue } (\text{next } S) \quad (11.17)$$

其中:

$$\begin{aligned} \text{queue } (S_f, f') &= (f', n, S_0, r, m) \quad \text{用逐步计算结果 } f' \text{ 替换 } f \\ \text{queue } S' &= (f, n, S', r, m) \end{aligned} \quad (11.18)$$

我们还需要实现 *abort* 函数,指示状态机,由于出队,可以少连接一个元素。

$$\begin{aligned} \text{abort } (S_c, 0, (x:a), f') &= (S_f, a) \\ \text{abort } (S_c, n, a, f') &= (S_c, n - 1, a, f') \\ \text{abort } (S_r, n, f' f, r' r) &= (S_r, n - 1, f', f, r', r) \\ \text{abort } S &= S \end{aligned} \quad (11.19)$$

### 练习 11.3

- 11.3.1. 在 *abort* 函数中,当  $n = 0$  时,我们实际上撤销了上一个链接元素的操作,去掉了  $x$  而返回  $a$  作为结果。为什么需要回滚一个元素?
- 11.3.2. 使用双数组实现实时队列。注意:当开始递进反转时,不能一次性复制数组,否则就会将性能降低到线性时间。请实现一个惰性复制,使得每步反转时仅复制一个元素。

答案: 383 页

## 11.6 惰性实时队列

实时队列的关键在于将耗时的  $f \# \text{reverse } r$  计算分解。利用惰性求值可以得到一个简化的实现。假设函数 *rotate* 可以逐步计算  $f \# \text{reverse } r$ 。也就是说,使用一个累积器  $a$ ,下面的两个函数等价:

$$\text{rotate } xs \ ys \ a = xs \# (\text{reverse } ys) \# a \quad (11.20)$$

我们将  $xs$  初始化为  $f$  列表,  $ys$  初始化为  $r$  列表,  $a$  初始化为空  $[]$ 。为了实现轮转,我们先考虑边界情况:

$$\text{rotate } [] \ [y] \ a = y:a \quad (11.21)$$

递归情况为:

$$\begin{aligned}
& rotate(x:xs)(y:ys) a \\
= & (x:xs) \# (reverse(y:ys)) \# a \quad \text{定义式 (11.20)} \\
= & x : (xs \# reverse(y:ys)) \# a \quad \text{连接的结合性} \\
= & x : (xs \# reverse ys \# (y:a)) \quad \text{反转的性质和连接的结合性} \\
= & x : rotate xs ys (y:a) \quad \text{反向用定义式 (11.20)}
\end{aligned} \tag{11.22}$$

归纳上面的两种情况,可以得到最终的轮转算法:

$$\begin{aligned}
rotate [] [y] a &= y:a \\
rotate (x:xs) (y:ys) a &= x : rotate xs ys (y:a)
\end{aligned} \tag{11.23}$$

在惰性执行环境中,(:)操作会推迟到出、入队时才执行,这样就将 *rotate* 计算自然分摊了。修改双列表的定义为  $(f,r,rot)$ ,其中 *rot* 表示正在进行的轮转计算  $f \# reverse r$ ,它初始为空 []。

$$\begin{cases}
push\ x\ (f,r,rot) = balance\ f\ (x:r)\ rot \\
pop\ (x:f,r,rot) = balance\ f\ r\ rot
\end{cases} \tag{11.24}$$

每次 *balance* 操作都会向前推进一次轮转,轮转结束时开始新一轮计算:

$$\begin{aligned}
balance\ f\ r\ [] &= (f',[],f') \quad \text{其中 } f' = rotate\ f\ r\ [] \\
balance\ f\ r\ (x:rot) &= (f,r,rot) \quad \text{推进轮转}
\end{aligned} \tag{11.25}$$

## 练习 11.4

11.4.1. 如何实现双向队列,在头部尾部都支持常数时间的元素添加和删除。

答案: 384 页

## 11.7 附录:例子程序

列表实现的出、入队:

```

Queue<K> enQ(Queue<K> q, K x) {
    var p = Node(x)
    p.next = null
    q.tail.next = p
    q.tail = p
    return q
}

K deQ(Queue<K> q) {
    var p = q.head.next //the next of S
    q.head.next = p.next
    if q.tail == p then q.tail = q.head //empty
    return p.key
}

```

循环缓冲区的定义:

```

data Queue<K> {
  [K] buf
  int head, cnt, size

  Queue(int max) {
    buf = Array<K>(max)
    size = max
    head = cnt = 0
  }
}

```

使用循环缓冲区的出、入队:

```

N offset(N i, N size) = if i < size then i else i - size

void enQ(Queue<K> q, K x) {
  if q.cnt < q.size {
    q.buf[offset(q.head + q.cnt, q.size)] = x;
    q.cnt = q.cnt + 1
  }
}

K head(Queue<K> q) = if q.cnt == 0 then null else q.buf[q.head]

K deQ(Queue<K> q) {
  K x = null
  if q.cnt > 0 {
    x = head(q)
    q.head = offset(q→head + 1, q→size);
    q.cnt = q.cnt - 1
  }
  return x
}

```

实时队列

```

data State a = Empty
  | Reverse Int [a] [a] [a] [a] — n, acc f, f, acc r, r
  | Concat Int [a] [a] — n, acc, reversed f
  | Done [a] — f' = f ++ reverse r

— f, n = length f, state, r, m = length r
data RealtimeQueue a = RTQ [a] Int (State a) [a] Int

push x (RTQ f n s r m) = balance f n s (x:r) (m + 1)

pop (RTQ (_:f) n s r m) = balance f (n - 1) (abort s) r m

top (RTQ (x:_) _ _ _ _) = x

balance f n s r m
  | m ≤ n = step f n s r m

```

```

| otherwise = step f (m + n) (next (Reverse 0 [] f [] r)) [] 0

step f n s r m = queue (next s) where
  queue (Done f') = RTQ f' n Empty r m
  queue s' = RTQ f n s' r m

next (Reverse n f' (x:f) r' (y:r)) = Reverse (n + 1) (x:f') f (y:r') r
next (Reverse n f' [] r' [y]) = next $ Concat n (y:r') f'
next (Concat 0 acc _) = Done acc
next (Concat n acc (x:f')) = Concat (n-1) (x:acc) f'
next s = s

abort (Concat 0 (_:acc) _) = Done acc — rollback 1 elem
abort (Concat n acc f') = Concat (n - 1) acc f'
abort (Reverse n f' f r' r) = Reverse (n - 1) f' f r' r
abort s = s

```

惰性实时队列:

```

data LazyRTQueue a = LQ [a] [a] [a] — front, rear, f ++ reverse r

empty = LQ [] [] []

push (LQ f r rot) x = balance f (x:r) rot

pop (LQ (_:f) r rot) = balance f r rot

top (LQ (x:_) _ _) = x

balance f r [] = let f' = rotate f r [] in LQ f' [] f'
balance f r (_:rot) = LQ f r rot

rotate [] [y] acc = y:acc
rotate (x:xs) (y:ys) acc = x : rotate xs ys (y:acc)

```

## 第十二章 序列

序列是数组和列表的一种组合。我们希望理想的序列满足：

1. 可以在头部、尾部以常数时间插入、删除元素；
2. 可以快速(优于线性时间)连接两个序列；
3. 可以快速随机访问、更改任何元素；
4. 可以快速在指定位置断开序列。

数组、列表仅部分满足这些要求,如下表所示。其中  $n$  为单个序列的长度,  $n_1$ 、 $n_2$  分别是两个序列的长度。

操作	数组	列表
在头部插入、删除	$O(n)$	$O(1)$
在尾部插入、删除	$O(1)$	$O(n)$
连接	$O(n_2)$	$O(n_1)$
随机访问位置 $i$	$O(1)$	$O(i)$
在位置 $i$ 删除	$O(n - i)$	$O(1)$

本章给出三种序列实现:二叉随机访问列表、可连接列表、手指树。

### 12.1 二叉随机访问列表

二叉随机访问列表是由二叉树森林实现的随机访问列表。森林包含若干完全二叉树。元素只保存在叶子节点中。对任何非负整数  $n$ , 将其表示为二进制, 我们就知道需要多少棵完全二叉树来存储  $n$  个元素。每个值为 1 的二进制位代表一棵二叉树, 树的大小对应着二进制位的高低。任给索引  $1 \leq i \leq n$ , 我们都可以快速在森林中定位到保存第  $i$  个节点的二叉树。如图 12.1 所示, 树  $t_1$ 、 $t_2$  表示序列  $[x_1, x_2, x_3, x_4, x_5, x_6]$ 。

记深度为  $i + 1$  的完全二叉树为  $t_i$ 。  $t_0$  只含有一个叶子节点。  $t_i$  含有  $2^i$  个叶子。对于  $n$  个元素的序列, 我们把  $n$  表示为二进制数  $n = (e_m e_{m-1} \dots e_1 e_0)_2$ , 其中  $e_i$  为 1 或 0。

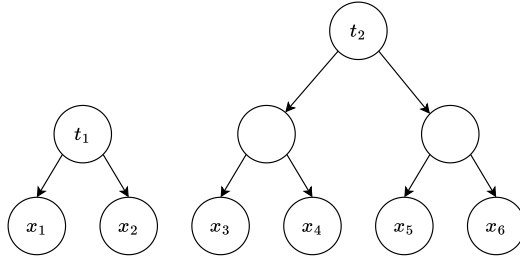


图 12.1: 含有 6 个元素的序列

$$n = 2^0 e_0 + 2^1 e_1 + \dots + 2^m e_m \quad (12.1)$$

如果  $e_i \neq 0$ , 就需要一棵大小为  $2^i$  的完全二叉树  $t_i$ 。在图 12.1 的例子中, 序列长度为  $6 = (110)_2$ 。最低位是 0, 我们不需要大小为 1 的树; 第 2 位是 1, 需要一棵大小为 2 的树  $t_1$ ; 最高位是 1, 需要一棵大小为 4 的树  $t_2$ 。这样就把序列  $[x_1, x_2, \dots, x_n]$  表示为树的列表。列表中每棵树的大小都是唯一的, 并按照从小到大排列。我们称之为**二叉随机访问列表**<sup>[3]</sup>。我们可以在二叉树定义的基础上稍作变化以实现这种列表: 1、元素只保存在叶子节点中, 2、在每棵子树中记录树的大小。这样每个分枝节点记为  $(s, l, r)$ , 其中  $s$  表示子树的大小,  $l, r$  分别表示左右子树。包含元素  $x$  的叶子节点记为  $(x)$ 。我们可以这样获取一棵树的大小:

$$\begin{aligned} \text{size}(x) &= 1 \\ \text{size}(s, l, r) &= s \end{aligned} \quad (12.2)$$

为了把新元素  $y$  插入到序列  $S$  的前面, 我们创建一棵只有一个叶子节点的  $t_0$  树:  $t' = (y)$ , 然后把它插入到森林中。  $\text{insert } y \text{ } S = \text{insert}_T(y) S$ , 或写成柯里化形式:

$$\text{insert } y = \text{insert}_T(y) \quad (12.3)$$

我们检查森林中的第一棵树  $t_i$ , 比较  $t_i$  和  $t'$  的大小, 如果  $t_i$  较大, 就将  $t'$  置于森林最前面(常数时间); 若  $t_i$  和  $t'$  相等, 我们将它们链接(常数时间)成一棵较大的树:  $t'_{i+1} = (2s, t_i, t')$ , 然后递归地将  $t'_{i+1}$  插入到森林中。如图 12.2 所示。

$$\begin{aligned} \text{insert}_T t [] &= [t] \\ \text{insert}_T t (t_1:ts) &= \begin{cases} \text{size } t < \text{size } t_1 : t:t_1:ts \\ \text{否则} : & \text{insert}_T(\text{link } t \ t_1) \ ts \end{cases} \end{aligned} \quad (12.4)$$

其中  $\text{link}$  将两棵大小相同的树链接起来:  $\text{link } t_1 \ t_2 = (\text{size } t_1 + \text{size } t_2, t_1, t_2)$ 。

若森林中包含  $m$  棵树,  $m$  的大小为  $O(\lg n)$ , 头部插入的性能为  $O(\lg n)$ 。稍后我们证明分摊性能为常数时间。

对称地, 我们利用插入的逆过程实现从序列头部删除元素。如果森林中第一棵树是  $t_0$ (单叶子节点), 我们直接将  $t_0$  删除; 否则, 递归地将第一棵树拆分直到获得  $t_0$ , 然后将其删除。如图 12.3 所示。

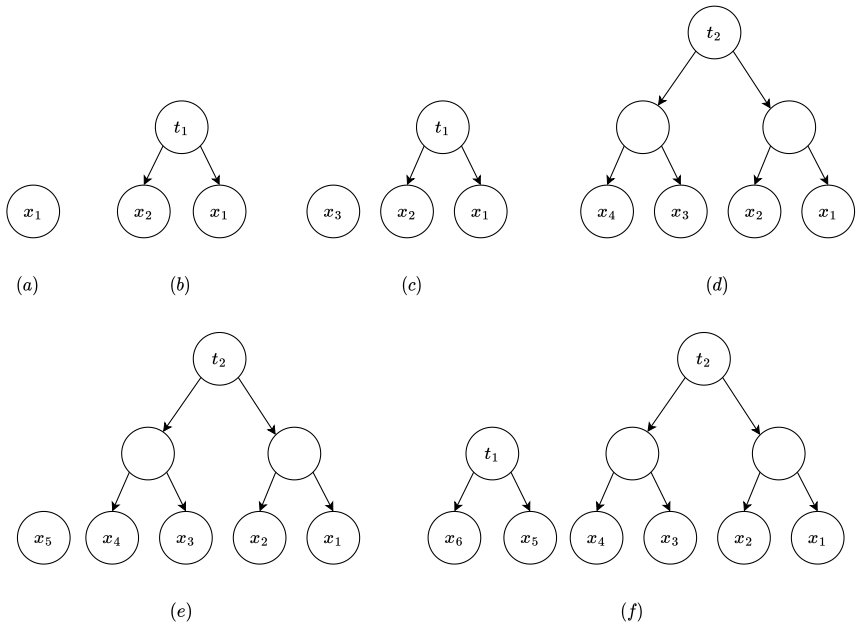


图 12.2: 插入  $x_1, x_2, \dots, x_6$ : (a) 插入  $x_1$ 。 (b) 插入  $x_2$ , 链接产生  $[t_1]$ 。 (c) 插入  $x_3$ , 结果为  $[t_0, t_1]$ 。 (d) 插入  $x_4$ 。 经两次链接后结果为  $[t_2]$ 。 (e) 插入  $x_5$ , 结果为:  $[t_0, t_2]$ 。 (f) 插入  $x_6$ , 结果为:  $[t_1, t_2]$ 。

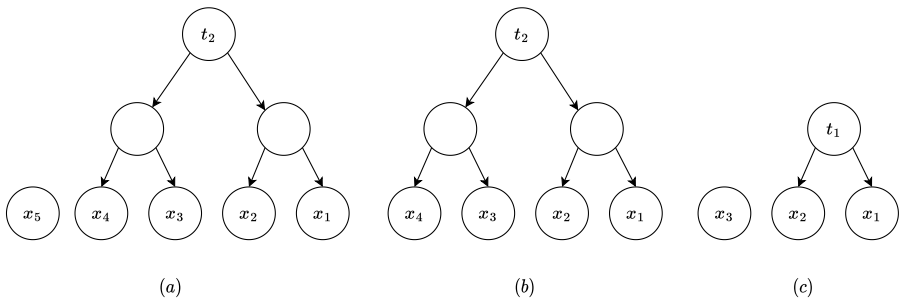


图 12.3: 从头部删除元素: (a)  $x_1, x_2, \dots, x_5$  对应  $[t_0, t_2]$ 。 (b) 删除  $x_5(t_0)$ 。 (c) 删除  $x_4$ 。 经过两次拆分后得到  $[t_0, t_0, t_1]$ , 删除后得  $[t_0, t_1]$ 。

$$\begin{aligned} \text{extract} ((x):ts) &= (x, ts) \\ \text{extract} ((s, t_1, t_2):ts) &= \text{extract} (t_1:t_2:ts) \end{aligned} \quad (12.5)$$

利用 *extract* 即可实现对头部元素的删除:

$$\begin{cases} \text{head} &= \text{fst} \circ \text{extract} \\ \text{tail} &= \text{snd} \circ \text{extract} \end{cases} \quad (12.6)$$

其中  $\text{fst}(a, b) = a$ ,  $\text{snd}(a, b) = b$  分别返回一对值中的两个部分。

森林中的树实际上将元素划分为大小不同的区块。给定任意索引  $1 \leq i \leq n$ , 我们先定位到对应的完全二叉树, 然后再进行一次树查找就可定位到元素。

1. 比较  $i$  和森林中第一棵树  $t$  的大小, 若  $i \leq \text{size}(t)$ , 则元素在  $t$  中, 接下来在树  $t$  中进行查找;
2. 否则, 令  $i' = i - \text{size}(t)$ , 然后递归地在剩余的树中查找第  $i'$  个元素。

$$(t:ts)[i] = \begin{cases} i \leq \text{size } t: & \text{lookup}_T i t \\ \text{否则}: & ts[i - \text{size } t] \end{cases} \quad (12.7)$$

其中  $\text{lookup}_T$  在树中进行二分查找。如果  $i = 1$ , 我们返回根节点; 否则, 我们将树拆半, 然后递归查找:

$$\begin{aligned} \text{lookup}_T 1(x) &= x \\ \text{lookup}_T i(s, t_1, t_2) &= \begin{cases} i \leq \lfloor \frac{s}{2} \rfloor: & \text{lookup}_T i t_1 \\ \text{否则}: & \text{lookup}_T (i - \lfloor \frac{s}{2} \rfloor) t_2 \end{cases} \end{aligned} \quad (12.8)$$

图 12.4 描述了在一个长度为 6 的序列中查找第 4 个元素的步骤。第一棵树大小为  $2 < 4$ , 继续检查第二棵树, 并将索引更新为  $i' = 4 - 2$ 。接下来的树大小为  $4 > i' = 2$ , 故待查找元素就在这棵树中。因为索引为 2, 不大于拆半的子树大小  $4/2 = 2$ , 所以接下来检查左子树, 然后检查右侧的孙子分支, 最终得到要访问的元素。类似地, 我们也可以修改任意位置  $i$  的元素。

根据完全二叉树的性质, 对于含有  $n$  个元素的序列, 树木的棵数为  $O(\lg n)$ 。对于索引  $i$ , 最多需要  $O(\lg n)$  时间来定位到树。接下来的搜索和树的高度成正比, 最多也是  $O(\lg n)$ 。因此随机访问的总体性能为  $O(\lg n)$ 。

## 练习 12.1

### 12.1.1. 如何处理索引越界情况?

答案: 384 页



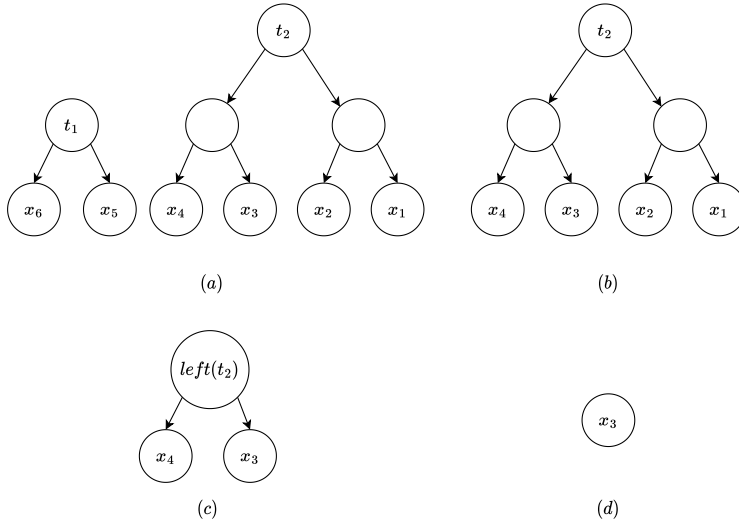


图 12.4: 获取  $S[4]$ : (a)  $S[4], 4 > size(t_1) = 2$  (b)  $S'[4 - 2] \Rightarrow lookup_T 2 t_2$  (c)  $2 \leq \lfloor \frac{size(t_2)}{2} \rfloor \Rightarrow lookup_T 2 left(t_2)$  (d)  $lookup_T 1 right(left(t_2))$ , 返回  $x_3$ 。

## 12.2 数字表示

非负整数  $n$  的二进制形式和森林之间存在关系:  $n = 2^0 e_0 + 2^1 e_1 + \dots + 2^m e_m$ , 其中  $e_i$  为第  $i$  位的值。若  $e_i = 1$ , 则存在一棵大小为  $2^i$  的完全二叉树。向序列头部插入元素, 对应于二进制数加 1; 而删除对应二进制数减 1。我们称这种关系为**数字表示**<sup>[3]</sup>。为了明确表示这种对应, 我们为每个二进制位定义两个状态: 状态零 *Zero* 表示不存在二叉树, 而状态一 *One t* 表示存在二叉树  $t$ 。这样森林就可以表示为一组二进制状态的列表, 从而把插入实现为二进制数的增加。

$$\begin{aligned}
 add\ t\ [] &= [One\ t] \\
 add\ t\ (Zero:ds) &= (One\ t):ds \\
 add\ t\ ((One\ t'):ds) &= Zero : add\ (link\ t\ t')\ ds
 \end{aligned}
 \tag{12.9}$$

将树  $t$  插入森林对应二进制加法: 若森林为空, 我们创建状态 *One t*, 它是二进制数中的唯一位。相当于  $0 + 1 = 1$ 。若森林不空, 如果二进制首位数字是 *Zero*, 我们创建一个状态 *One t* 替换掉 *Zero*。这相当于二进制加法  $(\dots digits \dots 0)_2 + 1 = (\dots digits \dots 1)_2$ 。例如  $6 + 1 = (110)_2 + 1 = (111)_2 = 7$ 。如果二进制首位是 *One t'*, 我们认为  $t$  和  $t'$  的大小相同。这是因为我们总是以一个叶子  $t_0 = (x)$  开始插入, 待插入树的大小逐渐增长, 呈一个序列  $1, 2, 4, \dots, 2^i, \dots$ 。我们将  $t$  和  $t'$  链接起来, 递归地插入到剩余的数字中。而之前的 *One t'* 被替换为 *Zero*。这相当于二进制加法  $(\dots digits \dots 1)_2 + 1 = (\dots digits' \dots 0)_2$ 。例如  $7 + 1 = (111)_2 + 1 = (1000)_2 = 8$ 。

接下来我们用二进制减法来表示删除。如果序列只含有一位 *One t*, 删除后序列变为空。这对应二进制减法  $1 - 1 = 0$ 。如果序列有多位, 并且首位是 *One t*, 我

们将其替换为 *Zero*。这相当于二进制减法  $(\dots \text{digits} \dots 1)_2 - 1 = (\dots \text{digits} \dots 0)_2$ 。例如  $7 - 1 = (111)_2 - 1 = (110)_2 = 6$ 。如果首位是 *Zero*，减法需要借位。我们递归地从剩余的数中抽取树，将其分拆成两棵树  $t_1, t_2$ ，将 *Zero* 替换成 *One*  $t_2$ ，并删除  $t_1$ 。这相当于二进制减法  $(\dots \text{digits} \dots 0)_2 - 1 = (\dots \text{digits}' \dots 1)_2$ 。例如  $4 - 1 = (100)_2 - 1 = (11)_2 = 3$ 。

$$\begin{aligned} \text{minus} [\text{One } t] &= (t, []) \\ \text{minus} ((\text{One } t):ts) &= (t, \text{Zero}:ts) \\ \text{minus} (\text{Zero}:ts) &= (t_1, (\text{One } t_2):ts'), \text{其中 } : (s, t_1, t_2) = \text{minus } ts \end{aligned} \quad (12.10)$$

数字表示并没有改变复杂度。我们使用聚合法，分析插入的分摊复杂度。考虑依次向空序列插入  $n = 2^m$  个元素的过程。森林的二进制表示如表 12.1：

i	二进制 (高 ... 低)
0	0, 0, ..., 0, 0
1	0, 0, ..., 0, 1
2	0, 0, ..., 1, 0
3	0, 0, ..., 1, 1
...	...
$2^m - 1$	1, 1, ..., 1, 1
$2^m$	1, 0, 0, ..., 0, 0
位变化次数	1, 1, 2, ..., $2^{m-1}, 2^m$

表 12.1: 插入  $2^m$  个元素的过程

二进制表示的最低位每次插入时都变化，总共需要  $2^m$  次计算；次低位每隔一次变化，执行一次树的链接操作。共需要  $2^{m-1}$  次计算；次高位总共只变化一次，将所有的树链接成一棵大树。最后一个元素插入后，最高位变为 1。将所有计算次数相加，得到  $T = 1 + 1 + 2 + 4 + \dots + 2^{m-1} + 2^m = 2^{m+1}$ 。平均每次插入操作的分摊复杂度为：

$$O(T/n) = O\left(\frac{2^{m+1}}{2^m}\right) = O(1) \quad (12.11)$$

因此插入的分摊复杂度为常数时间。

## 练习 12.2

- 12.2.1. 实现数值表示序列的随机访问  $S[i], 1 \leq i \leq n$ 。其中  $n$  是序列长度。
- 12.2.2. 分析数值表示序列删除操作的分摊复杂度。
- 12.2.3. 可以用长度为  $2^m$  的数组表示完全二叉树 ( $m$  是非负整数)。请用数组实现二叉树森林的插入和随机访问。

## 12.3 双数组序列

我们把 11.3 中的双数组队列扩展为双数组序列。如图 12.5, 按头对头的方式连接两个数组。在列表的头部插入时, 添加到  $f$  数组末尾; 向尾部插入时, 添加到  $r$  数组末尾。用一对数组表示列表  $S = (f, r)$ , 令  $\text{FRONT}(S) = f, \text{REAR}(S) = r$ 。前后插入实现如下:

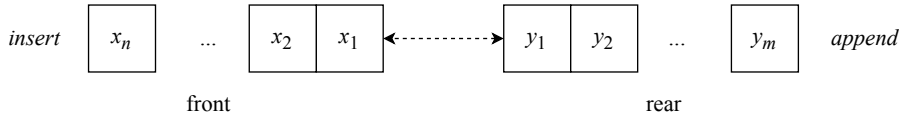


图 12.5: 双数组序列

```

1: function INSERT( $x, S$ )
2:   APPEND( $x, \text{FRONT}(S)$ )
3: function APPEND( $x, S$ )
4:   APPEND( $x, \text{REAR}(S)$ )

```

随机访问第  $i$  个元素时, 我们先判断  $i$  索引到  $f$  还是  $r$ , 然后再定位到元素。若  $i \leq |f|$ , 元素在  $f$  中。由于  $f$  和  $r$  头对头连接的, 所以  $f$  按照从右向左逆序索引, 用  $|f| - i + 1$  定位到元素; 如果  $i > |f|$ , 元素在  $r$  中, 用  $i - |f|$  定位到元素。

```

1: function GET( $i, S$ )
2:    $f, r \leftarrow \text{FRONT}(S), \text{REAR}(S)$ 
3:    $n \leftarrow \text{SIZE}(f)$ 
4:   if  $i \leq n$  then
5:     return  $f[n - i + 1]$ 
6:   else
7:     return  $r[i - n]$ 

```

▷ 反向索引

删除可能把数组  $f$  或  $r$  变空, 而另一个仍有元素, 需要恢复平衡。当  $f$  或  $r$  等于  $[\ ]$  时, 我们将另一数组分成两半, 然后将前半反转形成一对新的数组。  $f, r$  是对称的。我们也可以交换  $f, r$ , 递归调用 BALANCE, 再把  $f, r$  交换回来。

```

1: function BALANCE( $S$ )
2:    $f \leftarrow \text{FRONT}(S), r \leftarrow \text{REAR}(S)$ 
3:    $n \leftarrow \text{SIZE}(f), m \leftarrow \text{SIZE}(r)$ 
4:   if  $F = [\ ]$  then
5:      $k \leftarrow \lfloor \frac{m}{2} \rfloor$ 
6:     return (REVERSE( $r[1..k]$ ),  $r[(k + 1)..m]$ )
7:   if  $R = [\ ]$  then
8:      $k \leftarrow \lfloor \frac{n}{2} \rfloor$ 

```

```

9:     return (f[(k + 1)...n], REVERSE(f[1...k]))
10:    return (f, r)

```

在每次删除时,我们都检查  $f, r$  是否为空,并触发平衡操作:

```

1: function REMOVE-HEAD( $S$ )
2:   BALANCE( $S$ )
3:    $f, r \leftarrow$  FRONT( $S$ ), REAR( $S$ )
4:   if  $f = []$  then                                     ▷  $S = ([], [x])$ 
5:      $r \leftarrow []$ 
6:   else
7:     REMOVE-LAST( $f$ )

8: function REMOVE-TAIL( $S$ )
9:   BALANCE( $S$ )
10:   $f, r \leftarrow$  FRONT( $S$ ), REAR( $S$ )
11:  if  $r = []$  then                                     ▷  $S = ([x], [])$ 
12:     $f \leftarrow []$ 
13:  else
14:    REMOVE-LAST( $r$ )

```

由于要进行反转,双数组序列在最坏情况下性能为  $O(n)$ ,其中  $n$  是元素个数。但是分摊复杂度是常数时间的。

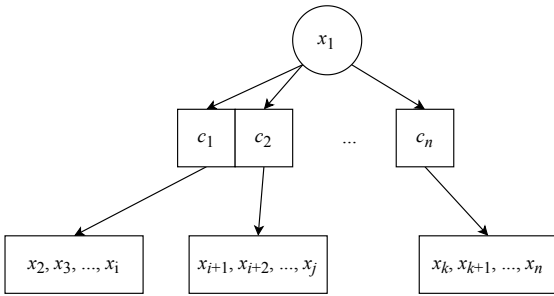
### 练习 12.3

12.3.1. 试分析双数组序列删除的分摊复杂度。

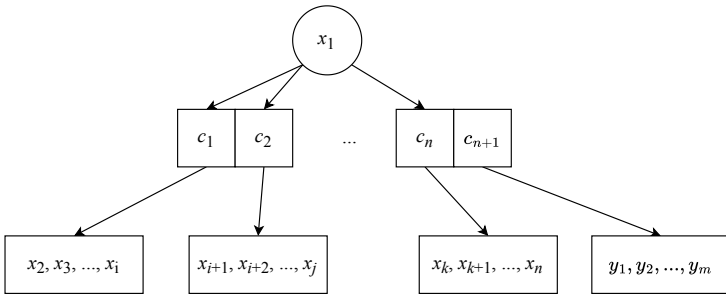
答案: 386 页

## 12.4 可连接列表

虽然可以用  $O(\lg n)$  时间在二叉树随机访问森林的头部进行插入、删除、索引,但连接两个序列并不容易。我们不能简单地将所有二叉树合并到一起,而需要不断链接大小相同的树。图 12.6 给出了一种可连接列表的实现。多叉树的根存储序列表的第一个元素  $x_1$ , 其它元素被分成若干片段保存在更小的序列中,每个片段是一棵子树。这些子树由一个实时队列(见 11.5)管理。我们把序列表示为  $(x_1, Q_x) = [x_1, x_2, \dots, x_n]$ 。当需要连接另一个列表  $(y_1, Q_y) = [y_1, y_2, \dots, y_m]$  时,我们将其入队到  $Q_x$  的尾部。连接运算定义如下。实时队列的入队性能为常数时间,因此列表连接的性能也是常数时间的。



(a)



(b)

图 12.6: 可连接列表:(a)  $(x_1, Q_x) = [x_1, x_2, \dots, x_n]$ , (b) 连接  $(y_1, Q_y) = [y_1, y_2, \dots, y_m]$  相当于入队  $c_{n+1}$  到  $Q_x$ 。

$$\begin{aligned}
 s \# \emptyset &= s \\
 \emptyset \# s &= s \\
 (x, Q) \# s &= (x, \text{push } s \ Q)
 \end{aligned}
 \tag{12.12}$$

插入新元素  $z$  时,我们先创建一个单元素列表  $(z, \emptyset)$ ,然后将其连接起来。

$$\begin{cases}
 \text{insert } z \ s &= (z, \emptyset) \# s \\
 \text{append } z \ s &= s \# (z, \emptyset)
 \end{cases}
 \tag{12.13}$$

从可连接列表的头部删除元素需要单独的设计。 $x_1$  为根节点,删除后剩下的所有子树也都是由多叉树表示的可连接列表。我们可以把它们全部连接到一起,形成一个新列表。

$$\begin{aligned}
 \text{concat } \emptyset &= \emptyset \\
 \text{concat } Q &= (\text{top } Q) \# \text{concat } (\text{pop } Q)
 \end{aligned}
 \tag{12.14}$$

全部子树存储在实时队列中。我们将第一棵子树  $c_1$  出队,然后递归地将剩余的子树连接在一起成为  $s$ ,最后把  $c_1$ 、 $s$  连接起来。我们使用  $\text{concat}$  从头部删除元素。

$$\text{tail } (x, Q) = \text{concat } Q
 \tag{12.15}$$

算法  $\text{concat}$  遍历了队列,逐步归并到一个最终的结果。这本质上相当于对  $Q$  进行叠加操作<sup>[10]</sup>。

$$\begin{aligned}
 \text{fold } f \ z \ \emptyset &= z \\
 \text{fold } f \ z \ Q &= f(\text{top } Q) (\text{fold } f \ z \ (\text{pop } Q))
 \end{aligned}
 \tag{12.16}$$

其中  $f$  是用于归并的二元函数, $z$  是零元。下面是一些队列叠加的例子,令  $Q = \{1, 2, \dots, 5\}$ 。

$$\begin{aligned}
 \text{fold } (+) \ 0 \ Q &= 1 + (2 + (3 + (4 + (5 + 0)))) = 15 \\
 \text{fold } (\times) \ 1 \ Q &= 1 \times (2 \times (3 \times (4 \times (5 \times 1)))) = 120 \\
 \text{fold } (\times) \ 0 \ Q &= 1 \times (2 \times (3 \times (4 \times (5 \times 0)))) = 0
 \end{aligned}$$

我们可以利用叠加来定义  $\text{concat}$ (柯里化形式):

$$\text{concat} = \text{fold } (\#) \ \emptyset
 \tag{12.17}$$

通常情况下插入、添加、删除、连接随机发生。删除操作的性能在最坏情况下是线性的。当连续向空序列添加  $n$  个元素后,立即执行一次删除。此时多叉树中的  $n - 1$  棵子树都是单元素的(只含有一个叶子节点)。 $\text{concat}$  需要  $O(n)$  时间进行归并。但其分摊复杂度是常数时间的。

## 12.5 手指树

二叉随机访问列表可以在头部用常数时间(分摊)插入、删除,以对数时间进行随机访问。但是难以向尾部添加元素、也无法进行快速连接。可连接列表能够用常数时间(分摊)进行连接,在头、尾部用常数时间插入。但不能用索引进行随机访问。这两个例子提示我们:1、需要某种方式快速访问头、尾以进行增删;2、带有递归的结构(例如树)可将随机访问转换成分而治之的搜索。手指树<sup>[66]</sup>利用了这两点来实现序列<sup>[65]</sup>。树是否平衡对搜索性能至关重要。手指树利用了 2-3 树(一种 B-树)。一棵 2-3 树包含二或三棵子树,如:  $(t_1, t_2)$  或  $(t_1, t_2, t_3)$ 。

```
data Node a = Br2 a a | Br3 a a a
```

定义手指树为下面三种情况之一:

1. 空  $\emptyset$ ;
2. 单元素叶子  $(x)$ ;
3. 包含三部分:一棵子树和左、右手指,记为  $(f, t, r)$ 。每个手指是一个列表<sup>1</sup>,至多含有 3 个元素。

```
data Tree a = Empty
            | Lf a
            | Tr [a] (Tree (Node a)) [a]
```

### 12.5.1 插入

如图 12.7 所示。(1) 为  $\emptyset$ , (2) 是插入一个元素后的结果。(3) 含有两个元素,分别在  $f$ 、 $r$  手指中。(4) 如果继续插入元素,  $f$  手指会超过 2-3 树的限制。(5) 恢复平衡,  $f$  手指中有 2 个元素,中间部分是一个含有一棵 2-3 树的叶子。这些例子可以表示为:

$\emptyset$	Empty
$(a)$	Lf a
$([b], \emptyset, [a])$	Tr [b] Empty [a]
$([d, c, b], \emptyset, [a])$	Tr [d, c, b] Empty [a]
$([f, e], (d, c, b), [a])$	Tr [f, e] Lf (Br3 d c b) [a]

注意最后一个例子中,中间部分的子树是一个叶子。手指树是递归的。除去  $f$ 、 $r$  手指的中间部分是一棵更深的手指树,定义为  $Tree (Node a)$ 。深度增加一级,就多嵌套一级。上面的例子实际描述了向手指树插入元素的过程。我们可以归纳如下。向一棵手指树  $T$  中插入  $a$  时:

<sup>1</sup>分别是英文前(front)、后(rear)的首字母。

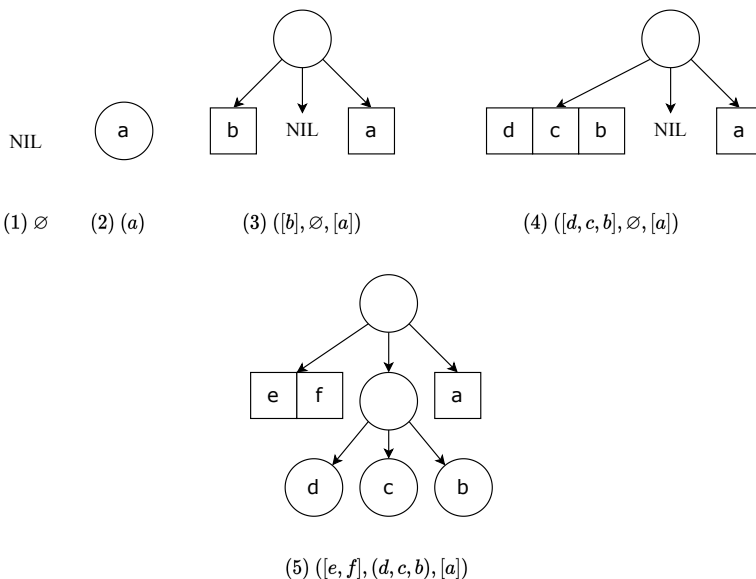


图 12.7: 手指树

1. 如果  $T = \emptyset$ , 则结果为单元素叶子  $(a)$ ;
2. 如果  $T = (b)$  是一个叶子, 结果为  $([a], \emptyset, [b])$ ;
3.  $T = (f, t, r)$ , 如果  $f$  中元素个数不足 3, 将  $a$  插入到  $f$  中; 否则将  $f$  中的后 3 个元素移入一棵新的 2-3 树  $t'$ , 递归地将  $t'$  插入到  $t$  中。最后将  $a$  插入到  $f$  中。

$$\begin{aligned}
 \text{insert } a \ \emptyset &= (x) \\
 \text{insert } a \ (b) &= ([a], \emptyset, [b]) \\
 \text{insert } a \ ([b, c, d], t, r) &= ([a], \text{insert } (b, c, d) \ t, r) \\
 \text{insert } a \ (f, t, r) &= (a:f, t, r)
 \end{aligned}
 \tag{12.18}$$

除了递归插入外, 其它情况插入都需要常数时间。递归深度取决于树的高度  $h$ , 由于使用 2-3 树并维持平衡, 因此  $h = O(\lg n)$ , 其中  $n$  是手指树中存储元素的个数。递归可以分摊到其它情况中, 插入的分摊复杂度为常数时间<sup>[3][65]</sup>。我们可以利用叠加连续将若干元素插入到树中:

$$xs \gg t = \text{foldr } \text{insert } t \ xs \tag{12.19}$$

### 练习 12.4

12.4.1. 消除递归, 用循环的方式实现手指树插入。

答案: 386 页



### 12.5.2 删除

从头部删除可以看作对 *insert* 进行逆操作。

$$\begin{aligned}
 \text{extract } (a) &= (a, \emptyset) \\
 \text{extract } ([a], \emptyset, [b]) &= (a, (b)) \\
 \text{extract } ([a], \emptyset, b:bs) &= (a, ([b], \emptyset, bs)) \\
 \text{extract } ([a], t, r) &= (a, (\text{toList } f, t', r)), \text{ 其中: } (f, t') = \text{extract } t \\
 \text{extract } (a:as, t, r) &= (a, (as, t, r))
 \end{aligned} \tag{12.20}$$

其中 *toList* 将一棵 2-3 树转换为列表:

$$\begin{aligned}
 \text{toList } (a, b) &= [a, b] \\
 \text{toList } (a, b, c) &= [a, b, c]
 \end{aligned} \tag{12.21}$$

我们略过了错误情况 (如从空树中删除)。如果手指树是单元素叶子, 结果为空树; 如果手指树只包含两个元素, 我们删除 *f* 中的元素, 结果为单元素的叶子; 如果 *f* 中只含有一个元素, 中间部分为空, 而 *r* 不空, 我们删除 *f* 中的唯一元素, 然后从 *r* 中“借”一个元素放入 *f*; 如果 *f* 只有一个元素, 而中间子树不空, 我们就递归地从子树中删除一个节点, 然后将这一节点中的内容转换成列表来代替 *f*。而原来 *f* 中的唯一元素被删除; 如果 *f* 包含一个以上的元素, 我们将第一个元素删除。图 12.8 展示了从序列头部删除两个元素的例子。使用 *extract*, 我们可以定义出 *head* 和 *tail*:

$$\begin{cases} \text{head} &= \text{fst} \circ \text{extract} \\ \text{tail} &= \text{snd} \circ \text{extract} \end{cases} \tag{12.22}$$

## 练习 12.5

12.5.1. 消除递归, 用循环实现删除。

答案: 387 页

### 12.5.3 尾部操作

可以对称地实现尾部添加、删除。

$$\begin{aligned}
 \text{append } \emptyset a &= (a) \\
 \text{append } (a) b &= ([a], \emptyset, [b]) \\
 \text{append } (f, t, [a, b, c]) d &= (f, \text{append } t (a, b, c), [d]) \\
 \text{append } (f, t, r) a &= (f, t, r \# [a])
 \end{aligned} \tag{12.23}$$

如果 *r* 中的元素不超过 3 个, 直接把新元素添加到 *r* 末尾。否则, 将 *r* 中的前三个元素取出, 构造一棵新的 2-3 树, 递归地添加到中间子树的尾部。类似地可以用左侧叠加连续将若干元素构造成一棵手指树:

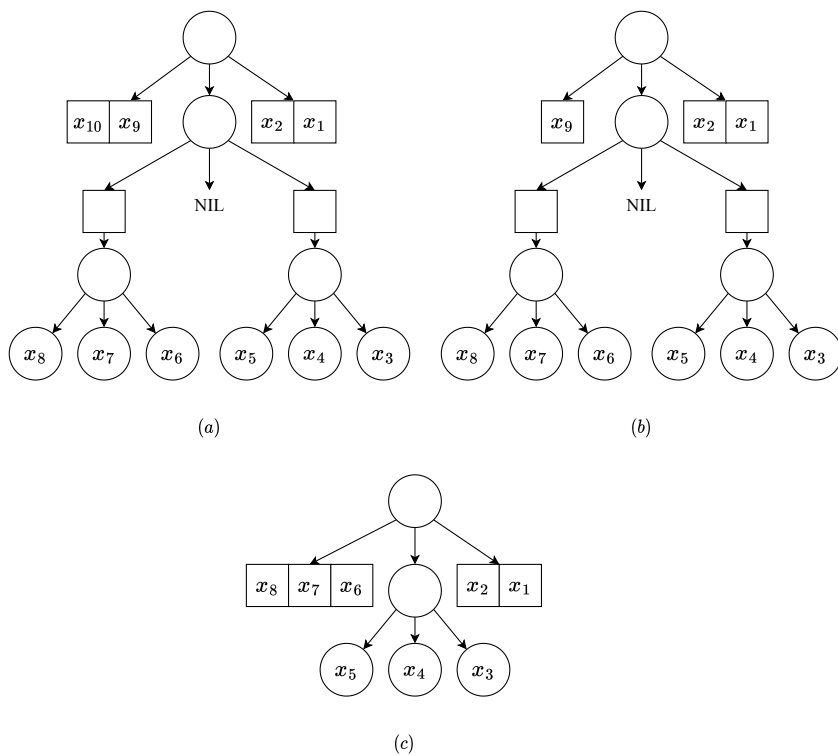


图 12.8: 删除:(a) 含有 10 个元素的树。(b) 删除一个元素后,  $f$  还剩一个元素。(c) 再次删除一个元素, 从中间子树“借”一个节点, 将它从 2-3 树转换成列表, 作为新的  $f$ 。

$$t \ll xs = foldl \text{ append } t \ xs \quad (12.24)$$

从尾部删除相当于添加的逆操作:

$$\begin{aligned} \text{remove } (a) &= (\emptyset, a) \\ \text{remove } ([a], \emptyset, [b]) &= ((a), b) \\ \text{remove } (f, \emptyset, [a]) &= ((\text{init } f, \emptyset, [\text{last } f]), a) \\ \text{remove } (f, t, [a]) &= ((f, t', \text{toList } r), a), \text{ 其中 } : (t', r) = \text{remove } t \\ \text{remove } (f, t, r) &= ((f, t, \text{init } r), \text{last } r) \end{aligned} \quad (12.25)$$

其中 *last* 获取列表的最后一个元素, *init* 返回其余部分(见式 (1.4))。

### 12.5.4 连接

$T_1 \# T_2$ 。考虑两棵手指树都不为空的情况:  $T_1 = (f_1, t_1, r_1)$ 、 $T_2 = (f_2, t_2, r_2)$ 。我们用  $f_1$  作为连接结果中的  $f$ , 用  $r_2$  作为结果中的  $r$ 。然后将  $t_1, r_1, f_2, t_2$  合并成中间子树。由于  $r_1$  和  $f_2$  都是节点的列表, 所以这等价于如下问题:

$$\text{merge } t_1 (r_1 \# f_2) t_2 = ?$$

$t_1$  和  $t_2$  也都是手指树, 但它们比  $T_1$  和  $T_2$  深一级, 若  $T_1$  中的元素类型为  $a$ , 则  $t_1$  中的元素类型为 *Node*  $a$ 。我们递归地进行合并: 保留  $t_1$  的  $f$  手指和  $t_2$  的  $r$  手指, 然后将  $t_1$  和  $t_2$  的中间部分,  $t_1$  的  $r$  手指和  $t_2$  的  $f$  手指合并。

$$\begin{aligned} \text{merge } \emptyset \ ts \ t_2 &= ts \gg t_2 \\ \text{merge } t_1 \ ts \ \emptyset &= t_1 \ll ts \\ \text{merge } (a) \ ts \ t_2 &= \text{merge } \emptyset (a:ts) \ t_2 \\ \text{merge } t_1 \ ts \ (a) &= \text{merge } t_1 (ts \# [a]) \ \emptyset \\ \text{merge } (f_1, t_1, r_1) \ ts \ (f_2, t_2, r_2) &= (f_1, \text{merge } t_1 (\text{nodes } (r_1 \# ts \# f_2)) \ t_2, r_2) \end{aligned} \quad (12.26)$$

其中 *nodes* 将若干元素组织成一组 2-3 树。这是因为中间子树中的元素类型, 比手指中的元素类深一级。

$$\begin{aligned} \text{nodes } [a, b] &= [(a, b)] \\ \text{nodes } [a, b, c] &= [(a, b, c)] \\ \text{nodes } [a, b, c, d] &= [(a, b), (c, d)] \\ \text{nodes } (a:b:c:ts) &= (a, b, c): \text{nodes } ts \end{aligned} \quad (12.27)$$

这样我们可以用 *merge* 来定义手指树的连接:

$$(f_1, t_1, r_1) \# (f_2, t_2, r_2) = (f_1, \text{merge } t_1 (r_1 \# f_2) \ t_2, r_2) \quad (12.28)$$

比较这一定义和式 (12.26), 连接操作本质上就是合并操作, 我们可以给出下面更加一致的定义:

$$T_1 \uplus T_2 = \text{merge } T_1 \ [] T_2 \quad (12.29)$$

连接的性能取决于递归的合并操作。递归的深度为两棵树中较小的一棵。由于 2-3 树的平衡性, 手指树的高度为  $O(\lg n)$  其中  $n$  为元素的个数。合并在边界条件下的性能和插入一样(最多调用 *insert* 8 次)为分摊常数时间, 最坏情况为  $O(m)$ , 其中  $m$  是两棵树的高度差。总体上算法的复杂度为  $O(\lg n)$ , 其中  $n$  是两棵手指树中含有的元素总数。

### 12.5.5 随机访问

我们的策略是把随机访问转换为树搜索。为了避免反复计算树的大小, 我们给每个分枝节点增加一个  $s$  变量记录其包含的元素个数:  $(s, f, t, r)$ 。

```
data Tree a = Empty
  | Lf a
  | Tr Int [a] (Tree (Node a)) [a]
```

$$\begin{aligned} \text{size } \emptyset &= 0 \\ \text{size } (x) &= \text{size } x \\ \text{size } (s, f, t, r) &= s \end{aligned} \quad (12.30)$$

这里  $\text{size } (x)$  并不一定是 1。这是因为  $x$  可能是更深的节点, 例如 *Node a*, 而只有当第一层时大小才是 1。为此我们可以在插入或添加时, 把每个元素  $x$  包装在一个单元里  $(x)_e$ , 规定这个单元的大小为 1, 即:  $\text{size } (x)_e = 1$  (参见附录例子)。

$$\begin{cases} x \triangleleft t = \text{insert } (x)_e t \\ t \triangleright x = \text{append } t (x)_e \end{cases} \quad (12.31)$$

以及:

$$\begin{cases} xs \ll t = \text{foldr } (\triangleleft) t xs \\ t \gg xs = \text{foldl } (\triangleright) t xs \end{cases} \quad (12.32)$$

我们还需要获取 2-3 树的大小:

$$\begin{aligned} \text{size } (t_1, t_2) &= \text{size } t_1 + \text{size } t_2 \\ \text{size } (t_1, t_2, t_3) &= \text{size } t_1 + \text{size } t_2 + \text{size } t_3 \end{aligned} \quad (12.33)$$

对于节点的列表(例如深一级的手指)我们可以用  $\text{sum} \circ (\text{map } \text{size})$  来计算大小。在插入和删除操作中, 我们也需要更新树大小。增加大小信息后, 给定一个位置  $i$ , 可以通过树搜索定位到相应的节点。手指树具有递归结构:  $(s, f, t, r)$ , 我们令这些子结构的大小为:  $s_f, s_t, s_r$ , 且  $s = s_f + s_t + s_r$ 。如果  $i \leq s_f$ , 则目标位于  $f$  中, 我们接下来在  $f$  中查找; 如果  $s_f < i \leq s_f + s_t$ , 则目标位于  $t$  中, 我们递归在  $t$  中搜索; 否则目标位于  $r$  中。除此之外, 我们还需要处理叶子节点  $(x)$  的情况。我们用一对值  $(i, t)$  表示在数据结构  $t$  中  $i$  的位置, 并定义查找操作  $\text{lookup}_T$  如下:

$$\begin{aligned}
 \text{lookup}_T i(x) &= (i, x) \\
 \text{lookup}_T i(s, f, t, r) &= \begin{cases} i < s_f : & \text{lookup}_s i f \\ s_f \leq i < s_f + s_t : & \text{lookup}_N(\text{lookup}_T(i - s_f) t) \\ \text{否则} : & \text{lookup}_s(i - s_f - s_t) r \end{cases}
 \end{aligned} \tag{12.34}$$

这里:  $s_f = \text{sum}(\text{map size } f)$ ,  $s_t = \text{size } t$ , 分别是手指树前两部分的大小。如果在叶子节点  $(x)$  中查找位于  $i$  的内容, 结果为  $(i, x)$ 。否则我们判断  $i$  位于  $(s, f, t, r)$  中哪一部分。如果位于前后手指  $f, r$  中, 我们依次查找手指列表中的每个元素。

$$\text{lookup}_s i(x:xs) = \begin{cases} i < \text{size } x : & (i, x) \\ \text{否则} : & \text{lookup}_s(i - \text{size } x) xs \end{cases} \tag{12.35}$$

如果  $i$  位于某个元素  $x$  中 ( $i < \text{size } x$ ), 我们返回  $(i, x)$ , 否则我们继续查找后面的元素。如果  $i$  不位于前后手指  $f, r$ , 而在中间部分  $t$ , 我们递归地在更深层查找, 得到位置  $(i', m)$ 。这里  $m$  是一个 2-3 树, 我们接下来在其中查找:

$$\begin{aligned}
 \text{lookup}_N i(t_1, t_2) &= \begin{cases} i < \text{size } t_1 : & (i, t_1) \\ \text{否则} : & (i - \text{size } t_1, t_2) \end{cases} \\
 \text{lookup}_N i(t_1, t_2, t_3) &= \begin{cases} i < \text{size } t_1 : & (i, t_1) \\ \text{size } t_1 \leq i < \text{size } t_1 + \text{size } t_2 : & (i - \text{size } t_1, t_2) \\ \text{否则} : & (i - \text{size } t_1 - \text{size } t_2, t_3) \end{cases}
 \end{aligned} \tag{12.36}$$

我们此前为了计算大小把每个元素  $x$  都封装在  $(x)_e$  中, 最终我们需要从中把  $x$  取回:

$$T[i] = \begin{cases} \text{若 } \text{lookup}_T i T = (i', (x)_e) : & \text{Just } x \\ \text{否则} : & \text{Nothing} \end{cases} \tag{12.37}$$

我们利用了类型  $\text{Maybe } a = \text{Nothing} \mid \text{Just } a$  来表示索引成功或没有找到<sup>2</sup>。随机访问需要递归在手指树中查找, 递归次数取决于树的深度。由于手指树是平衡的, 随机访问的复杂度为  $O(\lg n)$ , 其中  $n$  是存储的元素个数。

我们用手指树实现的序列在总体上有着均衡、良好的性能。头、尾操作的分摊复杂度为常数时间, 可以在对数时间内进行连接、分割、随机索引<sup>[67]</sup>。到本章为止, 我们介绍了最基本的数据结构。接下来可以使用它们解决一些典型问题。例如, 我们可以用序列实现 MTF<sup>3</sup> 编码算法<sup>[68]</sup>。MTF 把序列中任意位置  $i$  的元素移动到最前面(见练习 12.6.2):

$$\text{mtf } i S = x \triangleleft S', \text{ 其中 } (x, S') = \text{extractAt } i S$$

<sup>2</sup>很多编程环境提供了类似的处理, 例如 Java/C++ 中的 `Optional<T>` 类型

<sup>3</sup>英文 `move to front` 的缩写。它应用于 BWT (Burrows-Wheeler transform) 数据压缩算法。

在后面章节中,我们将介绍基本的分而治之的排序算法,包括快速排序、归并排序以及它们的变形;然后我们介绍字符串匹配算法和基本搜索算法。

## 练习 12.6

12.6.1. 在随机访问时,如何处理空树  $\emptyset$  和索引越界的情况?

12.6.2. 实现  $cut\ i\ S$ , 在位置  $i$  把序列  $S$  分割开。

答案: 389 页

## 12.6 附录:例子程序

随机访问列表(森林):

```

data Tree a = Leaf a
             | Node Int (Tree a) (Tree a)

type BRAList a = [Tree a]

size (Leaf _) = 1
size (Node sz _ _) = sz

link t1 t2 = Node (size t1 + size t2) t1 t2

insert x = insertTree (Leaf x) where
  insertTree t [] = [t]
  insertTree t (t':ts) = if size t < size t' then t:t':ts
                       else insertTree (link t t') ts

extract ((Leaf x):ts) = (x, ts)
extract ((Node _ t1 t2):ts) = extract (t1:t2:ts)

head' = fst ◦ extract
tail' = snd ◦ extract

getAt i (t:ts) | i < size t = lookupTree i t
               | otherwise = getAt (i - size t) ts
where
  lookupTree 0 (Leaf x) = x
  lookupTree i (Node sz t1 t2)
    | i < sz `div` 2 = lookupTree i t1
    | otherwise = lookupTree (i - sz `div` 2) t2

```

随机访问森林的数值表示:

```

data Digit a = Zero | One (Tree a)

type RAList a = [Digit a]

insert x = add (Leaf x) where
  add t [] = [One t]

```

```

add t (Zero:ts) = One t : ts
add t (One t' :ts) = Zero : add (link t t') ts

minus [One t] = (t, [])
minus (One t:ts) = (t, Zero:ts)
minus (Zero:ts) = (t1, One t2:ts') where
    (Node _ t1 t2, ts') = minus ts

head' ts = x where (Leaf x, _) = minus ts
tail' = snd ◦ minus

```

双数组序列:

```

Data Seq<K> {
    [K] front = [], rear = []
}

Int length(S<K> s) = length(s.front) + length(s.rear)

void insert(K x, Seq<K> s) = append(x, s.front)

void append(K x, Seq<K> s) = append(x, s.rear)

K get(Int i, Seq<K> s) {
    Int n = length(s.front)
    return if i < n then s.front[n - i - 1] else s.rear[i - n]
}

```

可连接列表:

```

data CList a = Empty | CList a (Queue (CList a))

wrap x = CList x emptyQ

x #+ Empty = x
Empty #+ y = y
(CList x q) #+ y = CList x (push q y)

fold f z q | isEmpty q = z
            | otherwise = (top q) `f` fold f z (pop q)

concat = fold (#+) Empty

insert x xs = (wrap x) #+ xs
append xs x = xs #+ wrap x

head (CList x _) = x
tail (CList _ q) = concat q

```

手指树:

```

— 2-3 树
data Node a = Tr2 Int a a
              | Tr3 Int a a a

```

— 手指树

```
data Tree a = Empty
  | Lf a
  | Br Int [a] (Tree (Node a)) [a] — size, front, mid, rear
```

```
newtype Elem a = Elem { getElem :: a } — 封装元素
```

```
newtype Seq a = Seq (Tree (Elem a)) — 序列
```

```
class Sized a where — 可计算大小
  size :: a → Int
```

```
instance Sized (Elem a) where
  size _ = 1 — 元素的大小总为 1
```

```
instance Sized (Node a) where
  size (Tr2 s _ _) = s
  size (Tr3 s _ _ _) = s
```

```
instance Sized a ⇒ Sized (Tree a) where
  size Empty = 0
  size (Lf a) = size a
  size (Br s _ _ _) = s
```

```
instance Sized (Seq a) where
  size (Seq xs) = size xs
```

```
tr2 a b = Tr2 (size a + size b) a b
tr3 a b c = Tr3 (size a + size b + size c) a b c
```

```
nodesOf (Tr2 _ a b) = [a, b]
nodesOf (Tr3 _ a b c) = [a, b, c]
```

— 左侧操作

```
x <| Seq xs = Seq (Elem x `cons` xs)
```

```
cons :: (Sized a) ⇒ a → Tree a → Tree a
cons a Empty = Lf a
cons a (Lf b) = Br (size a + size b) [a] Empty [b]
cons a (Br s [b, c, d] m r) = Br (s + size a) [a] ((tr3 b c d) `cons` m) r
cons a (Br s f m r) = Br (s + size a) (a:f) m r
```

```
head' (Seq xs) = getElem $ fst $ uncons xs
```

```
tail' (Seq xs) = Seq $ snd $ uncons xs
```

```
uncons :: (Sized a) ⇒ Tree a → (a, Tree a)
uncons (Lf a) = (a, Empty)
uncons (Br _ [a] Empty [b]) = (a, Lf b)
uncons (Br s [a] Empty (r:rs)) = (a, Br (s - size a) [r] Empty rs)
uncons (Br s [a] m r) = (a, Br (s - size a) (nodesOf f) m' r)
  where (f, m') = uncons m
uncons (Br s (a:f) m r) = (a, Br (s - size a) f m r)
```



— 右侧操作

```
Seq xs |> x = Seq (xs `snoc` Elem x)
```

```
snoc :: (Sized a) => Tree a -> a -> Tree a
```

```
snoc Empty a = Lf a
```

```
snoc (Lf a) b = Br (size a + size b) [a] Empty [b]
```

```
snoc (Br s f m [a, b, c]) d = Br (s + size d) f (m `snoc` (tr3 a b c)) [d]
```

```
snoc (Br s f m r) a = Br (s + size a) f m (r # [a])
```

```
last' (Seq xs) = getElem $ snd $ unsnoc xs
```

```
init' (Seq xs) = Seq $ fst $ unsnoc xs
```

```
unsnoc :: (Sized a) => Tree a -> (Tree a, a)
```

```
unsnoc (Lf a) = (Empty, a)
```

```
unsnoc (Br _ [a] Empty [b]) = (Lf a, b)
```

```
unsnoc (Br s f@(_:_:_) Empty [a]) = (Br (s - size a) (init f) Empty [last f], a)
```

```
unsnoc (Br s f m [a]) = (Br (s - size a) f m' (nodesOf r), a)
```

```
    where (m', r) = unsnoc m
```

```
unsnoc (Br s f m r) = (Br (s - size a) f m (init r), a) where a = last r
```

— 连接

```
Seq xs ## Seq ys = Seq (xs >< ys)
```

```
xs >< ys = merge xs [] ys
```

```
t <<< xs = foldl snoc t xs
```

```
xs >>> t = foldr cons t xs
```

```
merge :: (Sized a) => Tree a -> [a] -> Tree a -> Tree a
```

```
merge Empty es t2 = es >>> t2
```

```
merge t1 es Empty = t1 <<< es
```

```
merge (Lf a) es t2 = merge Empty (a:es) t2
```

```
merge t1 es (Lf a) = merge t1 (es#[a]) Empty
```

```
merge (Br s1 f1 m1 r1) es (Br s2 f2 m2 r2) =
```

```
    Br (s1 + s2 + (sum $ map size es)) f1 (merge m1 (trees (r1 # es # f2)) m2) r2
```

```
trees [a, b] = [tr2 a b]
```

```
trees [a, b, c] = [tr3 a b c]
```

```
trees [a, b, c, d] = [tr2 a b, tr2 c d]
```

```
trees (a:b:c:es) = (tr3 a b c):trees es
```

— 索引

```
data Place a = Place Int a
```

```
getAt :: Seq a -> Int -> Maybe a
```

```
getAt (Seq xs) i | i < size xs = case lookupTree i xs of
```

```
    Place _ (Elem x) -> Just x
```

```
    | otherwise = Nothing
```

```
lookupTree :: (Sized a) => Int -> Tree a -> Place a
```

```
lookupTree n (Lf a) = Place n a
```

```
lookupTree n (Br s f m r) | n < sf = lookups n f
```

```

        | n < sm = case lookupTree (n - sf) m of
            Place n' xs → lookupNode n' xs
        | n < s = lookups (n - sm) r
where sf = sum $ map size f
        sm = sf + size m

lookupNode :: (Sized a) ⇒ Int → Node a → Place a
lookupNode n (Tr2 _ a b) | n < sa = Place n a
                        | otherwise = Place (n - sa) b
    where sa = size a

lookupNode n (Tr3 _ a b c) | n < sa = Place n a
                        | n < sab = Place (n - sa) b
                        | otherwise = Place (n - sab) c
    where sa = size a
          sab = sa + size b

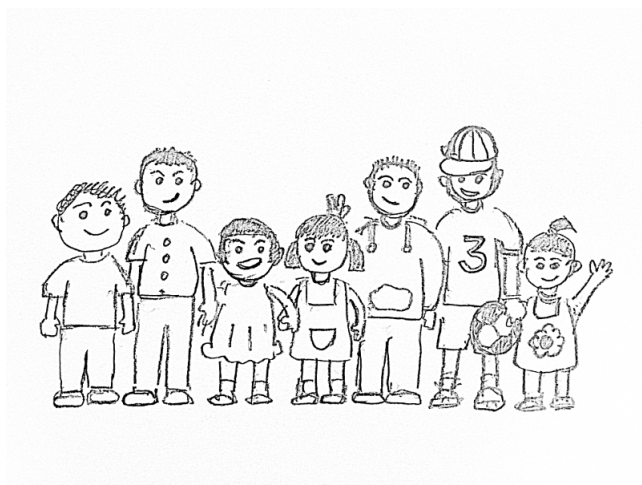
lookups :: (Sized a) ⇒ Int → [a] → Place a
lookups n (x:xs) = if n < sx then Place n x
                else lookups (n - sx) xs
    where sx = size x

```

## 第十三章 快速排序和归并排序

本章介绍两种分治排序算法:快速排序和归并排序,性能均可达到  $O(n \lg n)$ ,这是基于比较的排序算法性能上限<sup>[51]</sup>。我们还会介绍它们的若干变形,如自然归并排序,原地归并排序等。

### 13.1 快速排序



考虑安排一组小朋友按照个头排队。

1. 第一个小朋友举起手。所有比他矮的都站到他的左侧去;所有比他高的站到右侧去;
2. 所有站到左侧、右侧的小朋友重复这一步骤。

设孩子们的身高为(厘米): $[102, 100, 98, 95, 96, 99, 101, 97]$ ,表 13.1 描述了这一排队过程。第一步,身高为 102 厘米的孩子举手,以他为基准(标以下划线)。恰巧他最高,所有人都站到他左侧,如表中第二行所示。此时,身高为 102 厘米的小朋友站到了最终应站的位置(标以引号)。第二步,身高为 100 厘米的孩子举手。身高为 98、95、96、99 厘米的小朋友站到了他的左侧,身高为 101 厘米的小朋友站到了右侧。如表中第三行。第三步,身高为 98 厘米的小朋友成为左侧基准;身高为 101 厘米的小朋友成

为右侧基准。但右侧那组只有一人,无需继续排序。重复同样的方法,直到所有人都站到最终位置。

<u>102</u>	100	98	95	96	99	101	97
<u>100</u>	98	95	96	99	101	97	'102'
<u>98</u>	95	96	99	97	'100'	101	'102'
<u>95</u>	96	97	'98'	99	'100'	'101'	'102'
'95'	<u>96</u>	97	'98'	'99'	'100'	'101'	'102'
'95'	'96'	97	'98'	'99'	'100'	'101'	'102'
'95'	'96'	'97'	'98'	'99'	'100'	'101'	'102'

表 13.1: 按身高排队过程

我们可以归纳出快速排序的定义。对序列  $L$  进行排序时:

- 若  $L$  为空  $[]$ , 则排序结果为空  $[]$ ;
- 否则, 在  $L$  中任选一个元素作为基准  $p$ , 然后递归地将  $L$  中不大于  $p$  的元素排序, 将结果置于  $p$  的左侧, 同时将所有大于  $p$  的元素排序, 结果置于右侧。

我们说“同时”而不是“然后”。左右两侧的递归排序可以并行。快速排序由霍尔 (C. A. R. Hoare) 在 1960 年提出<sup>[51]、[78]</sup>。定义中并没有说明如何选择基准。这里有多种可能, 例如总选择第一个元素为基准  $p$ :

$$\begin{aligned} \text{sort } [] &= [] \\ \text{sort } (x:xs) &= \text{sort } [y|y \leftarrow xs, y \leq x] \# [x] \# \text{sort } [y|y \leftarrow xs, x < y] \end{aligned} \quad (13.1)$$

我们使用了 ZF 表达式(见 1.3.1、1.6)划分列表, 如下面的例子代码:

```
sort [] = []
sort (x:xs) = sort [y | y ← xs, y ≤ x] # [x] # sort [y | y ← xs, x < y]
```

我们假设按递增顺序排序。可以把比较条件( $\leq$ )抽象出来, 对数字、字符串等进行各种排序。我们并不要求一定是全序, 但是至少要满足**严格弱序**<sup>[79]、[52]</sup>(见 9.2)。

### 13.1.1 划分

我们遍历了两次: 第一次获得了所有  $\leq x$  的元素, 第二次获得所有  $> x$  的元素。可以将它们合并:

$$\begin{aligned} \text{part } p [] &= ([], []) \\ \text{part } p (x:xs) &= \begin{cases} p(x): (x:as, bs), \text{ 其中 } : (as, bs) = \text{part } p xs \\ \text{否则}: (as, x:bs) \end{cases} \end{aligned} \quad (13.2)$$

这样快速排序的定义变为:

$$\begin{aligned} \text{sort } [] &= [] \\ \text{sort } (x:xs) &= \text{sort } as \ ++ \ [x] \ ++ \ \text{sort } bs, \text{ 其中 } : (as, bs) = \text{part } (\leq x) \ xs \end{aligned} \quad (13.3)$$

也可以用叠加定义划分(柯里化):

$$\text{part } p = \text{foldr } f \ ([], []) \quad (13.4)$$

其中:

$$f \ (as, bs) \ x = \begin{cases} p(x) : \ (x:as, bs) \\ \text{否则} : \ (as, x:bs) \end{cases} \quad (13.5)$$

这本质上是向划分结果  $(as, bs)$  累积的过程。若  $p(x)$  成立, 则累积  $x$  到  $as$ , 否则累积到  $bs$ 。转化为尾递归:

$$\begin{aligned} \text{part } p \ [] \ as \ bs &= (as, bs) \\ \text{part } p \ (x:xs) \ as \ bs &= \begin{cases} p(x) : \ \text{part } p \ xs \ (x:as) \ bs \\ \text{否则} : \ \text{part } p \ xs \ as \ (x:bs) \end{cases} \end{aligned} \quad (13.6)$$

对  $x:xs$  进行划分的表达式为:  $(as, bs) = \text{part } (\leq x) \ xs \ [] \ []$ 。

利用累积变量  $s$  将  $\text{sort } as \ ++ \ [x] \ ++ \ \text{sort } bs$  中的  $++$  消除:

$$\begin{aligned} \text{sort}' \ s \ [] &= s \\ \text{sort}' \ s \ (x:xs) &= \text{sort}' \ (x : \text{sort}' \ s \ bs) \ as \end{aligned} \quad (13.7)$$

传入空列表启动排序:  $\text{sort} = \text{sort}' \ []$ 。划分完成后递归地对  $as, bs$  排序。我们可以先对  $bs$  排序, 将  $x$  链接到结果前, 作为新的“累积结果”传入后续的排序中。

```

sort = sort' []

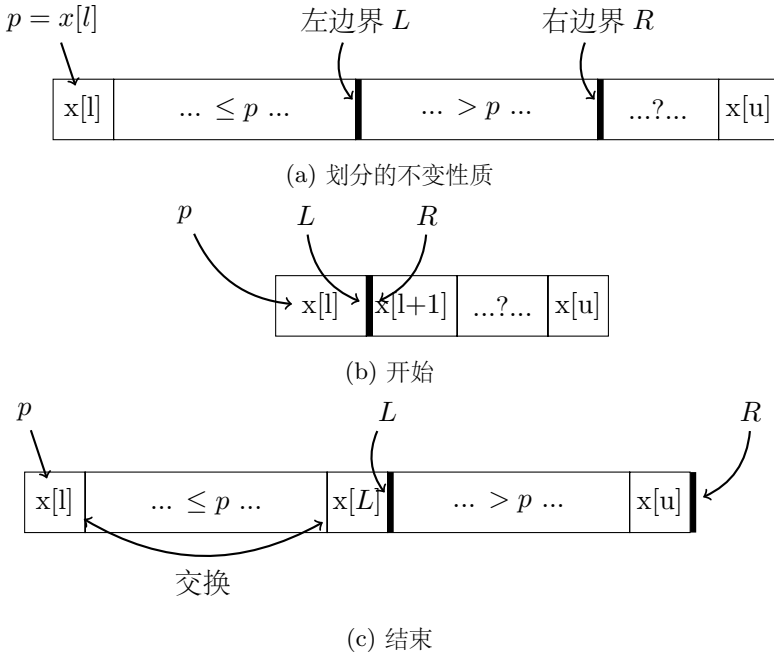
sort' acc [] = acc
sort' acc (x:xs) = sort' (x : sort' acc bs) as where
  (as, bs) = part xs [] []
  part [] as bs = (as, bs)
  part (y:ys) as bs | y ≤ x = part ys (y:as) bs
                    | otherwise = part ys as (y:bs)

```

### 13.1.2 原地排序

我们接下来考虑如何原地划分、排序。图 13.1 描述了一次遍历划分的方法<sup>[2][4]</sup>。从左向右扫描。任何时候数组都由图 13.1 (a) 所示的 4 部分组成:

1. 最左侧为基准  $p$ 。划分结束时,  $p$  被移动到最终位置;
2. 一段只包含  $\leq p$  的部分。这一段的右侧边界为  $L$ ;

图 13.1: 用首个元素作基准  $p$  划分一段数组

3. 一段只包含  $> p$  的部分。这一段的右侧边界为  $R$ 。 $L, R$  之间的元素都大于  $p$ ;
4.  $R$  后面的元素尚未处理。这部分的元素可能大于或不大于  $p$ 。

划分开始时,  $L$  指向  $p$ ,  $R$  指向  $p$  的下一个元素, 如图 13.1 (b) 所示。然后不断向右移动  $R$  进行处理, 直到  $R$  越过数组右侧。每次迭代, 都比较  $R$  指向的元素和  $p$ 。若  $x[R] > p$ , 它应位于  $L$  和  $R$  之间, 我们继续向前移动  $R$ ; 否则  $x[R] \leq p$ , 它应该位于  $L$  左侧。我们将  $L$  向前移动一步, 然后交换  $x[L]$  和  $x[R]$ 。当  $R$  越过最后一个位置时, 所有的元素都已处理完。 $> p$  的元素都被移动到了  $L$  右侧, 而其它元素位于  $L$  左侧。此时我们需要移动  $p$ , 使它位于这两段的中间。为此, 我们交换  $p$  和  $x[L]$ 。如图 13.1 (c) 中的双向箭头所示。 $L$  最终指向  $p$ , 将整个数组分成两部分。我们将  $L$  作为划分的结果返回。为了方便后继处理, 我们将  $L$  增加 1, 使得它指向第一个大于  $p$  的元素。令数组为  $A$ , 待划分区间的上下界为  $l, u$ , 原地划分实现如下:

```

1: function PARTITION( $A, l, u$ )
2:    $p \leftarrow A[l]$  ▷ 基准
3:    $L \leftarrow l$  ▷ 左侧
4:   for  $R$  in  $[l + 1, u]$  do ▷ 对右侧迭代
5:     if  $p \geq A[R]$  then
6:        $L \leftarrow L + 1$ 
7:       EXCHANGE  $A[L] \leftrightarrow A[R]$ 
8:   EXCHANGE  $A[L] \leftrightarrow p$ 

```

9: **return**  $L + 1$

▷ 返回划分的位置

表 13.2 给出了划分数组  $[3, 2, 5, 4, 0, 1, 6, 7]$  的步骤。

<u>3</u> (l)	2(r)	5	4	0	1	6	7	开始, $p = 3, l = 1, r = 2$
<u>3</u>	2(l)(r)	5	4	0	1	6	7	$2 < 3$ , 移动 $l(r = l)$
<u>3</u>	2(l)	5(r)	4	0	1	6	7	$5 > 3$ , 继续
<u>3</u>	2(l)	5	4(r)	0	1	6	7	$4 > 3$ , 继续
<u>3</u>	2(l)	5	4	0(r)	1	6	7	$0 < 3$
<u>3</u>	2	0(l)	4	5(r)	1	6	7	移动 $l$ , 然后和 $r$ 交换
<u>3</u>	2	0(l)	4	5	1(r)	6	7	$1 < 3$
<u>3</u>	2	0	1(l)	5	4(r)	6	7	移动 $l$ , 然后和 $r$ 交换
<u>3</u>	2	0	1(l)	5	4	6(r)	7	$6 > 3$ , 继续
<u>3</u>	2	0	1(l)	5	4	6	7(r)	$7 > 3$ , 继续
1	2	0	3	5(l+1)	4	6	7	$r$ 越过了边界, 交换 $p$ 和 $l$

表 13.2: 扫描并划分数组

使用 PARTITION, 可以实现快速排序如下:

```

1: procedure QUICK-SORT( $A, l, u$ )
2:   if  $l < u$  then
3:      $m \leftarrow$  PARTITION( $A, l, u$ )
4:     QUICK-SORT( $A, l, m - 1$ )
5:     QUICK-SORT( $A, m, u$ )

```

排序时传入数组上下界, 如: QUICK-SORT( $A, 1, |A|$ )。如果数组片段为空或只含有一个元素, 则直接返回。

## 练习 13.1

13.1.1. 改进基本快速排序的定义, 优化对单元素列表的处理。

答案: 391 页

### 13.1.3 性能分析

快速排序在实际应用中性能良好。最好情况下, 每次均分序列, 如图 13.2 所示, 共需要  $O(\lg n)$  次递归调用。第一层划分一次, 处理  $n$  个元素; 第二层划分两次, 每次处理  $n/2$  个元素, 总体执行时间为  $2O(n/2) = O(n)$ 。第三层划分四次, 每次处理  $n/4$  个元素, 总体执行时间也是  $O(n)$ ……最后一层总共有  $n$  个片段, 每个片段一个元素, 总时间也是  $O(n)$ 。将所有层相加得到最好情况下的性能为  $O(n \lg n)$ 。

最坏情况下, 划分极不平衡。一部分长  $O(1)$ , 另一部分长  $O(n)$ 。递归深度退化为  $O(n)$ 。最好情况下, 快速排序过程是一棵平衡二叉树; 最坏情况下, 是一棵极不平衡的

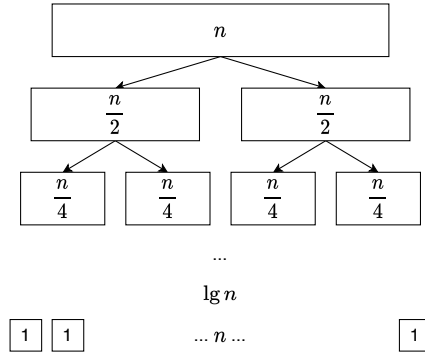


图 13.2: 最好情况, 每次均分。

树, 每个节点都只有一棵子树, 另一棵为空。二叉树退化成一条长度为  $O(n)$  的链表。而在每一层中, 所有的元素都被处理, 最坏情况下的性能为  $O(n^2)$ , 和插入排序、选择排序的性能相当。我们可以考虑几种极端情况, 例如大量元素相等、元素已序等都导致不平衡划分。我们无法彻底避免最坏情况。

### 平均情况 ★

快速排序在平均情况下性能良好。即使每次划分为 1:9 的两部分, 性能仍然为  $O(n \lg n)$ <sup>[4]</sup>。我们给出两种方法分析平均情况下的复杂度。方法一利用比较次数来考量性能<sup>[4]</sup>。在选择排序中, 任何两个元素都进行了比较。快速排序避免了很多不必要的比较。考虑列表  $[a_1, a_2, a_3, \dots, a_n]$ , 选择  $a_1$  作为基准, 划分结果为  $A = [x_1, x_2, \dots, x_k]$  和  $B = [y_1, y_2, \dots, y_{n-k-1}]$ 。此后  $A, B$  中的任何元素都不再相互比较 (任何  $x_i$  不再和  $y_j$  比较)。令最终排序结果为  $[a_1, a_2, \dots, a_n]$ , 若  $a_i < a_j$ , 当且仅当存在某一元素  $a_k$  满足  $a_i < a_k < a_j$ , 并且  $a_k$  在  $a_i$  或  $a_j$  之前被选为基准时, 不再对  $a_i$  和  $a_j$  进行比较。也就是说, 若  $a_i$  与  $a_j$  进行比较, 要么  $a_i$ 、要么  $a_j$  一定在所有  $a_{i+1} < a_{i+2} < \dots < a_{j-1}$  之前被选为基准。令  $P(i, j)$  代表  $a_i$  和  $a_j$  进行比较的概率, 我们有:

$$P(i, j) = \frac{2}{j - i + 1} \quad (13.8)$$

总共比较次数为:

$$C(n) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n P(i, j) \quad (13.9)$$

如果比较了  $a_i, a_j$ , 此后就不再比较  $a_j, a_i$ , 并且  $a_i$  从不会和自己比较。因此上式中  $i$  的上限为  $n-1$ ,  $j$  的下限为  $i+1$ 。将概率代入得:

$$\begin{aligned} C(n) &= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2}{j - i + 1} \\ &= \sum_{i=1}^{n-1} \sum_{k=1}^{n-i} \frac{2}{k + 1} \end{aligned} \quad (13.10)$$



使用调和级数<sup>[80]</sup>。

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \dots = \ln n + \gamma + \epsilon_n$$

得：

$$C(n) = \sum_{i=1}^{n-1} O(\lg n) = O(n \lg n) \quad (13.11)$$

第二种分析方法利用递归。令列表长度  $n$ ，划分后得到两个长度为  $i$  和  $n - i - 1$  的部分。划分过程比较基准  $p$  和每个元素，用时  $cn$ 。总时间为：

$$T(n) = T(i) + T(n - i - 1) + cn \quad (13.12)$$

其中  $T(n)$  是对长度为  $n$  的列表进行快速排序所用的时间。 $i$  以相同的概率在  $0, 1, \dots, n - 1$  中取值。对上式取数学期望：

$$\begin{aligned} T(n) &= E(T(i)) + E(T(n - i - 1)) + cn \\ &= \frac{1}{n} \sum_{i=0}^{n-1} T(i) + \frac{1}{n} \sum_{i=0}^{n-1} T(n - i - 1) + cn \\ &= \frac{1}{n} \sum_{i=0}^{n-1} T(i) + \frac{1}{n} \sum_{j=0}^{n-1} T(j) + cn \\ &= \frac{2}{n} \sum_{i=0}^{n-1} T(i) + cn \end{aligned} \quad (13.13)$$

两边同时乘以  $n$ ：

$$nT(n) = 2 \sum_{i=0}^{n-1} T(i) + cn^2 \quad (13.14)$$

将  $n$  用  $n - 1$  替换，得到另一等式：

$$(n - 1)T(n - 1) = 2 \sum_{i=0}^{n-2} T(i) + c(n - 1)^2 \quad (13.15)$$

用式 (13.14) 减去式 (13.15)，消去所有的  $T(i)$ ，其中  $0 \leq i < n - 1$ 。

$$nT(n) = (n + 1)T(n - 1) + 2cn - c \quad (13.16)$$

忽略常数项  $c$ ，化简为：

$$\frac{T(n)}{n + 1} = \frac{T(n - 1)}{n} + \frac{2c}{n + 1} \quad (13.17)$$

依次代入  $n - 1, n - 2, \dots$  得到  $n - 1$  个等式。

$$\frac{T(n - 1)}{n} = \frac{T(n - 2)}{n - 1} + \frac{2c}{n}$$

$$\begin{aligned} \frac{T(n-2)}{n-1} &= \frac{T(n-3)}{n-2} + \frac{2c}{n-1} \\ &\dots \\ \frac{T(2)}{3} &= \frac{T(1)}{2} + \frac{2c}{3} \end{aligned}$$

将所有等式相加,消去左右相同的部分,化简为  $n$  的函数。

$$\frac{T(n)}{n+1} = \frac{T(1)}{2} + 2c \sum_{k=3}^{n+1} \frac{1}{k} \quad (13.18)$$

利用调和级数,最终的结果为:

$$O\left(\frac{T(n)}{n+1}\right) = O\left(\frac{T(1)}{2} + 2c \ln n + \gamma + \epsilon_n\right) = O(\lg n) \quad (13.19)$$

因此

$$O(T(n)) = O(n \lg n) \quad (13.20)$$

#### 13.1.4 改进

PARTITION 在处理大量重复元素时性能不佳。考虑含有  $n$  个相等元素的特殊序列  $[x, x, \dots, x]$ :

1. 根据定义:任选某一元素作为基准  $p = x$ ,划分为一个空序列和长度为  $n-1$  的序列:  $[x, x, \dots, x]$ 。然后递归地对长为  $n-1$  的序列排序。总复杂度为  $O(n^2)$ 。
2. 用条件  $< x$  和  $> x$  划分。结果是两个空序列和  $n$  个等于  $x$  的元素。对两个空序列递归并立即结束。结果为  $[] \# [x, x, \dots, x] \# []$ 。总复杂度为  $O(n)$ 。

据此将二分划分改进为三分划分来处理重复元素。

$$\begin{aligned} \text{sort } [] &= [] \\ \text{sort } (x:xs) &= \text{sort } S \# \text{sort } E \# \text{sort } G \end{aligned} \quad (13.21)$$

其中:

$$\begin{cases} S = [y \leftarrow xs, y < x] \\ E = [y \leftarrow xs, y = x] \\ G = [y \leftarrow xs, y > x] \end{cases}$$

使用累积变量改进  $S, E, G$  的连接:  $qsort = \text{sort } []$ 。其中:

$$\begin{aligned} \text{sort } A [] &= A \\ \text{sort } A (x:xs) &= \text{sort } (E \# \text{sort } A G) S \end{aligned} \quad (13.22)$$

$E$  中元素全相等, 无需排序。先用累积变量  $A$  对  $G$  排序, 然后将  $E$  连接到结果前作为新的累积变量对  $S$  排序。划分也可以使用累积变量改进:

$$\begin{aligned} \text{part } S \ E \ G \ x \ [] &= (S, E, G) \\ \text{part } S \ E \ G \ x \ (y:ys) &= \begin{cases} y < x : (y:S, E, G) \\ y = x : (S, y:E, G) \\ y > x : (S, E, y:G) \end{cases} \end{aligned} \quad (13.23)$$

伯德给出了另一种改进<sup>[1]</sup>, 把递归排好的子列表收集起来再连接:

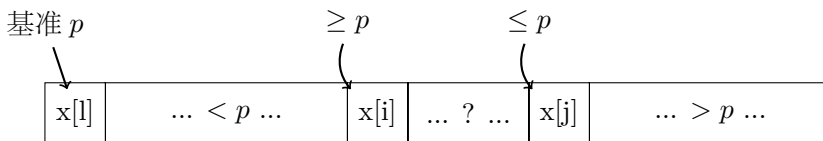
```

sort :: (Ord a) => [a] -> [a]
sort = concat o (pass [])

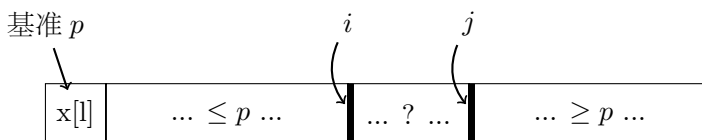
pass xss [] = xss
pass xss (x:xs) = step xs [] [x] [] xss where
  step [] as bs cs xss = pass (bs : pass xss cs) as
  step (x':xs') as bs cs xss | x' < x = step xs' (x':as) bs cs xss
                               | x' == x = step xs' as (x':bs) cs xss
                               | x' > x = step xs' as bs (x':cs) xss

```

塞奇维克(Sedgwick)给出了**双向扫描**划分方法<sup>[69][2]</sup>。使用两个指针  $i, j$  从左右两侧相向扫描。开始时  $i, j$  指向数组左右边界, 选择最左侧的元素作为基准  $p$ 。然后左指针  $i$  向右扫描直到遇到某个  $\geq p$  的元素; 另外<sup>1</sup> 右指针  $j$  向左扫描直到遇到某个  $\leq p$  的元素。此时, 所有  $i$  左边的元素  $< p$ ,  $j$  右边的元素  $> p$ 。  $i$  指向一个  $\geq p$  的元素, 而  $j$  指向一个  $\leq p$  的元素。如图 13.3 (a)。为了将全部  $\leq p$  的元素划分到左侧, 其余元素划分到右侧, 我们交换  $i$  和  $j$  指向的两个元素。然后继续扫描, 重复上面的步骤直到  $i$  和  $j$  相遇或者交错。在划分的任何时刻, 总保持不变条件: 所有  $i$  左边的元素(包括  $i$ ) 都  $\leq p$ ; 所有  $j$  右侧的元素(包括  $j$ ) 都  $\geq p$ 。  $i$  和  $j$  之间的元素尚未处理。如图 13.3 (b)。



(a) 指针  $i$  和  $j$  停止前进时



(b) 划分不变条件

图 13.3: 双向扫描

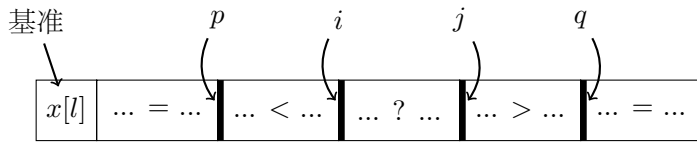
<sup>1</sup> 两轮扫描可以并发。

当  $i, j$  相遇或交错时,需要一次额外的交换,将最左侧的基准  $p$  交换到  $j$  指向的位置上。然后,对划分区间下界  $l$  和  $j$  之间的数组片段  $A[l...j)$ ;  $i$  和划分区间上界  $u$  之间的片段  $A[i...u)$  递归排序。

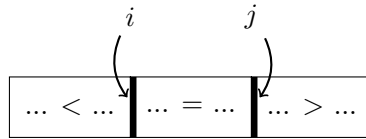
```

1: procedure SORT( $A, l, u$ )                                ▷ 排序区间:  $[l, u)$ 
2:   if  $u - l > 1$  then                                    ▷ 包含 1 个以上的元素
3:      $i \leftarrow l, j \leftarrow u$ 
4:      $p \leftarrow A[l]$                                   ▷ 基准
5:     loop
6:       repeat
7:          $i \leftarrow i + 1$ 
8:       until  $A[i] \geq p$                                   ▷ 忽略  $i \geq u$  的情况
9:       repeat
10:         $j \leftarrow j - 1$ 
11:      until  $A[j] \leq p$                                   ▷ 忽略  $j < l$  的情况
12:      if  $j < i$  then
13:        break
14:      EXCHANGE  $A[i] \leftrightarrow A[j]$ 
15:    EXCHANGE  $A[l] \leftrightarrow A[j]$                           ▷ 移动  $p$ 
16:    SORT( $A, l, j$ )
17:    SORT( $A, i, u$ )
    
```

考虑所有元素都相等的极端情况:数组被均分为两段,共进行  $\frac{n}{2}$  次交换。由于划分是平衡的,总体性能仍然为  $O(n \lg n)$ ,没有退化。和此前的实现对比,这一方法的交换次数更少。它跳过了那些在基准正确一侧的元素不进行处理。我们可以把双向扫描和三路划分结合起来。只对不等于基准的元素递归排序。本特利 (Jon Bentley) 和麦基尔罗伊 (Douglas McIlroy) 给出了一个方法:如图 13.4 (a) 所示,先把所有和基准相等的元素保存在两侧 [70] [71]。



(a) 三路划分的不变条件。



(b) 将等于基准的元素交换到中间。

图 13.4: 三路划分

仍然左右相向扫描,直到  $i$  遇到  $\geq$  基准的元素,并且  $j$  遇到  $\leq$  基准的元素。此时如果  $i, j$  没有相遇或者交错,不仅交换  $A[i] \leftrightarrow A[j]$ ,同时检查  $A[i], A[j]$  是否等于基准。如果相等,就交换  $A[i] \leftrightarrow A[p]$  或  $A[j] \leftrightarrow A[q]$ 。在划分结束前,需要把所有等于基准的元素从左右交换到中间。交换次数取决于重复元素的个数。如果所有元素都不同,则交换次数为零,不产生任何额外消耗。划分结果如图 13.4 (b) 所示。此后,我们只需要对“严格小于”和“严格大于”部分递归。

```

1: procedure SORT( $A, l, u$ )
2:   if  $u - l > 1$  then
3:      $i \leftarrow l, j \leftarrow u$ 
4:      $p \leftarrow l, q \leftarrow u$                                 ▷ 指向相等元素的边界
5:      $pivot \leftarrow A[l]$ 
6:     loop
7:       repeat
8:          $i \leftarrow i + 1$ 
9:       until  $A[i] \geq pivot$                                     ▷ 忽略  $i \geq u$  的错误处理
10:      repeat
11:         $j \leftarrow j - 1$ 
12:      until  $A[j] \leq pivot$                                     ▷ 忽略  $j < l$  的错误处理
13:      if  $j \leq i$  then
14:        break
15:      EXCHANGE  $A[i] \leftrightarrow A[j]$ 
16:      if  $A[i] = pivot$  then                                    ▷ 处理相等的元素
17:         $p \leftarrow p + 1$ 
18:        EXCHANGE  $A[p] \leftrightarrow A[i]$ 
19:      if  $A[j] = pivot$  then
20:         $q \leftarrow q - 1$ 
21:        EXCHANGE  $A[q] \leftrightarrow A[j]$ 
22:      if  $i = j$  且  $A[i] = pivot$  then
23:         $j \leftarrow j - 1, i \leftarrow i + 1$ 
24:      for  $k$  from  $l$  to  $p$  do                                    ▷ 将相等的元素交换到中间
25:        EXCHANGE  $A[k] \leftrightarrow A[j]$ 
26:         $j \leftarrow j - 1$ 
27:      for  $k$  from  $u - 1$  down-to  $q$  do
28:        EXCHANGE  $A[k] \leftrightarrow A[i]$ 
29:         $i \leftarrow i + 1$ 
30:      SORT( $A, l, j + 1$ )
31:      SORT( $A, i, u$ )

```

双向扫描加三路划分的逻辑变得复杂了,需要仔细处理各种边界条件。此前的单向扫描具有简单、直观的特点。我们也可以直接改进它。首先调整不变条件。选择第一个元素作为基准  $p$ ,如图 13.5 所示。任何时刻,左侧片段包含  $< p$  的元素;接下来的片段包含  $= p$  的元素;最右侧片段包含  $> p$  的元素。三个片段的边界分别为  $i, k, j$ 。 $[k, j)$  之间是尚未扫描的元素。从左向右逐一扫描。开始时, $< p$  的部分为空; $= p$  的部分只有一个元素。 $i$  指向数组的下界, $k$  指向  $i$  的下一个元素。 $> p$  的部分也为空, $j$  指向数组上界。

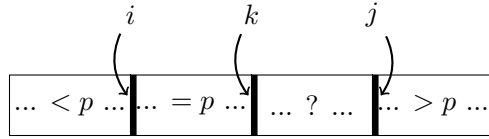


图 13.5: 单向扫描三路划分

逐一检查  $k$  指向的元素。如果  $A[k] = p$ , 就向前移动  $k$ ; 如果  $A[k] > p$ , 交换  $A[k] \leftrightarrow A[j - 1]$ , 这样  $> p$  的区间长度增加一。它的边界  $j$  向左移动一步。由于不确定移动到位置  $k$  的元素是否仍然大于  $p$ , 我们需要再次比较并重复上述过程。否则, 如果  $A[k] < p$ , 交换  $A[k] \leftrightarrow A[i]$ , 其中  $A[i]$  是  $= p$  的区间中的第一个元素。当  $k, j$  相遇时划分结束。

```

1: procedure SORT( $A, l, u$ )
2:   if  $u - l > 1$  then
3:      $i \leftarrow l, j \leftarrow u, k \leftarrow l + 1$ 
4:      $pivot \leftarrow A[i]$ 
5:     while  $k < j$  do
6:       while  $pivot < A[k]$  do
7:          $j \leftarrow j - 1$ 
8:         EXCHANGE  $A[k] \leftrightarrow A[j]$ 
9:       if  $A[k] < pivot$  then
10:        EXCHANGE  $A[k] \leftrightarrow A[i]$ 
11:         $i \leftarrow i + 1$ 
12:         $k \leftarrow k + 1$ 
13:     SORT( $A, l, i$ )
14:     SORT( $A, j, u$ )

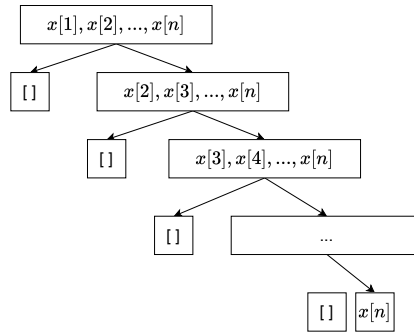
```

和双向扫描加三路划分相比,这一实现相对简单但需要更多的交换次数。

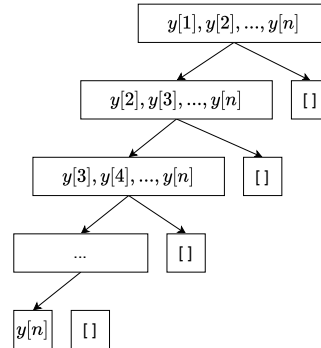
### 特殊情况

虽然三路划分能应对重复元素,但仍有一些困难的情况。例如,序列中的大部分元素已序时(无论是升序还是降序),划分结果不平衡。图 13.6 是两种极端情

况:  $[x_1 < x_2 < \dots < x_n]$  和  $[y_1 > y_2 > \dots > y_n]$  的划分结果。还有更多的例子, 如  $[x_m, x_{m-1}, \dots, x_2, x_1, x_{m+1}, x_{m+2}, \dots, x_n]$ , 其中  $[x_1 < x_2 < \dots < x_n]$ , 以及  $[x_n, x_1, x_{n-1}, x_2, \dots]$ 。如图 13.7 所示。



(a)  $[x_1 < x_2 < \dots < x_n]$  的划分树。≤  $p$  的部分总为空。



(b)  $[y_1 > y_2 > \dots > y_n]$  的划分树, ≥  $p$  的部分总为空。

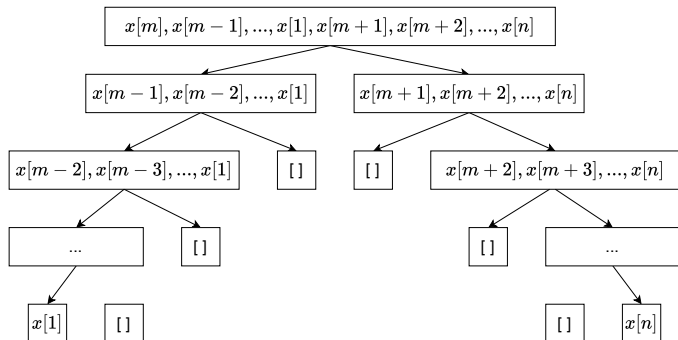
图 13.6: 特殊情况

这几种情况中, 用第一个元素作为基准造成不平衡划分。塞奇维克给出了一种改进<sup>[69]</sup>: 不在固定的位置上选择基准, 而是进行简单的抽样以减小不平衡的可能性。检查头部、中间、末尾元素, 选择这三者的中数作为基准。有两种方法: 一种最多需要三次比较<sup>[70]</sup>; 另一种通过交换将三者的最小值移动到头部, 将最大值移动到尾部, 将中数移动到中间。

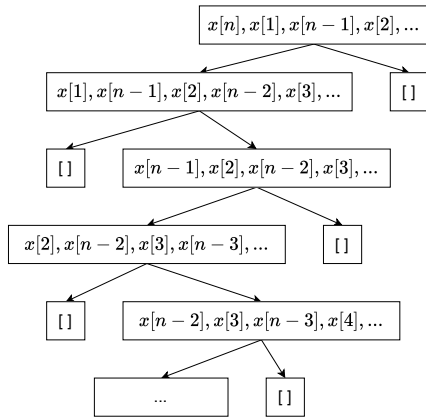
```

1: procedure SORT( $A, l, u$ )
2:   if  $u - l > 1$  then
3:      $m \leftarrow \lfloor \frac{l+u}{2} \rfloor$                                 ▷ 或使用  $l + \frac{u-l}{2}$  避免溢出
4:     if  $A[m] < A[l]$  then                                ▷ 确保  $A[l] \leq A[m]$ 
5:       EXCHANGE  $A[l] \leftrightarrow A[m]$ 
6:     if  $A[u-1] < A[l]$  then                                ▷ 确保  $A[l] \leq A[u-1]$ 
7:       EXCHANGE  $A[l] \leftrightarrow A[u-1]$ 
8:     if  $A[u-1] < A[m]$  then                                ▷ 确保  $A[m] \leq A[u-1]$ 
9:       EXCHANGE  $A[m] \leftrightarrow A[u-1]$ 

```



(a) 除了第一次划分,其它都不平衡。



(b) 一个之字形的划分树。

图 13.7: 更多特殊情况



```

10:      EXCHANGE  $A[l] \leftrightarrow A[m]$ 
11:       $(i, j) \leftarrow \text{PARTITION}(A, l, u)$ 
12:      SORT( $A, l, i$ )
13:      SORT( $A, j, u$ )

```

这一实现可以应对上述四种特殊情况,被称为“三点中值”法。另外,我们也可以随机选择基准:

```

1: procedure SORT( $A, l, u$ )
2:   if  $u - l > 1$  then
3:     EXCHANGE  $A[l] \leftrightarrow A[\text{RANDOM}(l, u)]$ 
4:      $(i, j) \leftarrow \text{PARTITION}(A, l, u)$ 
5:     SORT( $A, l, i$ )
6:     SORT( $A, j, u$ )

```

函数  $\text{RANDOM}(l, u)$  返回  $l$  和  $u$  之间的随机整数  $l \leq i < u$ 。交换  $A[i]$  到头部作为基准。此方法称为**随机快速排序**<sup>[4]</sup>。无论是三点中值还是随机快速排序都无法完全避免最差情况。如果序列随机分布,无论选择第一个还是其它位置上的元素作为基准,在效果上都是相同的。尽管如此,这些改进在实际中被广泛应用。

除改进划分外,还有一些工程实践。塞奇维克观察到在序列较短时,快速排序没有明显优势,而插入排序反而更快<sup>[2][70]</sup>。他和本特利、麦基尔罗伊测试了不同的序列长度,定义了一个阈值。如果元素个数少于阈值,就转而使用插入排序。

```

1: procedure SORT( $A, l, u$ )
2:   if  $u - l > \text{CUT-OFF}$  then
3:     QUICK-SORT( $A, l, u$ )
4:   else
5:     INSERTION-SORT( $A, l, u$ )

```

### 13.1.5 快速排序与树排序

“真正的快速排序”复合应用了多种改进——当序列较短时转为插入排序,就地交换元素,使用三点中值选择基准,双向扫描三路划分……简短的递归定义仅诠释了思路,却没有使用任何改进。有人认这样的快速排序本质上是树排序。快速排序和树排序的确有紧密的关系。伯德指出可以通过“砍伐”概念<sup>2</sup>,从二叉树排序推导出快速排序<sup>[72]</sup>。定义 *unfold* 函数将列表转换为二叉搜索树:

$$\begin{aligned}
 \text{unfold} [ ] &= \emptyset \\
 \text{unfold} (x:xs) &= (\text{unfold} [a \leftarrow xs, a \leq x], x, \text{unfold} [a \leftarrow xs, a > x])
 \end{aligned}
 \tag{13.24}$$

和二叉搜索树(见第二章)插入算法相比,*unfold* 产生树的方式大相径庭。如果列

<sup>2</sup>deforestation

表为空, 结果为一棵空树; 否则, 将列表中第一个元素  $x$  作为节点的值, 然后递归地构造左右子树。其中左子树是  $\leq x$  的元素; 而右子树是  $> x$  的元素。而将一棵二叉搜索树通过中序遍历转换成列表的定义为:

$$\begin{aligned} toList \ \emptyset &= [] \\ toList \ (l, k, r) &= toList \ l \ ++ [k] \ ++ toList \ r \end{aligned} \quad (13.25)$$

将上述两个函数组合起来, 就可定义出快速排序算法:

$$sort = toList \circ unfold \quad (13.26)$$

我们先通过 *unfold* 构造出二叉搜索树, 将其作为中间结果送入 *toList*, 结束后丢弃。如果将临时的中间结果树消除, 就得到了基本的快速排序算法<sup>[73]</sup>。

## 13.2 归并排序

快速排序性能优异, 但无法完全避免最差情况。归并排序在所有情形下都保证  $O(n \lg n)$  的复杂度, 适用于数组和列表, 在算法设计和分析上具有重要意义。很多编程环境用归并排序作为标准工具<sup>3</sup>。归并排序也利用了分治策略。每次将序列均分, 递归排序、归并。

$$\begin{aligned} sort \ [] &= [] \\ sort \ [x] &= [x] \\ sort \ xs &= merge \ (sort \ as) \ (sort \ bs), \text{ 其中: } (as, bs) = halve \ xs \end{aligned} \quad (13.27)$$

其中 *halve* 将序列对半分开。对于数组, 我们可以直接在中间位置分割:  $splitAt \ \lfloor \frac{|xs|}{2} \rfloor \ xs$ 。但列表需要线性时间移动到中点(见式 (1.45)):

$$splitAt \ n \ xs = shift \ n \ [] \ xs \quad (13.28)$$

其中:

$$\begin{aligned} shift \ 0 \ as \ bs &= (as, bs) \\ shift \ n \ as \ (b:bs) &= shift \ (n - 1) \ (b:as) \ bs \end{aligned} \quad (13.29)$$

考虑均分无需要保持元素间顺序, 我们可以利用奇偶位置简化。奇偶要么同样多, 要么仅差一个, 总能保证平衡。  $halve = split \ [] \ []$ , 其中:

$$\begin{aligned} split \ as \ bs \ [] &= (as, bs) \\ split \ as \ bs \ [x] &= (x:as, bs) \\ split \ as \ bs \ (x:y:xs) &= split \ (x:as) \ (y:bs) \ xs \end{aligned} \quad (13.30)$$

我们也可以用叠加进一步简化如下, 每次把元素  $x$  添加到  $as$  上, 然后交换  $as \leftrightarrow bs$ :

<sup>3</sup>如 Haskell, Python 和 Java

```
halve = foldr f ([], []) where
  f x (as, bs) = (bs, x : as)
```

归并的思想如图 13.8 所示。考虑两按照身高排队的人,他们按身高顺序依次通过一扇门,每次一人。由于两队已序,我们比较两个排头,个子较小的通过门;重复这一步骤,直到任何一队都已经通过,然后剩下的一队再逐一通过。

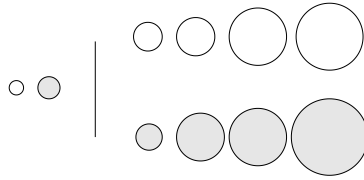


图 13.8: 归并

$$\begin{aligned}
 \text{merge } [] \text{ } bs &= bs \\
 \text{merge } as \text{ } [] &= as \\
 \text{merge } (a:as) \text{ } (b:bs) &= \begin{cases} a < b : a : \text{merge } as \text{ } (b:bs) \\ \text{否则} : b : \text{merge } (a:as) \text{ } bs \end{cases} \quad (13.31)
 \end{aligned}$$

对于数组,可以直接在中点分割,分别递归排序后再归并:

```
1: procedure SORT(A)
2:   n ← |A|
3:   if n > 1 then
4:     m ← ⌊n/2⌋
5:     X ← COPY-ARRAY(A[1...m])
6:     Y ← COPY-ARRAY(A[m + 1...n])
7:     SORT(X)
8:     SORT(Y)
9:     MERGE(A, X, Y)
```

我们使用了和数组  $A$  同样大小的空间进行归并。这因为 MERGE 不是原地修改元素的。归并时不断检查  $X$ 、 $Y$  中的元素,选择较小的放入  $A$ ,继续向前直到处理完任一数组。最后把另一数组中的剩余元素添加到  $A$  中。

```
1: procedure MERGE(A, X, Y)
2:   i ← 1, j ← 1, k ← 1
3:   m ← |X|, n ← |Y|
4:   while i ≤ m 且 j ≤ n do
5:     if X[i] < Y[j] then
6:       A[k] ← X[i]
7:       i ← i + 1
```

```

8:     else
9:         A[k] ← Y[j]
10:        j ← j + 1
11:       k ← k + 1
12:   while i ≤ m do
13:       A[k] ← X[i]
14:       k ← k + 1
15:       i ← i + 1
16:   while j ≤ n do
17:       A[k] ← Y[j]
18:       k ← k + 1
19:       j ← j + 1

```

为了简化归并  $X, Y$  的逻辑, 可以将  $\infty$  添加到末尾<sup>4</sup>。

```

1: procedure MERGE(A, X, Y)
2:   APPEND(X, ∞)
3:   APPEND(Y, ∞)
4:   i ← 1, j ← 1, n ← |A|
5:   for k ← from 1 to n do
6:     if X[i] < Y[j] then
7:       A[k] ← X[i]
8:       i ← i + 1
9:     else
10:      A[k] ← Y[j]
11:      j ← j + 1

```

### 13.2.1 性能分析

归并排序分为两步: 划分和归并。由于总是均分, 划分树是平衡二叉树, 如图 13.2 所示。树的高度和递归深度都是  $O(\lg n)$ 。每层都进行归并, 逐一比较两个序列的元素, 当其中一个处理完, 将另一序列中的剩余元素复制到结果中。归并的复杂度为线性时间。对长度为  $n$  的序列, 令  $T(n)$  为排序所时间, 有以下递归关系:

$$T(n) = T\left(\frac{n}{2}\right) + T\left(\frac{n}{2}\right) + cn = 2T\left(\frac{n}{2}\right) + cn \quad (13.32)$$

排序时间包含三部分: 对前半部分排序  $T\left(\frac{n}{2}\right)$ , 对后半部分排序  $T\left(\frac{n}{2}\right)$ , 归并  $cn$ , 其中  $c$  是某个常数。解此方程得到结果为  $O(n \lg n)$ 。另一重要性能指标是空间复杂度。不同归并方法的空间消耗大相径庭。对于上面的基本实现, 每次递归都需要和输入数

<sup>4</sup>降序使用  $-\infty$

组同样大小的空间,用以复制元素。递归返回后,可以释放这些空间。最大消耗出现在最深层递归时,为  $O(n \lg n)$ 。

### 13.2.2 改进

归并时反复获取空间复制数组是一个瓶颈<sup>[2]</sup>。我们可以一次性准备好大小为  $n = |A|$  的工作区用于归并,最后再把工作区内容复制回原数组。

```

1: procedure SORT(A)
2:    $n \leftarrow |A|$ 
3:   SORT'(A, CREATE-ARRAY( $n$ ), 1,  $n$ )

4: procedure SORT'(A, B, l, u)
5:   if  $u - l > 0$  then
6:      $m \leftarrow \lfloor \frac{l+u}{2} \rfloor$ 
7:     SORT'(A, B, l, m)
8:     SORT'(A, B, m + 1, u)
9:     MERGE'(A, B, l, m, u)

```

需要修改 MERGE' 使用传入的工作区:

```

1: procedure MERGE'(A, B, l, m, u)
2:    $i \leftarrow l, j \leftarrow m + 1, k \leftarrow l$ 
3:   while  $i \leq m$  且  $j \leq u$  do
4:     if  $A[i] < A[j]$  then
5:        $B[k] \leftarrow A[i]$ 
6:        $i \leftarrow i + 1$ 
7:     else
8:        $B[k] \leftarrow A[j]$ 
9:        $j \leftarrow j + 1$ 
10:     $k \leftarrow k + 1$ 
11:   while  $i \leq m$  do
12:      $B[k] \leftarrow A[i]$ 
13:      $k \leftarrow k + 1$ 
14:      $i \leftarrow i + 1$ 
15:   while  $j \leq u$  do
16:      $B[k] \leftarrow A[j]$ 
17:      $k \leftarrow k + 1$ 
18:      $j \leftarrow j + 1$ 
19:   for  $i \leftarrow$  from  $l$  to  $u$  do
20:      $A[i] \leftarrow B[i]$ 

```

▷ 复制回

这一改进将空间复杂度从  $O(n \lg n)$  降低到  $O(n)$ 。对 10 万个整数排序,性能可提升 20%到 25%。

### 13.2.3 原地归并排序

为了避免使用额外空间,考虑如何复用原数组作为工作区。如图 13.9 所示,子数组  $X$  和  $Y$  已排序好,原地归并时,令  $l$  之前为已归并部分(已序)。若  $A[l] < A[m]$ ,向前移动  $l$  一步;否则若  $A[l] \geq A[m]$  需要把  $A[m]$  放入归并结果中,位于  $l$  之前。为此需要把所有位于区间  $[l, m)$  中<sup>5</sup>的元素向后平移一个位置。

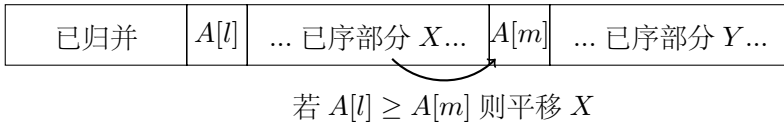


图 13.9: 原地平移归并

```

1: procedure MERGE( $A, l, m, u$ )
2:   while  $l \leq m \leq u$  do
3:     if  $A[l] < A[m]$  then
4:        $l \leftarrow l + 1$ 
5:     else
6:        $x \leftarrow A[m]$ 
7:       for  $i \leftarrow m$  down-to  $l + 1$  do ▷ 平移
8:          $A[i] \leftarrow A[i - 1]$ 
9:        $A[l] \leftarrow x$ 

```

但平移使得时间复杂度退化成  $O(n^2)$ 。数组移动是线性时间的操作,和  $X$  中的元素个数成正比。对子数组排序时,我们使用数组剩余部分作为工作区。对那些已在工作区内的元素,要避免它们被覆盖丢失。比较已序子数组  $X, Y$ ,每当把较小的元素放入工作区时,同时将工作区中的元素交换出来。归并完成后, $X$  和  $Y$  就保存了此前工作区的内容。如图 13.10 所示。

$X, Y$  和工作区  $Z$  都是原数组的一部分。归并时我们传入  $X, Y$  的起始、结束位置:区间  $[i, m), [j, n)$  和工作区的起始位置  $k$ 。

```

1: procedure MERGE( $A, [i, m), [j, n), k$ )
2:   while  $i < m$  且  $j < n$  do
3:     if  $A[i] < A[j]$  then
4:       EXCHANGE  $A[k] \leftrightarrow A[i]$ 
5:        $i \leftarrow i + 1$ 

```

<sup>5</sup> $[a, b)$  表示左闭右开区间,包括  $a$ ,但不包括  $b$ 。

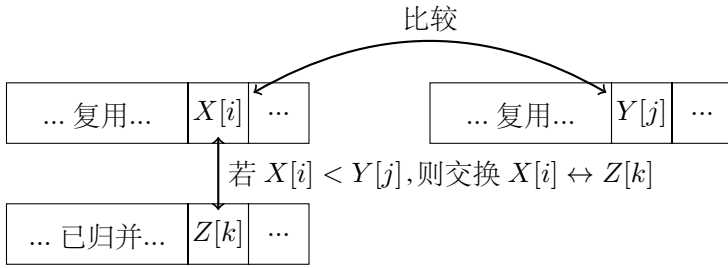


图 13.10: 归并时将工作区内容交换出来

```

6:     else
7:         EXCHANGE A[k] ↔ A[j]
8:         j ← j + 1
9:         k ← k + 1
10:    while i < m do
11:        EXCHANGE A[k] ↔ A[i]
12:        i ← i + 1
13:        k ← k + 1
14:    while j < m do
15:        EXCHANGE A[k] ↔ A[j]
16:        j ← j + 1
17:        k ← k + 1
    
```

工作区需要满足两个条件:1)必须足够大,以容纳交换进来的元素而不越界;2)工作区可以和任一子数组重叠,但须保证不覆盖尚未归并的元素。可以用一半数组作为工作区来归并排序另一半,如图 13.11 所示。

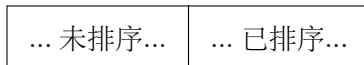


图 13.11: 归并排序一半数组

如果再递归地对一半工作区排序,就只剩下  $\frac{1}{4}$  了,如图 13.12 所示。我们必须在某个时刻归并  $X(\frac{1}{2}$  数组)和  $Y(\frac{1}{4}$  数组)。但剩余工作区大小只能容纳  $\frac{1}{4}$  数组,不能容纳  $X + Y$  的结果。

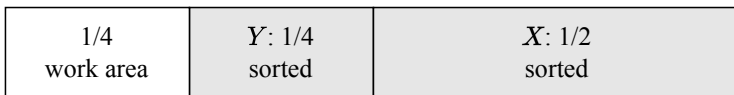


图 13.12: 工作区不足以归并 X 和 Y

工作区的第二条规则启发我们:通过某种设计,使得未归并的元素不被覆盖,从而

利用工作区和子数组的重叠部分来解决问题。我们先对工作区的后 1/2 排序,结果  $Y$  被交换到前 1/2 部分,新工作区就位于  $X, Y$  中间,如图 13.13 上方所示。这样的安排使得工作区和子数组  $X$  产生了重叠<sup>[74]</sup>。考虑两种极端情况:

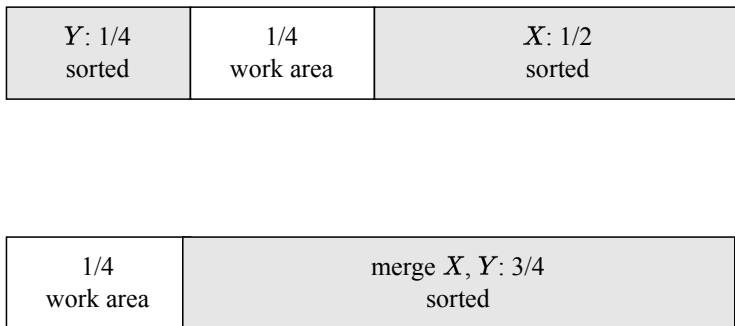


图 13.13: 利用工作区归并  $X, Y$

1. 所有  $Y$  中元素  $y$  都小于  $X$  中元素  $x$ 。归并后,  $Y$  和工作区内容交换( $Y$  和工作区大小相等);
2. 所有  $X$  中元素  $x$  都小于  $Y$  中元素  $y$ 。归并不断交换  $X$  和工作区中的内容。当工作区被  $X$  中一半元素填满后,开始覆盖  $X$  中的内容。幸运的是,被覆盖的是已归并的元素。工作区的右侧边界不断向数组的末尾移动到 3/4 的位置。此后,归并算法开始交换  $Y$  和工作区的内容。最终工作区被移动到了数组最左侧,如图 13.13 下方所示。

其它情况介于两种极端情况之间,最终工作区被移动到前 1/4。我们重复这一步骤,总是对未排序部分的后 1/2 排序,将结果交换到前 1/2,而使新工作区位于中间。不断将工作区的大小减半,从  $\frac{1}{2}$  到  $\frac{1}{4}$  到  $\frac{1}{8}$ ……归并的规模不断下降。当工作区中只剩下一个元素时结束。归并只有一个元素的数组等价于插入。也可以使用插入排序来处理最后的几个元素。

- ```

1: procedure SORT( $A, l, u$ )
2:   if  $u - l > 0$  then
3:      $m \leftarrow \lfloor \frac{l + u}{2} \rfloor$ 
4:      $w \leftarrow l + u - m$ 
5:     SORT'( $A, l, m, w$ ) ▷ 对一半数组排序
6:     while  $w - l > 1$  do
7:        $u' \leftarrow w$ 
8:        $w \leftarrow \lceil \frac{l + u'}{2} \rceil$  ▷ 工作区减半
9:       SORT'( $A, w, u', l$ ) ▷ 对剩余一半排序
10:      MERGE( $A, [l, l + u' - w], [u', u], w$ )
11:    for  $i \leftarrow w$  down-to  $l$  do ▷ 改用插入排序

```



```

12:          $j \leftarrow i$ 
13:         while  $j \leq u$  且  $A[j] < A[j - 1]$  do
14:             EXCHANGE  $A[j] \leftrightarrow A[j - 1]$ 
15:              $j \leftarrow j + 1$ 

```

为了保证工作区足够大,我们使用上限取整。我们将包含结束位置的区间信息传入了 MERGE 算法。接下来,我们需要定义 SORT' 算法,它反过来递归调用 SORT 来交换工作区和已序部分。

```

1: procedure SORT'(A, l, u, w)
2:   if  $u - l > 0$  then
3:      $m \leftarrow \lfloor \frac{l + u}{2} \rfloor$ 
4:     SORT(A, l, m)
5:     SORT(A, m + 1, u)
6:     MERGE(A, [l, m), [m + 1, u), w)
7:   else ▷ 将所有元素交换到工作区
8:     while  $l \leq u$  do
9:       EXCHANGE  $A[l] \leftrightarrow A[w]$ 
10:       $l \leftarrow l + 1$ 
11:       $w \leftarrow w + 1$ 

```

这一方法不需要平移子数组。未排序部分不断递减:  $\frac{n}{2}, \frac{n}{4}, \frac{n}{8}, \dots$ , 总共需要  $O(\lg n)$  步完成排序。每次递归对剩余部分的一半排序,然后使用线性时间进行归并。令  $n$  个元素排序时间为  $T(n)$ ,有如下递归关系:

$$T(n) = T\left(\frac{n}{2}\right) + c\frac{n}{2} + T\left(\frac{n}{4}\right) + c\frac{3n}{4} + T\left(\frac{n}{8}\right) + c\frac{7n}{8} + \dots \quad (13.33)$$

对于一半的元素,花费的时间为:

$$T\left(\frac{n}{2}\right) = T\left(\frac{n}{4}\right) + c\frac{n}{4} + T\left(\frac{n}{8}\right) + c\frac{3n}{8} + T\left(\frac{n}{16}\right) + c\frac{7n}{16} + \dots \quad (13.34)$$

两式相减式 (13.33) - 式 (13.34):

$$T(n) - T\left(\frac{n}{2}\right) = T\left(\frac{n}{2}\right) + cn\left(\frac{1}{2} + \frac{1}{2} + \dots\right)$$

共有  $\lg n$  个  $\frac{1}{2}$  相加,由此得:

$$T(n) = 2T\left(\frac{n}{2}\right) + \frac{c}{2}n \lg n$$

使用裂项求和法解方程(或利用主定理),得到结果  $O(n \lg^2 n)$ 。

### 13.2.4 自然归并排序

高德纳给出了另外一种实现。思路如同从两 endpoint 燃蜡烛<sup>[51]</sup>,称为自然归并排序。对于任何序列,可以从任何位置开始找到一个非递减子序列。特别地,我们总可以在最左侧找到这样的子序列,如下表。

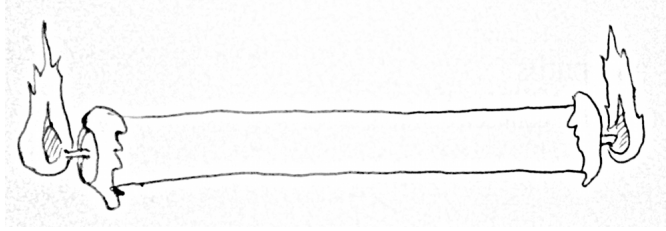


图 13.14: 从两端向中间燃烧的蜡烛

15, 0, 4, 3, 5, 2, 7, 1, 12, 14, 13, 8, 9, 6, 10, 11  
8, 12, 14, 0, 1, 4, 11, 2, 3, 5, 9, 13, 10, 6, 15, 7  
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

表中第一行是最短情况,子序列只包含一个元素;第三行是最长情况,子序列包含全部元素,整个序列已序。对称地,我们总可以从右向左找到非递减子序列,然后将两个子序列,一头一尾归并。这一思路的特点是利用元素间的自然顺序进行划分。

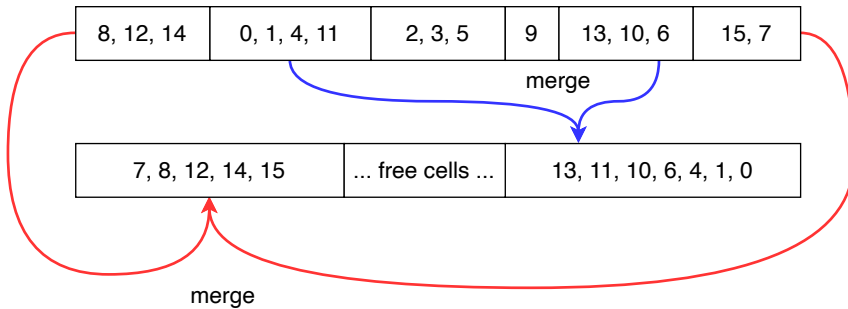


图 13.15: 自然归并排序

如图 13.15, 从两侧扫描, 分别找到最长的非递减子序列。然后归并到工作区左侧。重复这一步骤, 继续从两侧向中心进行扫描。这一次, 将两个子序列归并到工作区右侧, 从右向左放置。这样左右交替向工作区中心推进。当所有元素都被扫描归并到工作区后, 交换工作区和原数组, 再次从两侧向中心扫描归并。开始新一轮扫描时, 如果最长非递减子序列一直伸展到末尾, 说明整个序列已序, 排序结束。这一方法从左右两个方向处理数组, 并且使用子序列的自然顺序, 称为**两路自然归并排序**。如图 13.16 所示, 任何时候,  $a$  之前  $d$  之后的元素已处理完。将非递减子序列  $[a, b)$  向右扩展到最长, 同时将子序列  $[c, d)$  向左扩展到最长。对于工作区,  $f$  之前和  $r$  之后的元素已归并好(包含若干子序列)。奇数轮时, 将子序列  $[a, b)$  和  $[c, d)$  从  $f$  起向右归并; 偶数轮时, 从  $r$  起向左归并。

排序开始前, 准备好和数组同样大小的工作区。  $a$  和  $b$  指向最左侧,  $c$  和  $d$  指向最右侧。  $f$ 、 $r$  指向工作区左右两端。

1: **function** SORT( $A$ )

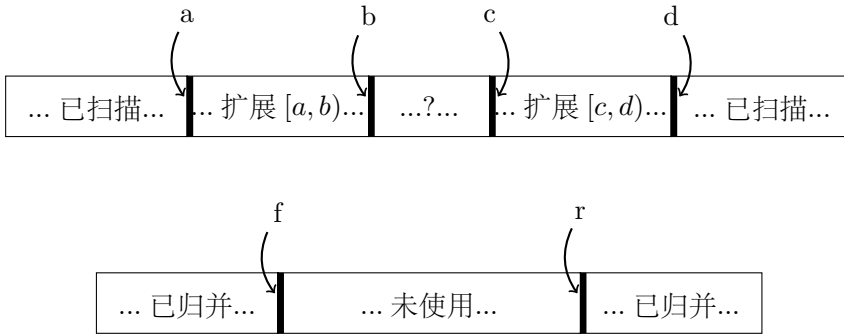


图 13.16: 自然归并排序时的不变性质

```

2:  if  $|A| > 1$  then
3:      $n \leftarrow |A|$ 
4:      $B \leftarrow \text{CREATE-ARRAY}(n)$  ▷ 创建工作区
5:     loop
6:          $[a, b] \leftarrow [1, 1]$ 
7:          $[c, d] \leftarrow [n + 1, n + 1]$ 
8:          $f \leftarrow 1, r \leftarrow n$  ▷ 指向工作区首尾
9:          $t \leftarrow 1$  ▷ 奇偶轮
10:        while  $b < c$  do ▷ 存在需要扫描的元素
11:            repeat ▷ 扩展 [a, b]
12:                 $b \leftarrow b + 1$ 
13:            until  $b \geq c$  或  $A[b] < A[b - 1]$ 
14:            repeat ▷ 扩展 [c, d]
15:                 $c \leftarrow c - 1$ 
16:            until  $c \leq b$  或  $A[c - 1] < A[c]$ 
17:            if  $c < b$  then ▷ 避免交错
18:                 $c \leftarrow b$ 
19:            if  $b - a \geq n$  then ▷ [a, b] 覆盖全数组时排序结束
20:                return  $A$ 
21:            if  $t$  是奇数 then ▷ 从左归并
22:                 $f \leftarrow \text{MERGE}(A, [a, b], [c, d], B, f, 1)$ 
23:            else ▷ 从右归并
24:                 $r \leftarrow \text{MERGE}(A, [a, b], [c, d], B, r, -1)$ 
25:                 $a \leftarrow b, d \leftarrow c$ 
26:                 $t \leftarrow t + 1$ 
27:        EXCHANGE  $A \leftrightarrow B$  ▷ 切换工作区

```

28: **return**  $A$

归并时需要将归并方向传入。

```

1: function MERGE( $A, [a, b), [c, d), B, w, \Delta$ )
2:   while  $a < b$  且  $c < d$  do
3:     if  $A[a] < A[d - 1]$  then
4:        $B[w] \leftarrow A[a]$ 
5:        $a \leftarrow a + 1$ 
6:     else
7:        $B[w] \leftarrow A[d - 1]$ 
8:        $d \leftarrow d - 1$ 
9:      $w \leftarrow w + \Delta$ 
10:  while  $a < b$  do
11:     $B[w] \leftarrow A[a]$ 
12:     $a \leftarrow a + 1$ 
13:     $w \leftarrow w + \Delta$ 
14:  while  $c < d$  do
15:     $B[w] \leftarrow A[d - 1]$ 
16:     $d \leftarrow d - 1$ 
17:     $w \leftarrow w + \Delta$ 
18:  return  $w$ 

```

自然归并排序的性能和元素顺序无关。考虑极端情况：第一轮扫描时，非递减子序列的长度总为 1。这轮结束后，工作区中归并的已序子数组的长度为 2。假设下一轮仍然如此，但是非递减子序列的长度已不可能小于 2。这一轮后，工作区包含长度至少为 4 的归并结果……每一轮后，归并的已序子数组的长度都加倍，因此最多需要  $O(\lg n)$  轮扫描归并。每一轮扫描所有的元素，总性能为  $O(n \lg n)$ 。对于列表，我们无法从首尾两端扫描。考虑列表由若干非递减子列表组成，可以每次取两个子列表归并。反复取出归并后，非递减子列表的数目不断减半，最后得到排序结果。定义(柯里化)：

$$\text{sort} = \text{sort}' \circ \text{group} \quad (13.35)$$

其中  $\text{group}$  将元素分组为非递减子列表：

$$\begin{aligned} \text{group} [] &= [[]] \\ \text{group} [x] &= [[x]] \\ \text{group} (x:y:xs) &= \begin{cases} x < y : (x:g):gs, \text{其中 } (g:gs) = \text{group} (y:xs) \\ \text{否则} : [x]:g:gs \end{cases} \end{aligned} \quad (13.36)$$

$$\begin{aligned} \text{sort}' [] &= [] \\ \text{sort}' [g] &= g \\ \text{sort}' gs &= \text{sort}' (\text{mergePairs } gs) \end{aligned} \quad (13.37)$$

其中:

$$\begin{aligned} \text{mergePairs } (g_1:g_2:gs) &= \text{merge } g_1 \ g_2 : \text{mergePairs } gs \\ \text{mergePairs } gs &= gs \end{aligned} \quad (13.38)$$

## 练习 13.2

13.2.1. 定义叠加合并:  $\text{sort}' = \text{foldr merge } []$ , 它与两两合并( $\text{mergePairs}$ )的复杂度相同么? 如果相同, 请给出证明; 如果不同, 哪个更快?

答案: 391 页

### 13.2.5 自底向上归并排序

自然归并排序的复杂度分析给出了一种自底向上的排序方法, 可以很方便地用迭代实现。首先将序列变成  $n$  个列表, 每个列表包含一个元素。然后将相邻的子序列两两归并, 得到  $\frac{n}{2}$  个长度为 2 的已序列表; 如果  $n$  是奇数, 会剩余一个长度为 1 的子序列。我们不断成对归并相邻子序列, 最后得到排序的结果。高德纳称之为“直接两路归并排序”<sup>[51]</sup>, 如图 13.17 所示。

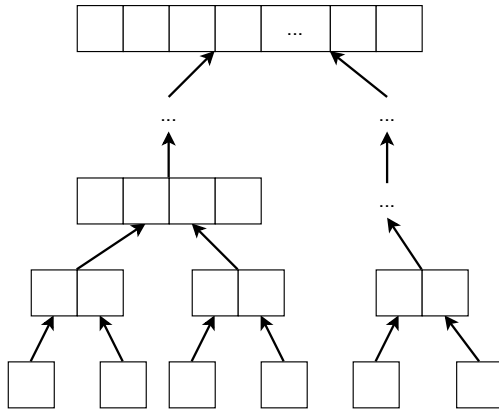


图 13.17: 自底向上归并排序

我们无需在每次递归时分割列表。将  $[x_1, x_2, \dots, x_n]$  转换为  $[[x_1], [x_2], \dots, [x_n]]$ , 然后不断归并。

$$\text{sort} = \text{sort}' \circ \text{map}(x \mapsto [x]) \quad (13.39)$$

我们复用自然归并排序的  $\text{mergePairs}$ , 直到归并为一个<sup>[3]</sup>列表。自底向上归并排序和自然归并排序很像, 仅仅是分组方法不同。本质上, 它可以由自然归并排序的特殊情况(最差情况)推导出来。自然归并排序总是将非递减子列表扩展到最长, 而自底向上归并排序仅把子列表的长度扩展到 1。我们可以把尾递归形式转换成迭代:

1: **function** SORT( $A$ )

```

2:   $n \leftarrow |A|$ 
3:   $B \leftarrow \text{CREATE-ARRAY}(n)$ 
4:  for  $i$  from 1 to  $n$  do
5:       $B[i] = A[i]$ 
6:  while  $n > 1$  do
7:      for  $i \leftarrow$  from 1 to  $\lfloor \frac{n}{2} \rfloor$  do
8:           $B[i] \leftarrow \text{MERGE}(B[2i-1], B[2i])$ 
9:      if  $\text{ODD}(n)$  then
10:          $B[\lceil \frac{n}{2} \rceil] \leftarrow B[n]$ 
11:      $n \leftarrow \lceil \frac{n}{2} \rceil$ 
12:  if  $B = [ ]$  then
13:     return  $[ ]$ 
14:  return  $B[1]$ 

```

### 练习 13.3

13.3.1. 将两两组合的操作抽象成叠加 *foldp*, 并使用它定义自底向上的归并排序。

答案: 391 页

## 13.3 并行处理

在快速排序中, 当划分完成时, 可以并行对两个子序列排序。这一策略对归并排序也适用。我们可以增多并行任务, 将序列分割成  $p$  个子序列, 其中  $p$  为处理器个数。理想情况下, 如果可以并行在  $T'$  时间内完成排序, 并且满足  $O(n \lg n) = pT'$ , 就称为**线性加速**, 这样的算法叫做最优化并行算法。但是, 简单地扩展基本快速排序算法: 选取  $p-1$  个基准, 划分为  $p$  个子序列, 然后并行对它们排序, 并不是最优化的。瓶颈出现在划分阶段, 我们只能得到平均  $O(n)$  的性能。另一方面, 并发归并排序算法的瓶颈出现在归并阶段。为了达到最优化的并行加速, 需要对并行归并排序和快速排序进行更好的设计。总体上说, 归并排序和快速排序的分治特性使它们相对容易并行化。科尔在 1986 年开发出了  $n$  个处理器, 性能为  $O(\lg n)$  的并行归并排序算法<sup>[76]</sup>。并行处理是一个复杂的题目, 超出了本书“基本算法”的范围<sup>[76][77]</sup>。

## 13.4 小结

本章介绍了两种常用的分治排序算法: 快速排序和归并排序。它们都达到了比较排序的性能上限  $O(n \lg n)$ 。塞奇维克评价快速排序是 20 世纪发现的最伟大的算法。大量的编程环境都使用快速排序作为标准排序工具。某些环境中, 特别是在处理动态

抽象序列的情况下,存储模型不是简单的数组,人们往往使用归并排序<sup>7</sup>。这一现象的原因,可以部分地在本章中找到。快速排序在大多数情况下表现优异。和其它算法相比,需要较少的交换操作。但在另一些环境中,交换并不是最有效的。底层的数据结构比较复杂,不是向量化的数组。而归并排序则适合这类环境,在各种情况下都能保证性能。反之快速排序的性能在最差情况下出现退化。在命令式环境中对数组排序时,归并排序的性能不如快速排序,并且需要额外的归并空间。在某些情况下,例如嵌入式系统,空间往往受到限制。原地归并排序仍然是一个活跃的研究领域。

快速排序和归并排序存在着联系。快速排序可以看作树排序的优化形式。归并排序也可以由树排序推导出来<sup>[75]</sup>。排序算法有多种分类<sup>[51]</sup>,其中一种是根据划分和归并的难易程度<sup>[72]</sup>。快速排序易于归并。基准前的子序列都小于等于基准后的子序列,归并实际上就是序列的连接。与此相反,归并排序的归并过程复杂,但划分却很简单,无论是等分、奇偶分割、自然分割、还是自底向上分割。快速排序很难保证完美分割,即使使用三点中值法、随机快速排序、三路划分等,仍无法彻底避免最差情况。

到本章为止,我们给出了基本排序算法,包括插入排序、树排序、选择排序、堆排序、快速排序、归并排序。排序仍然是计算机科学中活跃的研究领域。撰写本章时,人们正经历着“大数据”挑战,传统的排序方法无法在有限的时间和资源下处理越来越巨大的数据。在某些领域,处理几百 G 的数据已经成为日常工作中的任务。

## 练习 13.4

13.4.1. 使用归并排序的策略从一个序列产生二叉搜索树。

答案: 392 页

## 13.5 附录:例子程序

原地划分:

```

Int partition([K] xs, Int l, Int u) {
  for (Int pivot = l, Int r = l + 1; r < u; r = r + 1) {
    if xs[pivot] ≥ xs[r] {
      l = l + 1
      swap(xs[l], xs[r])
    }
  }
  swap(xs[pivot], xs[l])
  return l + 1
}

void sort([K] xs, Int l, Int u) {
  if l < u {
    Int m = partition(xs, l, u)
    sort(xs, l, m - 1)
    sort(xs, m, u)
  }
}

```

<sup>7</sup>实际中,大部分排序工具都是某种混合算法,在序列较短时使用插入排序来保持良好的性能

```

}
}

```

双向扫描:

```

void sort([K] xs, Int l, Int u) {
  if l < u - 1 {
    Int pivot = l, Int i = l, Int j = u
    loop {
      while i < u and xs[i] < xs[pivot] {
        i = i + 1
      }
      while j ≥ l and xs[pivot] < xs[j] {
        j = j - 1
      }
      if j < i then break
      swap(xs[i], xs[j])
    }
    swap(xs[pivot], xs[j])
    sort(xs, l, j)
    sort(xs, i, u)
  }
}

```

归并排序:

```

[K] sort([K] xs) {
  Int n = length(xs)
  if n > 1 {
    var ys = sort(xs[0 ... n/2 - 1])
    var zs = sort(xs[n/2 ...])
    xs = merge(xs, ys, zs)
  }
  return xs
}

[K] merge([K] xs, [K] ys, [K] zs) {
  Int i = 0
  while ys ≠ [] and zs ≠ [] {
    xs[i] = if ys[0] < zs[0] then pop(ys) else pop(zs)
    i = i + 1
  }
  xs[i...] = if ys ≠ [] then ys else zs
  return xs
}

```

使用工作区进行归并排序:

```

Void sort([K] xs) = msort(xs, copy(xs), 0, length(xs))

Void msort([K] xs, [K] ys, Int l, Int u) {
  if (u - l > 1) {
    Int m = l + (u - l) / 2
    msort(xs, ys, l, m)
  }
}

```



```

    msort(xs, ys, m, u)
    merge(xs, ys, l, m, u)
}
}

Void merge([K] xs, [K] ys, Int l, Int m, Int u) {
    Int i = l, Int k = l; Int j = m
    while i < m and j < u {
        ys[k++] = if xs[i] < xs[j] then xs[i++] else xs[j++]
    }
    while i < m {
        ys[k++] = xs[i++]
    }
    while j < u {
        ys[k++] = xs[j++]
    }
    while l < u {
        xs[l] = ys[l]
        l++
    }
}
}

```

原地归并排序:

```

Void merge([K] xs, (Int i, Int m), (Int j, Int n), Int w) {
    while i < m and j < n {
        swap(xs, w++, if xs[i] < xs[j] then i++ else j++)
    }
    while i < m {
        swap(xs, w++, i++)
    }
    while j < n {
        swap(xs, w++, j++)
    }
}

Void wsort([K] xs, (Int l, Int u), Int w) {
    if u - l > 1 {
        Int m = l + (u - l) / 2
        imsort(xs, l, m)
        imsort(xs, m, u)
        merge(xs, (l, m), (m, u), w)
    }
    else {
        while l < u { swap(xs, l++, w++) }
    }
}

Void imsort([K] xs, Int l, Int u) {
    if u - l > 1 {
        Int m = l + (u - l) / 2
        Int w = l + u - m
        wsort(xs, l, m, w)
    }
}

```

```

while w - l > 2 {
    Int n = w
    w = l + (n - l + 1) / 2;
    wsort(xs, w, n, l);
    merge(xs, (l, l + n - w), (n, u), w);
}
for Int n = w; n > l; --n {
    for Int m = n; m < u and xs[m] < xs[m-1]; m++ {
        swap(xs, m, m - 1)
    }
}
}
}

```

迭代式自底向上归并排序:

```

[K] sort([K] xs) {
    var ys = [[x] | x in xs]
    while length(ys) > 1 {
        ys += merge(pop(ys), pop(ys))
    }
    return if ys == [] then [] else pop(ys)
}

[K] merge([K] xs, [K] ys) {
    [K] zs = []
    while xs ≠ [] and ys ≠ [] {
        zs += if xs[0] < ys[0] then pop(xs) else pop(ys)
    }
    return zs ++ (if xs ≠ [] then xs else ys)
}

```

# 第十四章 搜索 I

人们借助现代计算机系统实现了复杂的搜索功能。生产线上的工业机器人从物料堆中找出零件进行组装,带有卫星导航的汽车从地图中找出最佳行驶路线,智能手机从网上商店中找出最便宜的购物方案……本章介绍最基本的查找和匹配搜索算法。

## 14.1 $k$ 选择问题

如何在  $n$  个元素中寻找第  $k$  大(或小)的元素。这里大小的含义是抽象的序关系。我们先用  $\leq$  关系找出第  $k$  小的元素,然后再推广到一般情况。最简单的方法是先找到最小的,然后在剩余元素中寻找第二小的。进行  $k$  次就可以找到第  $k$  小的元素。在  $n$  个元素中寻找最小元素是线性时间  $O(n)$  的。因此总时间为  $O(kn)$ 。我们也可以利用堆。这样可以在  $O(\lg n)$  时间内更新、获取顶部,重复  $k$  次可以在  $O(k \lg n)$  时间内获得答案。

$$\text{top } k \text{ } xs = \text{find } k \text{ } (\text{heapify } xs) \quad (14.1)$$

或写成柯里化形式:

$$\text{top } k = (\text{find } k) \circ \text{heapify} \quad (14.2)$$

其中:

$$\begin{aligned} \text{find } 0 &= \text{top} \\ \text{find } k &= (\text{find } (k - 1)) \circ \text{pop} \end{aligned} \quad (14.3)$$

我们还能找到更好的方法。使用分治策略将元素划分为  $A$ 、 $B$ ,使得  $A$  中元素都不大于( $\leq$ ) $B$  中元素。令  $m = |A|$  为  $A$  的大小,比较  $m$  和  $k$ :

1. 若  $k < m$ ,则第  $k$  小的元素在  $A$  中,我们丢弃  $B$ ,然后在  $A$  中递归查找;
2. 若  $m < k$ ,则第  $k$  小的元素在  $B$  中,我们丢弃  $A$ ,然后在  $B$  中递归查找第  $(k - m)$  小的元素。

理想情况下划分是平衡的,  $A$ 、 $B$  大小相当,每次问题规模减半,复杂度为  $O(n + n/2 + n/4 + \dots) = O(n)$ 。我们利用上一章快速排序中的划分方法  $part$ ,随机选择一

个元素(如第一个)作为基准  $p$ 。将所有  $\leq p$  的元素放入  $A$ , 剩余元素放入  $B$ 。这样若  $m = k - 1$ , 则  $p$  就是第  $k$  小的元素, 否则我们递归在  $A$  或  $B$  中查找。

$$\text{top } k(x:xs) = \begin{cases} m = k - 1: & x, \text{ 其中 } m = |A|, (A, B) = \text{part}(\leq x) xs \\ m < k - 1: & \text{top}(k - m - 1) B \\ \text{否则}: & \text{top } k A \end{cases} \quad (14.4)$$

和快速排序一样, 最差情况下划分结果总不平衡, 性能退化为  $O(kn)$  或  $O((n - k)n)$ 。平均情况下可以在线性时间内找到答案。我们也可以应用快速排序中的各种改进, 如三点中值法<sup>1</sup>和随机划分法:

```

1: function TOP( $k, A, l, u$ )
2:   EXCHANGE  $A[l] \leftrightarrow A[\text{RANDOM}(l, u)]$            ▷ 随机在  $[l, u]$  内选择
3:    $p \leftarrow \text{PARTITION}(A, l, u)$ 
4:   if  $p - l + 1 = k$  then
5:     return  $A[p]$ 
6:   if  $k < p - l + 1$  then
7:     return TOP( $k, A, l, p - 1$ )
8:   return TOP( $k - p + l - 1, A, p + 1, u$ )

```

可以调整这一方法获取全部前  $k$  个值( $k$  个值间的顺序是任意的), 如下面的例子程序:

```

tops _ [] = []
tops 0 _ = []
tops n (x:xs) | len == n = as
              | len < n = as ++ [x] ++ tops (n - len - 1) bs
              | otherwise = tops n as
where
  (as, bs) = partition ( $\leq x$ ) xs
  len = length as

```

## 14.2 二分查找

中学时老师表演过一个“数学魔术”: 学生随便想一个 1000 以内的数, 老师问十个问题, 学生回答是或否。然后老师就能猜出这个数。例如: 是偶数么? 是素数么? 所有位上的数字都相同么? 能被 3 整除么? 等等。在 1000 以内, 如果每次通过问答排除一半数字, 10 次内就能找出答案。因为  $2^{10} = 1024 > 1000$ 。“是偶数么?” 就是一个非常好的问题, 它能去掉一半的数字<sup>2</sup>。设计奇偶性这样的问题需要一定的技巧, 但存在着

<sup>1</sup>Blum、Floyd、Pratt、Rivest 和 Tarjan 在 1973 年给出了一个线性时间方法<sup>[4][81]</sup>。将列表划分为若干组, 每组最多 5 个元素, 总共选出  $n/5$  个中值。重复这一步骤选出中值的中值。

<sup>2</sup>社交网络上曾有一个“读心术”游戏。玩家暗想世界上任何一个人, 人工智能机器人提 16 个问题, 玩家回答是或否, 最后机器人能说出玩家暗想的人是谁。

简单的方法每次淘汰掉一半。我们经常在电视里看到这样的竞猜节目。主持人展示一件商品,然后让观众猜价格。并告之是猜高了还是低了。如果能够在 30 秒内猜对,就获得奖励。例如:1000 元,高了;500 元,低了;750 元,低了;890 元,低了;990 元,正确!这位观众使用了“二分查找”策略。为了在已序序列  $A$  中寻找  $x$ ,我们找到位于序列中点上的  $y$ ,和  $x$  比较。如果  $x = y$ ,查找结束;如果  $x < y$ ,由于  $A$  是已序的,我们只需要在前半部分继续查找;否则在后半部分查找。这样不断缩小  $A$  的规模,如果当  $A = []$  时仍未找到,则  $x$  不存在。二分查找要求  $A$  是已序的,我曾经看到有人试图对未排序的数组二分查找,深陷其中却得不到答案。高德纳说:“虽然二分查找的基本思想相对直观,具体细节却复杂得不可思议……”。本特利发现大多数二分查找的实现中含有错误。有趣的是他本人在《编程珠玑》第一版中给出的实现也包含了一个错误,直到 20 多年后才被发现<sup>[2]</sup>。下面给出了二分查找的实现,令数组  $A$  的上下界为  $l, u$  (不含  $u$ )。

$$bsearch\ x\ A\ (l, u) = \begin{cases} u < l: & \text{Nothing} \\ x = A[m]: & m, \text{其中: } m = l + \lfloor \frac{u-l}{2} \rfloor \\ x < A[m]: & bsearch\ x\ A\ (l, m-1) \\ \text{否则:} & bsearch\ x\ A\ (m+1, u) \end{cases} \quad (14.5)$$

我们也可以消除递归,通过更新边界实现二分查找:

```

1: function BINARY-SEARCH( $x, A, l, u$ )
2:   while  $l < u$  do
3:      $m \leftarrow l + \lfloor \frac{u-l}{2} \rfloor$                                 ▷ 避免  $\lfloor \frac{l+u}{2} \rfloor$  溢出
4:     if  $A[m] = x$  then
5:       return  $m$ 
6:     if  $x < A[m]$  then
7:        $u \leftarrow m - 1$ 
8:     else
9:        $l \leftarrow m + 1$ 
10:  Not found

```

由于每次将数组减半,二分查找的复杂度为  $O(\lg n)$ 。二分查找还可以搜索单调函数的解。例如方程  $a^x = y$ , 其中  $a \leq y, a, y$  都是自然数,我们寻找  $x$  的整数解。我们可以穷举:从 0 开始依次尝试  $a^0, a^1, a^2, \dots$ , 直到发现某个  $a^i = y$ , 或者  $a^i < y < a^{i+1}$ , 表示方程无整数解。对很大的  $a$  和  $x$ , 如果保持精度,则计算  $a^x$  会消耗一定的时间<sup>3</sup>。我们可以用二分查找来进行改进。首先估计解的上限。由于  $a^y \geq y$ , 我们可以在区间  $[0, 1, \dots, y]$  内搜索。由于函数  $f(x) = a^x$  是非减函数,对于自变量  $x$ ,我们先检查区间中点  $x_m = \lfloor \frac{0+y}{2} \rfloor$ , 如果  $a^{x_m} = y$ , 则  $x_m$  是方程的解;如果  $a^{x_m} < y$ , 我们丢弃  $x_m$  前的部分;否则丢弃  $x_m$  后的部分;两种情况下都将搜索范围减半,直到找到解或查找范围变成空,表示无整数解。下面是二分查找解的定义。我们将单调函数抽象为

<sup>3</sup>当然,我们可以复用  $a^n$  的结果来计算  $a^{n+1} = aa^n$ 。这里我们考虑一般意义下的单调函数  $f(n)$ 。

一个参数  $f$ , 调用  $bsearch\ f\ y\ (0, y)$ 。我们只需要计算  $O(\lg y)$  次  $f(x)$ , 好于穷举法。

$$bsearch\ f\ y\ (l, u) = \begin{cases} u < l: & \text{Nothing} \\ f(m) = y: & m, \text{ 其中: } m = \lfloor \frac{l+u}{2} \rfloor \\ f(m) < y: & bsearch\ f\ y\ (m+1, u) \\ f(m) > y: & bsearch\ f\ y\ (l, m-1) \end{cases} \quad (14.6)$$

### 14.3 二维查找

考虑把二分查找从一维扩展到二维或更高维。令矩阵  $M$  的大小为  $m \times n$ 。每行、每列的元素都是单调增加的自然数。如图 14.1 所示。如何快速地在矩阵中找到所有等于  $z$  的元素呢? 我们需要给出一个算法, 返回一组位置  $(i, j)$  的列表, 使得所有的  $M_{i,j} = z$

$$[(x, y) | x \leftarrow [1, 2, \dots, m], y \leftarrow [1, 2, \dots, n], M_{x,y} = z] \quad (14.7)$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 & \dots \\ 2 & 4 & 5 & 6 & \dots \\ 3 & 5 & 7 & 8 & \dots \\ 4 & 6 & 8 & 9 & \dots \\ \dots & & & & \end{bmatrix}$$

图 14.1: 每行、每列都单调增加

理查德·伯德曾用这一问题面试学生<sup>[4]</sup>。那些在中学就接触编程的学生往往尝试二分查找来解题, 但却很容易陷入困境。按照二分查找的思路, 通常先检查位于  $M_{\frac{m}{2}, \frac{n}{2}}$  上的元素。如果它小于  $z$ , 我们丢弃左上区域; 如果大于  $z$ , 丢弃右下区域, 如图 14.2 所示, 灰色的区域表示可丢弃。两种情况下, 搜索区域都从一个矩形变成了一个 L 形, 无法直接递归。我们把这类二维搜索问题抽象为: 已知函数  $f(x, y)$ , 在某一范围内寻找整数解  $(x, y)$ , 使得  $f(x, y) = z$ 。矩阵搜索可以特化为下面的函数:

$$f(x, y) = \begin{cases} 1 \leq x \leq m, 1 \leq y \leq n: & M_{x,y} \\ \text{其它}: & -1 \end{cases}$$

如果  $f(x, y)$  是单调增函数, 例如  $f(x, y) = x^a + y^b$ ,  $a, b$  是自然数, 有效解法不是从左下角出发, 而是从左上角出发开始查找<sup>[82]</sup>。如图 14.3 所示, 搜索从  $(0, z)$  开始, 对于每个点  $(p, q)$ , 我们比较  $f(p, q)$  和  $z$  的关系:

1. 如果  $f(p, q) < z$ : 由于  $f$  单调增, 所有的  $0 \leq y < q$ , 有  $f(p, y) < z$ 。我们丢弃垂直线段上的所有点(红色线段);

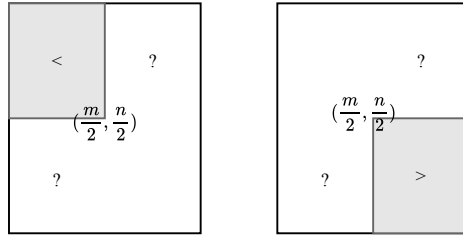


图 14.2: 左:中点元素小于  $z$ ,灰色区域都小于  $z$ ;右:中点元素大于  $z$ 。灰色区域都大于  $z$ 。

2. 如果  $f(p, q) > z$ :所有的  $p < x \leq z$ ,有  $f(x, q) > z$ 。我们丢弃水平线段上的所有点(蓝色线段);
3. 如果  $f(p, q) = z$ :( $p, q$ ) 是一个解,两条线段上的点都可以丢弃。

这样,我们就可以逐步缩小矩形搜索区域。每次要么丢弃一行,要么丢弃一列,或者同时丢弃行、列。

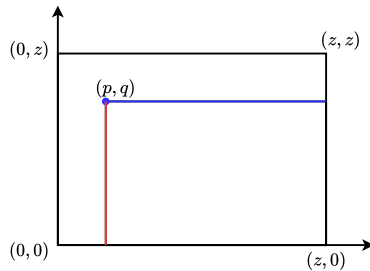


图 14.3: 从左上角搜索

定义  $search$  函数,并传入左上角开始搜索: $search\ f\ z\ 0\ z$ 。

$$search\ f\ z\ p\ q = \begin{cases} p > z \text{ 或 } q < 0: & [] \\ f(p, q) < z: & search\ f\ z\ (p + 1)\ q \\ f(p, q) > z: & search\ f\ z\ p\ (q - 1) \\ f(p, q) = z: & (p, q) : search\ f\ z\ (p + 1)\ (q - 1) \end{cases} \quad (14.8)$$

每次查找后, $p, q$  至少有一个会向右、下前进一步。最多需要  $2(z + 1)$  次查找完成搜索。最好情况有三种:(1)每次  $p, q$  同时前进一步,只要  $z + 1$  步就完成搜索;(2)不断水平向右前进,最后  $p$  超过  $z$ ;(3)不断垂直向下前进,最终  $q$  变为负。图 14.4 描述了最好和最坏的情况。图 14.4 (a) 中,对角线上的每个点  $(x, z - x)$  都满足  $f(x, z - x) = z$ ,总共需要  $z + 1$  步到达  $(z, 0)$ ;(b) 中,上方水平线的每个点  $(x, z)$  都使得  $f(x, z) < z$ , $z + 1$  步后,搜索结束;(c) 中,左侧垂线的每个点  $(0, x)$  都使得  $f(0, x) > z$ , $z + 1$  步后,搜索结束;(d) 描述的是最差情况。如果我们将搜索路径上的所有水平线段投射  $x$  轴

上,所有垂直线段投射到  $y$  轴上,就可以得到总共的搜索步数为  $2(z+1)$ 。和复杂度为  $O(z^2)$  的穷举法相比,这一改进将复杂度提高到线性时间  $O(z)$ 。

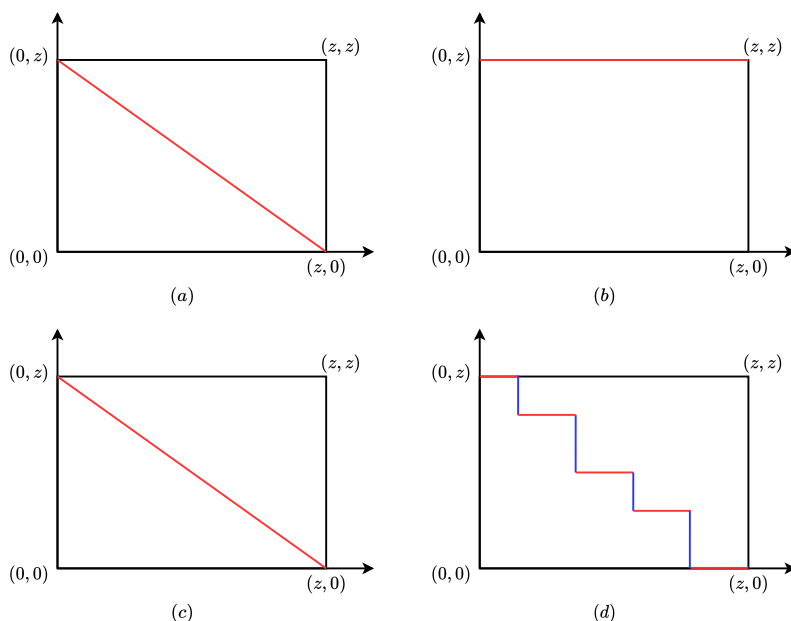


图 14.4: 最好和最差情况

这一方法叫做“马鞍搜索”。函数  $f$  的三维图像中,左下部的最小值和右上部的最大值,以及两侧的翼形图像,合起来像一个马鞍。如图 14.5 所示。马鞍搜索从左上角  $(0, z)$  向右下角  $(z, 0)$  搜索。这一范围可进一步缩小。 $f$  单调增,我们可以沿  $y$  轴找到最大  $m$ , 使得  $f(0, m) \leq z$ ; 沿  $x$  轴找到最大  $n$ , 使得  $f(n, 0) \leq z$ ; 这样搜索区域就从  $(0, z) - (z, 0)$  缩小到  $(0, m) - (n, 0)$ , 如图 14.6 所示。

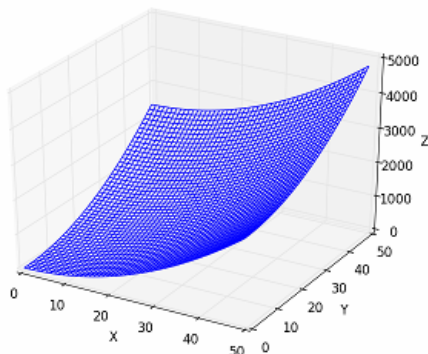


图 14.5:  $f(x, y) = x^2 + y^2$  的图像



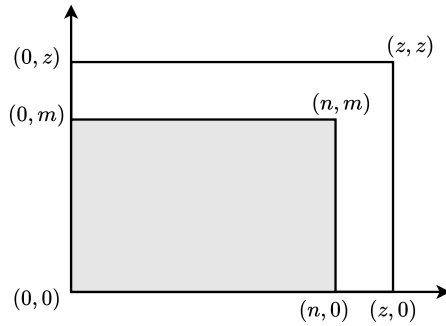


图 14.6: 缩小的灰色搜索区域

$$\begin{cases} m = \max [y | 0 \leq y \leq z, f(0, y) \leq z] \\ n = \max [x | 0 \leq x \leq z, f(x, 0) \leq z] \end{cases} \quad (14.9)$$

我们可以用二分查找搜索  $m$ 、 $n$  (固定  $x = 0$  查找  $m$ , 固定  $y = 0$  查找  $n$ )。改动式 (14.6) 的定义, 给定  $y$ , 寻找  $l \leq x \leq u$  满足  $f(x) \leq y < f(x+1)$

$$bsearch\ f\ y\ (l, u) = \begin{cases} u \leq l: & l \\ f(m) \leq y < f(m+1): & m, \text{ 其中 } m = \lfloor \frac{l+u}{2} \rfloor \\ f(m) \leq y: & bsearch\ f\ y\ (m+1, u) \\ f(m) > y: & bsearch\ f\ y\ (l, m-1) \end{cases} \quad (14.10)$$

这样就可以二分查找  $m$ 、 $n$ :

$$\begin{cases} m = bsearch\ (y \mapsto f(0, y))\ z\ (0, z) \\ n = bsearch\ (x \mapsto f(x, 0))\ z\ (0, z) \end{cases} \quad (14.11)$$

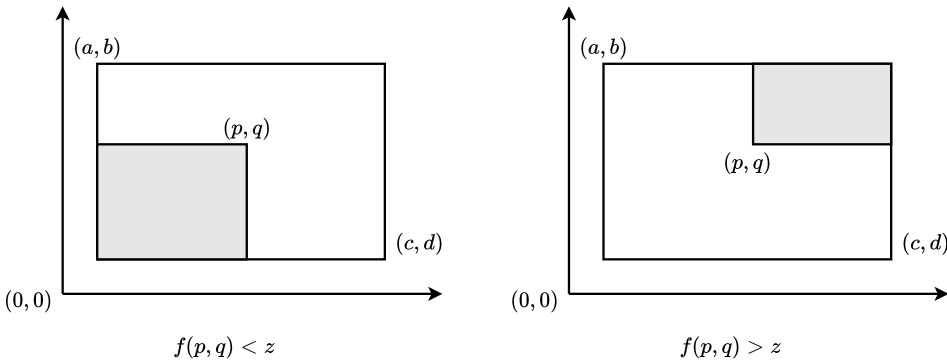
接下来在缩小的矩形内进行马鞍搜索:  $solve(f, z) = search\ f\ z\ 0\ m$

$$search\ f\ z\ p\ q = \begin{cases} p > n \text{ 或 } q < 0: & [] \\ f(p, q) < z: & search\ f\ z\ (p+1)\ q \\ f(p, q) > z: & search\ f\ z\ p\ (q-1) \\ f(p, q) = z: & (p, q) : search\ f\ z\ (p+1)\ (q-1) \end{cases} \quad (14.12)$$

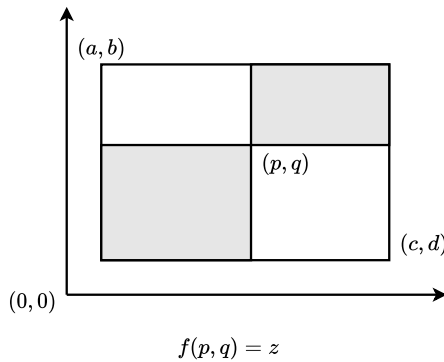
这一改进首先用两轮二分查找得到  $m$  和  $n$ , 每轮计算了  $O(\lg z)$  次  $f$ ; 马鞍搜索在最坏情况下计算  $O(m+n)$  次  $f$ ; 最好情况下计算  $O(\min(m, n))$  次。总体复杂度如下表。某些函数如  $f(x, y) = x^a + y^b$ , 对于自然数  $a, b$ , 边界  $m, n$  很小, 整体性能接近  $O(\lg z)$ 。

|      | 计算 $f$ 的次数              |
|------|-------------------------|
| 最坏情况 | $2 \log z + m + n$      |
| 最好情况 | $2 \log z + \min(m, n)$ |

如图 14.7, 考虑搜索区域  $(a, b) - (c, d)$  中的一点  $(p, q)$ , 若  $f(p, q) \neq z$ , 只能丢弃灰色部分 ( $\leq 1/4$ )。如果  $f(p, q) = z$ , 由于  $f$  单调增, 我们可同时丢弃左下、右上部分, 和  $p$  列、 $q$  行上的其它点。这样只剩下  $1/2$  的区域, 可以快速缩小搜索范围。为了找到  $f(p, q) = z$  的点, 我们沿着矩形中心水平线或中心垂直线应用二分查找。在线段  $L$  上二分查找的复杂度为  $O(\lg |L|)$ , 我们选取较短的中线搜索, 如图 14.8 所示。



(a) 如果  $f(p, q) \neq z$ , 只能丢弃左下或右上的灰色区域, 剩余区域变成了 L 形。



(b) 如果  $f(p, q) = z$ , 可同时丢弃两个灰色部分, 搜索区域减半。

图 14.7: 缩小搜索区域

如果中线上不存  $f(p, q) = z$  的点, 我们寻找满足  $f(p, q) < z < f(p+1, q)$  的点 (水平中线, 对于垂直中线为  $f(p, q) < z < f(p, q+1)$ )。此时我们不能将  $p$  列、 $q$  行上的点完全丢弃。综合起来, 我们用二分查找搜索水平中线上满足  $f(p, q) \leq z < f(p+1, q)$  的点, 或中垂直线上满足  $f(p, q) \leq z < f(p, q+1)$  的点。如果线段上所有点都使得  $f(p, q) < z$ , 则返回上界; 如果所有点都使得  $f(p, q) > z$ , 则返回下界。此时, 我们丢弃

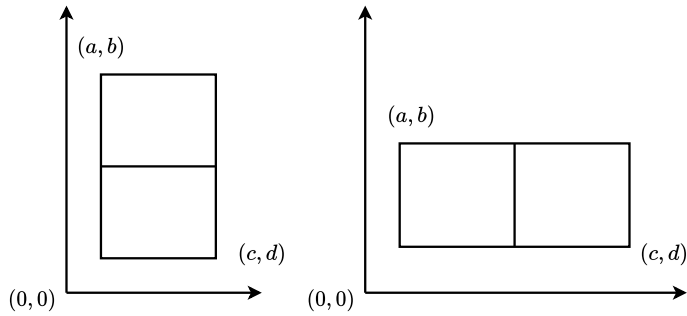


图 14.8: 沿较短的中线二分查找

中线一侧的区域。下面是改进的马鞍搜索：

1. 沿  $x, y$  轴二分搜索, 确定搜索区域  $(0, m) - (n, 0)$ ;
2. 若搜索区域  $(a, b) - (c, d)$  高大于宽, 沿水平中线二分查找; 否则沿中垂线二分查找, 结果为点  $(p, q)$ ;
3. 若  $f(p, q) = z$ ,  $(p, q)$  为一个解。递归搜索子区域  $(a, b) - (p - 1, q + 1)$  和  $(p + 1, q - 1) - (c, d)$ ;
4. 若  $f(p, q) \neq z$ , 递归搜索两个子区域外和一条线段。线段为  $(p, q + 1) - (p, b)$ , 如图 14.9 (a); 或  $(p + 1, q) - (c, q)$ , 如图 14.9 (b)。

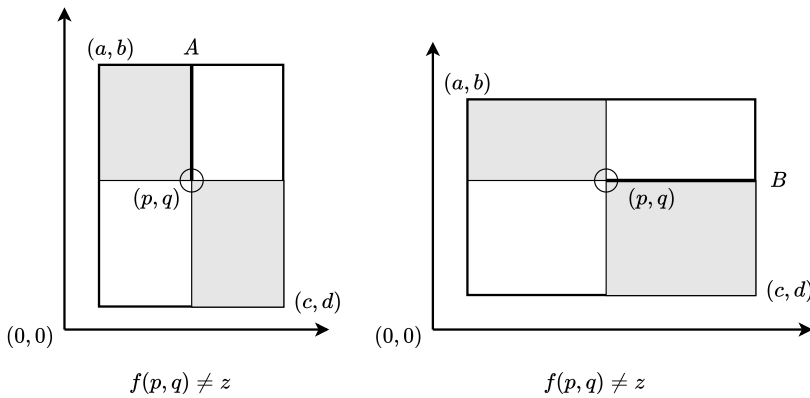


图 14.9: 递归搜索灰色的区域, 如果  $f(p, q) \neq z$ , 还需要搜索加粗的线段。

$$search(a, b)(c, d) = \begin{cases} c < a \text{ 或 } d < b : & [] \\ c - a < b - d : & csearch \\ \text{否则} : & rsearch \end{cases} \quad (14.13)$$

其中  $csearch$  在水平中线上二分查找  $(p, q)$  使得  $f(p, q) \leq z < f(p+1, q)$ , 如图 14.9 (a) 所示。如果中线上所有函数值大于  $z$ , 返回下界  $(a, \lfloor \frac{b+d}{2} \rfloor)$ 。丢弃中线上侧(含中线), 如图 14.10 (a) 所示。

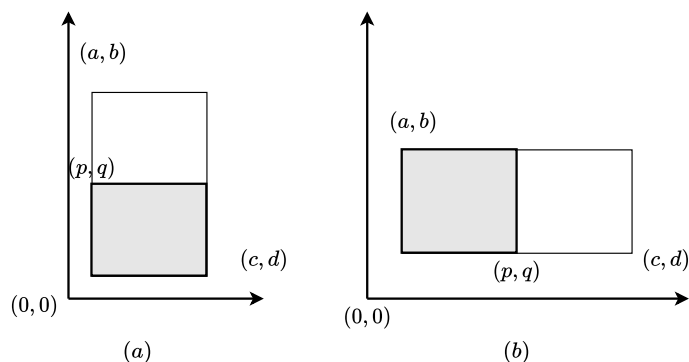


图 14.10: 沿中线二分查找时的特殊情况

令:

$$\begin{cases} q = \lfloor \frac{b+d}{2} \rfloor \\ p = bsearch(x \mapsto f(x, q)) z(a, c) \end{cases}$$

$$csearch = \begin{cases} f(p, q) > z : search(p, q-1)(c, d) \\ f(p, q) = z : search(a, b)(p-1, q+1) \# [(p, q)] \# search(p+1, q-1)(c, d) \\ f(p, q) < z : search(a, b)(p, q+1) \# search(p+1, q-1)(c, d) \end{cases} \quad (14.14)$$

$rsearch$  与此类似, 沿中垂线搜索。下面的例子程序实现了改进的马鞍搜索:

```
solve f z = search f z (0, m) (n, 0) where
  m = bsearch (f 0) z (0, z)
  n = bsearch (\x -> f x 0) z (0, z)

search f z (a, b) (c, d)
  | c < a || b < d = []
  | c - a < b - d = let q = (b + d) `div` 2 in
    csearch (bsearch (\x -> f x q) z (a, c), q)
  | otherwise = let p = (a + c) `div` 2 in
    rsearch (p, bsearch (f p) z (d, b))

where
  csearch (p, q)
    | z < f p q = search f z (p, q - 1) (c, d)
    | f p q == z = search f z (a, b) (p - 1, q + 1) #
      (p, q) : search f z (p + 1, q - 1) (c, d)
    | otherwise = search f z (a, b) (p, q + 1) #
      search f z (p + 1, q - 1) (c, d)
  rsearch (p, q)
    | z < f p q = search f z (a, b) (p - 1, q)
```

```

| f p q == z = search f z (a, b) (p - 1, q + 1) +
      (p, q) : search f z (p + 1, q - 1) (c, d)
| otherwise = search f z (a, b) (p - 1, q + 1) +
      search f z (p + 1, q) (c, d)

```

搜索区域每次减半, 总共搜索  $O(\lg(mn))$  轮。沿中线二分查找  $(p, q)$ , 需要计算  $f$  共  $O(\lg(\min(m, n)))$  次。令在  $m \times n$  区域搜索时间为  $T(m, n)$ , 有如下的递归关系:

$$T(m, n) = \lg(\min(m, n)) + 2T\left(\frac{m}{2}, \frac{n}{2}\right) \quad (14.15)$$

不妨设  $m = 2^i > n = 2^j$ , 使用裂项求和法:

$$\begin{aligned}
 T(2^i, 2^j) &= j + 2T(2^{i-1}, 2^{j-1}) \\
 &= \sum_{k=0}^{i-1} 2^k (j - k) \\
 &= O(2^i (j - i)) \\
 &= O(m \lg(n/m))
 \end{aligned} \quad (14.16)$$

伯德证明了这是在  $m \times n$  区域内搜索的最优下界<sup>[1]</sup>。

### 练习 14.1

- 14.1.1. 证明  $k$  选择的平均复杂度为  $O(n)$ 。
- 14.1.2. 为了查找  $A$  中的前  $k$  小元素, 我们可以获取  $x = \max(\text{take } k A), y = \min(\text{drop } k A)$ 。如果  $x < y$ , 则  $A$  的前  $k$  个元素就是答案; 否则我们用  $x$  划分前  $k$  个元素, 用  $y$  划分剩余元素, 然后在子序列  $[a \leftarrow A, x < a < y]$  中递归查找前  $k'$  个元素, 其中  $k' = k - \lfloor [a \leftarrow A, a \leq x] \rfloor$ 。请实现这一算法, 并分析复杂度。
- 14.1.3. 查找两个已序数组  $A$  和  $B$  的“简化中值”, 满足时间复杂度  $O(\lg(m+n))$ , 其中  $m = |A|, n = |B|$ , 分别为两个数组的长度。数组下标从 0 开始。简化中值定义为:  $\text{median}(A, B) = C[\lfloor \frac{m+n}{2} \rfloor]$ , 其中  $C = \text{merge}(A, B)$  为合并后的已序数组<sup>4</sup>。
- 14.1.4. 消除递归, 通过更新搜索边界实现马鞍查找。
- 14.1.5. 实现二维搜索算法并分析复杂度: 矩形搜索区域的左下角是最小值, 右上角是最大值。若搜索值  $z$  小于最小值或者大于最大值无解; 否则从中心划一个十字, 分割成 4 个小矩形, 然后递归搜索。

答案: 393 页

<sup>4</sup>简化定义与常见的统计学定义不同。后者对于含有  $n$  个元素升序序列  $x$  的中值定义为:

$$\text{median}(x) = \begin{cases} \text{odd}(n) : & x[\frac{n+1}{2}] \\ \text{even}(n) : & \frac{1}{2}(x[\frac{n}{2}] + x[\frac{n}{2} + 1]) \end{cases}$$

## 14.4 众数问题

人们常常投票并用计算机统计结果。某个小岛正在选举总统,只有赢得半数以上选票的候选人才可当选。从投票结果序列,如 A, B, A, C, B, B, D, ... 统计结果。能否高效找出当选总统? 可以利用字典数据结构遍历选票来解决(见第 2 章)<sup>6</sup>:

```
Optional<T> majority([T] xs) {
  Map<T, Int> m
  for var x in xs {
    if x in m then m[x]++ else mx[x] = 0
  }
  var (r, v) = (Optional<T>.Nothing, length(xs) / 2 - 1)
  for var (x, c) in m {
    if c > v then (r, v) = (Optional.of(x), c)
  }
  return r
}
```

可以用红黑树或散列表实现的字典数据结构,如果有  $m$  名候选人  $n$  张选票,这一实现的复杂度如下表:

| 字典  | 时间           | 空间        |
|-----|--------------|-----------|
| 红黑树 | $O(n \lg m)$ | $O(m)$    |
| 散列表 | $O(n)$       | 最少 $O(m)$ |

超过一半的元素叫做“众数”。鲍耶和摩尔在 1980 年给出了一种方法,可以扫描一遍找出众数(如果存在)。算法的时间复杂度为  $O(n)$ ,空间复杂度为  $O(1)$ ,被称做鲍耶——摩尔众数算法<sup>[83]</sup>。它建立在这样的想法上:众数最多只有 1 个,我们不断删掉两个不同的元素,直到最后剩下的元素都相等。如果众数存在,最后剩下的一定是众数。设第一张选票上的候选人为目前的获胜者,赢得的票数为 1。依次检查后继选票,若选票还投给目前的获胜者,则获胜者的得票加 1,否则获胜者得票减 1。若得票减到 0 则不再是获胜者了。选择下一张选票上的候选人为新获胜者并继续,如下表所示。若存在众数  $m$ ,则  $m$  不可能被其它元素超过落选。但如果众数不存在(选举无效,未决出优胜),则最后记录的“获胜者”并无意义。需要再扫描一轮进行验证。

<sup>6</sup>2004 年,人们发现了一种概率算法,称为 Count-min sketch 算法,使用 sub-linear 空间进行计数<sup>[84]</sup>。

| 获胜者 | 得票 | 扫描位置                                          |
|-----|----|-----------------------------------------------|
| A   | 1  | <b>A</b> , B, C, B, B, C, A, B, A, B, B, D, B |
| A   | 0  | A, <b>B</b> , C, B, B, C, A, B, A, B, B, D, B |
| C   | 1  | A, B, <b>C</b> , B, B, C, A, B, A, B, B, D, B |
| C   | 0  | A, B, C, <b>B</b> , B, C, A, B, A, B, B, D, B |
| B   | 1  | A, B, C, B, <b>B</b> , C, A, B, A, B, B, D, B |
| B   | 0  | A, B, C, B, B, <b>C</b> , A, B, A, B, B, D, B |
| A   | 1  | A, B, C, B, B, C, <b>A</b> , B, A, B, B, D, B |
| A   | 0  | A, B, C, B, B, C, A, <b>B</b> , A, B, B, D, B |
| A   | 1  | A, B, C, B, B, C, A, B, <b>A</b> , B, B, D, B |
| A   | 0  | A, B, C, B, B, C, A, B, A, <b>B</b> , B, D, B |
| B   | 1  | A, B, C, B, B, C, A, B, A, B, <b>B</b> , D, B |
| B   | 0  | A, B, C, B, B, C, A, B, A, B, B, <b>D</b> , B |
| B   | 1  | A, B, C, B, B, C, A, B, A, B, B, D, <b>B</b>  |

$$\begin{aligned} \text{maj } [ ] &= \emptyset \\ \text{maj } (x:xs) &= \text{scan } (x, 1) \text{ } xs \end{aligned} \quad (14.17)$$

其中  $\text{scan}$  定义为:

$$\begin{aligned} \text{scan } (m, v) [ ] &= m \\ \text{scan } (m, v) (x:xs) &= \begin{cases} m = x : & \text{scan } (m, v + 1) \text{ } xs \\ v = 0 : & \text{scan } (x, 1) \text{ } xs \\ \text{否则} : & \text{scan } (m, v - 1) \text{ } xs \end{cases} \end{aligned} \quad (14.18)$$

或者利用叠加实现:  $\text{maj} = \text{foldr } f (\emptyset, 0)$ , 其中:

$$f \ x \ (m, v) = \begin{cases} x = m : & (m, v + 1) \\ v = 0 : & (x, 1) \\ \text{否则} : & (m, v - 1) \end{cases} \quad (14.19)$$

最后还需要验证是否过半数:

$$\text{verify } m = \text{if } 2|\text{filter } (= m) \text{ } xs| > |xs| \text{ then } \text{Just } m \text{ else } \emptyset \quad (14.20)$$

下面是对应的迭代实现:

- 1: **function** MAJORITY( $A$ )
- 2:    $c \leftarrow 0, m \leftarrow \emptyset$
- 3:   **for** each  $a$  in  $A$  **do**

```

4:     if  $c = 0$  then
5:          $m \leftarrow a$ 
6:     if  $a = m$  then
7:          $c \leftarrow c + 1$ 
8:     else
9:          $c \leftarrow c - 1$ 
10:     $c \leftarrow 0$ 
11:    for each  $a$  in  $A$  do
12:        if  $a = m$  then
13:             $c \leftarrow c + 1$ 
14:    if  $c > \%50|A|$  then
15:        return  $x$ 
16:    else
17:        return  $\emptyset$ 

```

## 练习 14.2

14.2.1. 扩展众数算法, 在  $A$  中寻找出现次数超过  $\lfloor n/k \rfloor$  的  $k$  个众数, 其中  $n = |A|$ 。提示: 每次删掉  $k$  个不同元素, 直到最后剩下的元素种类不足  $k$  个。如果某个元素是  $k$ -众数(多于  $\lfloor n/k \rfloor$  个), 则一定会剩下来。

答案: 396 页

## 14.5 最大子序列和

序列  $V$  中的一段连续的部分  $V[i\dots j]$  叫做子序列。定义子序列和  $S = V[i] + V[i+1] + \dots + V[j]$ , 空序列  $[\ ]$  是任何序列的子序列, 它的和是 0。如何找到  $V$  的最大的子序列和<sup>[2]</sup>? 例如  $[3, -13, 19, -12, 1, 9, 18, -16, 15, -15]$  中, 子序列  $[19, -12, 1, 9, 18]$  的和最大, 为 35。如果序列中的元素都是正数, 最大和就是全部元素和。如果所有元素都是负数, 则空序列的和最大, 为 0。显然可以通过穷举找出答案:

```

1: function MAX-SUM( $V$ )
2:      $m \leftarrow 0, n \leftarrow |V|$ 
3:     for  $i \leftarrow 1$  to  $n$  do
4:          $s \leftarrow 0$ 
5:         for  $j \leftarrow i$  to  $n$  do
6:              $s \leftarrow s + V[j]$ 
7:              $m \leftarrow \text{MAX}(m, s)$ 
8:     return  $m$ 

```

穷举法的复杂度为  $O(n^2)$ , 其中  $n$  是序列长度。我们可以借鉴众数算法的思想, 一



边扫描一边记录下以当前位置  $i$  结尾的子序列的和  $A$ , 同时记录下目前所找到的最大子序列和  $B$ , 如图 14.11 所示。  $A, B$  不一定相等。总保持  $B \leq A$  的关系。当  $B$  和下一个元素  $V[i]$  相加后超过  $A$  时, 我们就用更大的  $B$  替换  $A$ 。当  $B + V[i] < 0$  时, 将  $B$  重置为 0。下表给出了扫描处理  $[3, -13, 19, -12, 1, 9, 18, -16, 15, -15]$  时的步骤。

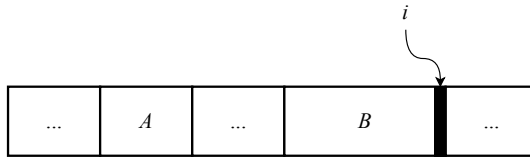


图 14.11:  $A$  是目前找到的最大子序列和,  $B$  是以  $i$  结尾的子序列和。

| 最大和 | 以 $i$ 结尾的子序列和 | 尚未扫描部分                                      |
|-----|---------------|---------------------------------------------|
| 0   | 0             | $[3, -13, 19, -12, 1, 9, 18, -16, 15, -15]$ |
| 3   | 3             | $[-13, 19, -12, 1, 9, 18, -16, 15, -15]$    |
| 3   | 0             | $[19, -12, 1, 9, 18, -16, 15, -15]$         |
| 19  | 19            | $[-12, 1, 9, 18, -16, 15, -15]$             |
| 19  | 7             | $[1, 9, 18, -16, 15, -15]$                  |
| 19  | 8             | $[9, 18, -16, 15, -15]$                     |
| 19  | 17            | $[18, -16, 15, -15]$                        |
| 35  | 35            | $[-16, 15, -15]$                            |
| 35  | 19            | $[15, -15]$                                 |
| 35  | 34            | $[-15]$                                     |
| 35  | 19            | $[\ ]$                                      |

```

1: function MAX-SUM( $V$ )
2:    $A \leftarrow 0, B \leftarrow 0, n \leftarrow |V|$ 
3:   for  $i \leftarrow 1$  to  $n$  do
4:      $B \leftarrow \text{MAX}(B + V[i], 0)$ 
5:      $A \leftarrow \text{MAX}(A, B)$ 
6:   return  $A$ 

```

我们也可以用叠加查找子序列最大和:  $S_{max} = fst \circ foldr f (0, 0)$ , 其中  $f$  更新最大和:

$$f x (S_m, S) = (S'_m = \max(S_m, S'), S' = \max(0, x + S)) \quad (14.23)$$

### 练习 14.3

- 14.3.1. 修改最大子序列和的实现,返回对应最大和的子序列。
- 14.3.2. 本特利<sup>[2]</sup>给出了一个分而治之的方法求子数组最大和。复杂度为  $O(n \log n)$ 。思路是将列表在中点分成两份。我们可以递归地找出前半部分的最大和,和后半部分的最大和,和跨越中点部分的最大和。实现这一方法。
- 14.3.3. 在  $m \times n$  的二维整数矩阵中寻找子矩阵,使得各元素相加后的和最大。

答案: 397 页

## 14.6 字符串搜索

字符串搜索是一类实用问题。所有文本编辑软件都带有字符串搜索功能。在基数树、前缀树中(第 6 章),我们利用特殊的数据结构进行搜索。我们也可以直接在字符串序列中进行搜索,如图 14.12 所示<sup>7</sup>。

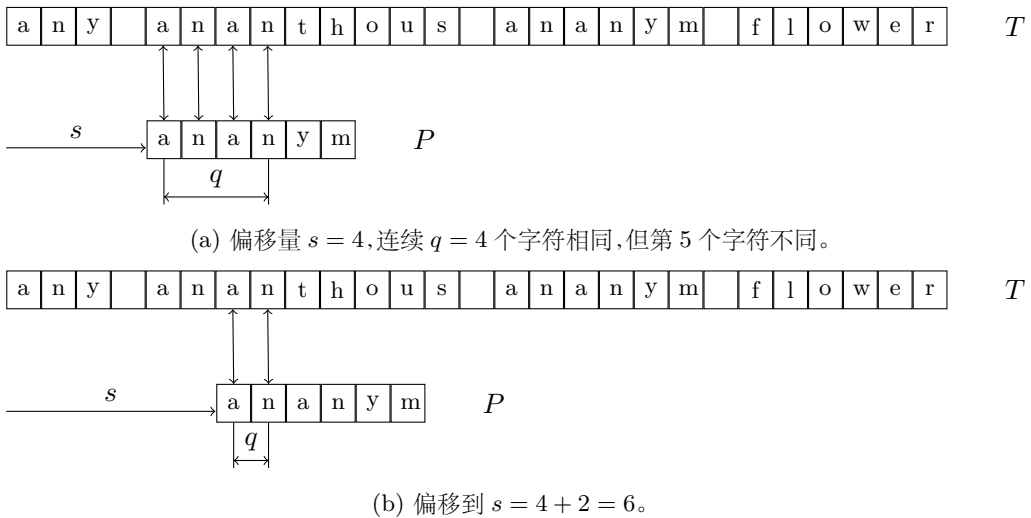


图 14.12: 在文本“any ananthous anany m flower”中寻找“anany m”

在文本  $T$  中搜索字符串  $P$ 。如图 14.12 (a) 所示,在偏移量  $s = 4$  时,逐一检查  $P$  和  $T$  中的字符。前 4 个相等,第 5 个在  $P$  中是  $y$ ,但在  $T$  中是  $t$ 。此时逐一比较立即终止,将  $s$  加 1 ( $P$  向右移动 1 个位置)然后重新比较  $anany m$  和  $nantho \dots$ 。我们发现  $s$  的增量可以超过 1。前两个字符  $an$  恰好是  $anan$  的后缀。我们可以将  $s$  增加 2 ( $P$  向右移动 2),如图 14.12 (b) 所示。我们复用了此前已经比较过 4 个字符的信息,跳过了无需比较的位置。高德纳、莫里斯、普拉特根据这一想法给出了一个高效的字符串匹配算法<sup>[85]</sup>,人们把三人名字的首字母合在一起,称作 KMP 算法。

记文本  $T$  中前  $k$  个字符组成的串为  $T_k$  ( $T$  的  $k$  个字符前缀)。为了尽可能多地把  $P$  向右移动  $s$  个位置,我们需要利用已成功匹配的  $q$  个字符。如图 14.13 所示,若  $P$

<sup>7</sup>编程环境一般会提供一些逐一搜索工具,如 C 标准库中的 `strstr`, C++ 标准库中的 `find`,以及 Java 标准库中的 `indexOf`。

可向右移动,则一定存在某个  $k$ ,使得  $P$  中的前  $k$  个字符和前缀  $P_q$  的最后  $k$  个字符相同。也就是说,前缀  $P_k$  同时是  $P_q$  的后缀。定义空串“”是任何串的前缀和后缀,则总存在最小的  $k = 0$ 。我们要找到同时既是前缀又是后缀的最大  $k$ 。定义前缀函数  $\pi(q)$ ,它告诉我们当第  $q + 1$  个字符不匹配时应该回退的位置<sup>[4]</sup>。

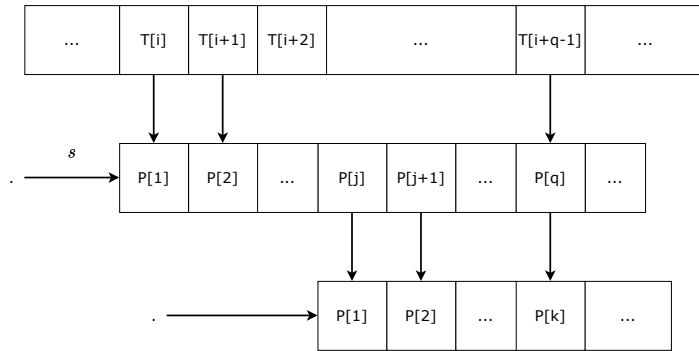


图 14.13:  $P_k$  同时是  $P_q$  的前缀和后缀

$$\pi(q) = \max\{k | 0 \leq k < q, \text{且 } P_k \text{ 是 } P_q \text{ 的后缀}\} \quad (14.24)$$

当在文本  $T$  中偏移  $s$  匹配  $P$  时,若前  $q$  个字符相同,而下一个不同,我们通过  $q' = \pi(q)$  找到一个回退的位置  $q'$ 。重新比较  $P[q']$  和文本:

```

1: function KMP( $T, P$ )
2:    $\pi \leftarrow \text{BUILD-PREFIXES}(P)$ 
3:    $n \leftarrow |T|, m \leftarrow |P|, q \leftarrow 0$ 
4:   for  $i \leftarrow 1$  to  $n$  do
5:     while  $q > 0$  且  $P[q + 1] \neq T[i]$  do
6:        $q \leftarrow \pi(q)$ 
7:     if  $P[q + 1] = T[i]$  then
8:        $q \leftarrow q + 1$ 
9:     if  $q = m$  then
10:      位置  $i - m$  是一个解
11:       $q \leftarrow \pi(q)$ 

```

▷ 寻找更多的匹配位置

直接利用式 (14.24) 构造  $\pi(q)$  并不实用。我们进一步复用信息,构造前缀函数。如果第一个字符就不匹配,此时最长前缀同时也是后缀的显然是空串:  $\pi(1) = 0$ 。即:  $P_k = P_0 = []$ 。当扫描到  $P$  中第  $q$  个字符时,前缀函数值  $\pi(i), i = 1, 2, \dots, q - 1$  都已算好,并且目前最长的前缀  $P_k$  同时也是  $P_{q-1}$  的后缀。如图 14.14 所示,若  $P[q] = P[k + 1]$ ,则找到了一个更大的  $k$ ,我们将  $k$  的最大值加一;否则  $P[q] \neq P[k + 1]$ ,我们利用  $\pi(k)$  回退到一个较短的  $P_{k'}$ ,其中  $k' = \pi(k)$ ,然后比较这个新前缀的下一个字符是否和第  $q$  个字符相等。重复这一步骤,直到  $k$  变成 0(空串),或者和第  $q$  个字符相等。下表给出了 ananym 的前缀函数值, $k$  列是满足式 (14.24) 的最大值。

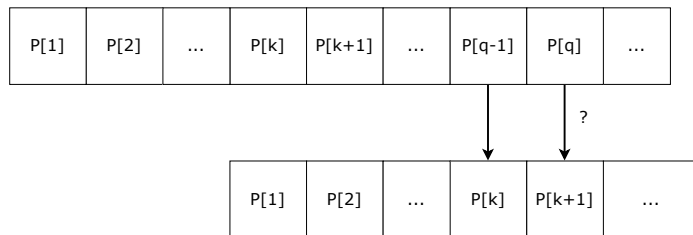


图 14.14:  $P_k$  是  $P_{q-1}$  的后缀, 比较  $P[q]$  和  $P[k+1]$

| $q$ | $P_q$  | $k$ | $P_k$ |
|-----|--------|-----|-------|
| 1   | a      | 0   | ""    |
| 2   | an     | 0   | ""    |
| 3   | ana    | 1   | a     |
| 4   | anan   | 2   | an    |
| 5   | anany  | 0   | ""    |
| 6   | anonym | 0   | ""    |

```

1: function BUILD-PREFIXES( $P$ )
2:    $m \leftarrow |P|, k \leftarrow 0$ 
3:    $\pi(1) \leftarrow 0$ 
4:   for  $q \leftarrow 2$  to  $m$  do
5:     while  $k > 0$  且  $P[q] \neq P[k+1]$  do
6:        $k \leftarrow \pi(k)$ 
7:     if  $P[q] = P[k+1]$  then
8:        $k \leftarrow k + 1$ 
9:      $\pi(q) \leftarrow k$ 
10:  return  $\pi$ 

```

KMP 预处理待搜索串, 构建前缀函数的分摊复杂度为  $O(m)$ <sup>[4]</sup>。搜索本身的分摊复杂度是  $O(n)$ 。总体分摊复杂度为  $O(m+n)$ , 并需要  $O(m)$  空间记录前缀函数值。待搜索串的形式并不影响 KMP 算法的性能。考虑在“aaa...a” ( $n$  个) 中搜索长为  $m$  的子串“aaa...ab”。第  $m$  个字符不匹配, 我们只能回退一个字符, 并且此后不断回退 1 个字符。即使在这种情况下, KMP 算法仍然是线性时间的。

## 14.7 附录: 例子程序

查找前  $k$  小元素:

```

Optional<K> top(Int k, [K] xs, Int l, Int u) {
  if  $l < u$  {

```

```

    swap(xs, l, rand(l, u))
    var p = partition(xs, l, u)
    if p - l + 1 == k
        return Optional.of(xs[p])
    return if k < p - l + 1 then top(k, xs, l, p)
        else top(k - p + l - 1, xs, p + 1, u)
}
return Optional.Nothing
}

Int partition([K] xs, Int l, Int u) {
    var p = l
    for var r = l + 1 to u {
        if not xs[p] < xs[r] {
            l = l + 1
            swap(xs, l, r)
        }
    }
    swap(xs, p, l)
    return l
}
}

```

马鞍搜索:

```

solve f z = search 0 m where
    search p q | p > n || q < 0 = []
                | z' < z = search (p + 1) q
                | z' > z = search p (q - 1)
                | otherwise = (p, q) : search (p + 1) (q - 1)
    where z' = f p q
m = bsearch (f 0) z (0, z)
n = bsearch (\x→f x 0) z (0, z)

bsearch f y (l, u) | u ≤ l = l
                  | f m ≤ y = if f (m + 1) ≤ y then bsearch f y (m + 1, u) else m
                  | otherwise = bsearch f y (l, m - 1)
where m = (l + u) `div` 2

```

众数查找:

```

Optional<T> majority([T] xs) {
    var (m, c) = (Optional<T>.Nothing, 0)
    for var x in xs {
        if c == 0 then (m, c) = (Optional.of(x), 0)
        if x == m then c++ else c--
    }
    c = 0
    for var x in xs {
        if x == m then c++
    }
    return if c > length(xs)/2 then m else Optional<T>.Nothing
}
}

```

用叠加实现的众数查找:

```
majority xs = verify $ foldr maj (Nothing, 0) xs where
  maj x (Nothing, 0) = (Just x, 1)
  maj x (Just y, v) | x == y = (Just y, v + 1)
                    | v == 0 = (Just x, 1)
                    | otherwise = (Just y, v - 1)
  verify (Nothing, _) = Nothing
  verify (Just m, _) = if 2 * (length $ filter (==m) xs) > length xs
                       then Just m else Nothing
```

最大子序列和:

```
maxSum :: (Ord a, Num a) => [a] -> a
maxSum = fst o foldr f (0, 0) where
  f x (m, mSofar) = (m', mSofar') where
    mSofar' = max 0 (mSofar + x)
    m' = max mSofar' m
```

KMP 字符串匹配算法:

```
[Int] match([T] w, [T] p) {
  n = length(w), m = length(p)
  [Int] fallback = prefixes(p)
  [Int] r = []
  Int k = 0
  for i = 0 to n {
    while k > 0 and p[k] ≠ w[i] {
      k = fallback[k]
    }
    if p[k] == w[i] then k = k + 1
    if k == m {
      add(r, i + 1 - m)
      k = fallback[k - 1]
    }
  }
  return r
}

[Int] prefixes([T] p) {
  m = length(p)
  [Int] t = [0] * m //fallback table
  Int k = 0
  for i = 2 to m {
    while k > 0 and p[i-1] ≠ p[k] {
      k = t[k-1] #fallback
    }
    if p[i-1] == p[k] then k = k + 1
    t[i] = k
  }
  return t
}
```

# 第十五章 搜索 II

在人工智能的早期阶段,人们发展出了许多方法搜索解。不同于字符串匹配,解并不一定直接存在于某个候选集中。往往需要一边构造解一边尝试。某些问题可解,某些问题无解,也可能存在多个解(例如存在多条走出迷宫的路线)。人们还要求出某种意义下的最优解。

## 15.1 深度优先和广度优先搜索

深度优先搜索(DFS)和广度优先搜索(BFS)常被叫做“图搜索”算法。我们简单介绍这两种搜索算法而略过图的概念。

### 15.1.1 迷宫

走迷宫是一类历史悠久的趣题。有窍门说:“当遇到分叉时总向右转。”这个方法有缺陷。如图 15.1(b) 所示,不断向右会绕着中心的大方块转圈。存在多个选项时,决策直接影响到最终的解。童话故事提供了一种思路:携带一块面包进入迷宫。遇到岔路时任选一条道路,并留下一小块面包屑记录下这次尝试。遇到了死路就沿着留下的面包屑走回上次选择的地方,然后换一条路。一旦发现地上已经有面包屑了,就说明进入了循环,必须进一步回退尝试。不断这样的“尝试——检查”,要么最终走出迷宫,要么退回起点(无解)。我们用  $m \times n$  的矩阵  $M$  描述迷宫,元素值为 0、1,表示是否有路。图 15.1(b) 的迷宫可以用下面的矩阵定义:

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

给定起点  $s = (i, j)$ 、终点  $e = (p, q)$ ,我们要找出所有从  $s$  到  $e$  的通路。我们先找出所有和  $s$  连通的相邻点,对每个邻点  $k$ ,递归找出从  $k$  到  $e$  所有路径。然后将通路

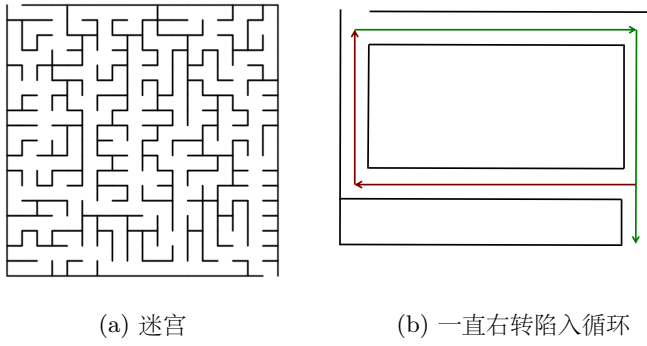


图 15.1: 迷宫

$s$ - $k$  连接到每个从  $k$  到  $e$  的路径前。我们需要留下一些“面包屑”标记以避免重复。用一个列表  $P$  记录走过的所有位置。每次检查此列表,只尝试新路径。

$$\text{solveMaze } M \ s \ e = \text{solve } s \ [[]] \quad (15.1)$$

其中:

$$\text{solve } s \ P = \begin{cases} s = e : \text{map } (\text{reverse} \circ (s :)) \ P \\ \text{否则} : \text{concat } [\text{solve } k \ (\text{map } (s :) \ P) | k \leftarrow \text{adj } s, k \notin P] \end{cases} \quad (15.2)$$

$P$  中保存的路径是逆序的,最后需要用  $\text{reverse}$  反转。 $\text{adj } p$  找出  $p$  相连通的邻点: 水平、垂直方向上相邻为 0 的点:

$$\text{adj } (x, y) = [(x', y') \leftarrow [(x - 1, y), (x + 1, y), (x, y - 1), (x, y + 1)], \\ 1 \leq x' \leq m, 1 \leq y' \leq n, M_{x'y'} = 0] \quad (15.3)$$

这是一种穷举解法,搜索所有路径。为了走出迷宫,只需要找到一条路径。我们需要某种数据结构保存“面包屑”,记录此前做出的决策。我们总在最新决策基础上搜索通路,这是后进先出的顺序。可以用一个栈来实现。开始时,栈中只有起点  $[s]$ 。将其弹出,找出和  $s$  连通的点,例如  $a, b, \dots$  将可能的路径  $[a, s], [b, s]$  入栈。接下来将路径  $[a, s]$  弹出,检查和  $a$  连通的点。然后把所有 3 步可达的路径入栈。重复这一过程。栈中记录了一条条逆序路径:从可达的最远位置反向到起点,如图 15.2 所示。当栈变为空时,我们已尝试了所有的可能,搜索结束;否则弹出栈顶的候选路径,扩展未曾走过的连通邻点,然后入栈。

$$\text{solveMaze } M \ s \ e = \text{solve } [[s]] \quad (15.4)$$

其中:



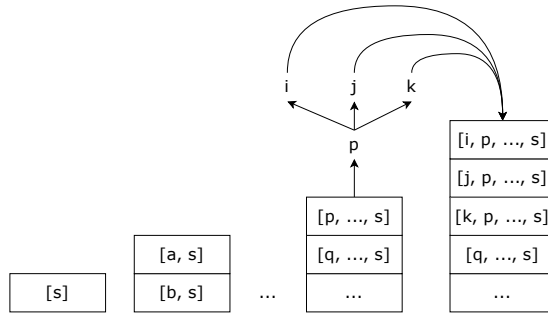


图 15.2: 用栈记录路径

$$\begin{aligned}
 \text{solve } [] &= [] \\
 \text{solve } ((p:ps):cs) &= \begin{cases} c = e : & \text{reverse } (p:ps) \\ ks = [] : & \text{solve } cs, \text{ 其中 } ks = \text{filter } (\neq ps) (\text{adj } p) \\ ks \neq [] : & \text{solve } ((\text{map } (: p:ps) ks) \# cs) \end{cases}
 \end{aligned} \tag{15.5}$$

对应的迭代实现如下:

```

1: function SOLVE-MAZE( $M, s, e$ )
2:    $S \leftarrow [s], L = []$ 
3:   while  $S \neq []$  do
4:      $P \leftarrow \text{POP}(S)$ 
5:      $p \leftarrow \text{LAST}(P)$ 
6:     if  $e = p$  then
7:        $\text{ADD}(L, P)$  ▷ 找到一个解
8:     else
9:       for each  $k$  in  $\text{ADJACENT}(M, p)$  do
10:        if  $k \notin P$  then
11:           $\text{PUSH}(S, P \# [k])$ 
12:   return  $L$ 

```

每步有上下左右 4 个选项可能入栈, 并且都在回溯时检查。复杂度看似是  $O(4^n)$ , 其中  $n$  是路径长度。实际复杂度不会这样大, 我们跳过了已走的位置。最坏情况下, 所有可达点都恰好被访问过一次, 复杂度为  $O(n)$ , 其中  $n$  是连通点的数量。由于使用了栈来保存路径, 空间复杂度为  $O(n^2)$ 。

### 练习 15.1

15.1.1. 使用栈来找出迷宫问题的所有解。

答案: 399 页

### 15.1.2 八皇后问题

虽然国际象棋历史悠久,但直到 1848 年贝策尔 (Max Bezzel) 才提出八皇后趣题<sup>[89]</sup>。皇后威力巨大,可以攻击同一行、列、斜线上的任意棋子。如何在棋盘上摆下八个皇后,而她们之间不互相攻击。图 15.3 (a) 描述了皇后可以攻击到的范围。图 15.3 (b) 给出了八皇后问题的某一种解。

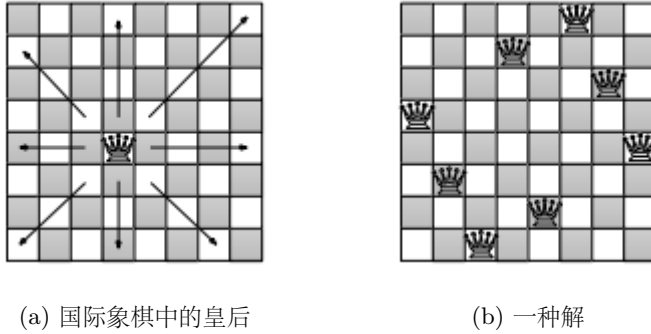


图 15.3: 八皇后问题

在 64 个格子中放入 8 个皇后,共有  $P_{64}^8$  种排列,约为  $4 \times 10^{10}$ 。每行、列只能摆放 1 个皇后,一个解的布局必然是  $[1, 2, 3, 4, 5, 6, 7, 8]$  的某种排列。例如布局  $[6, 2, 7, 1, 3, 5, 8, 4]$  表示第一行的皇后放在第 6 列;第二行的皇后摆在第 2 列上……第 8 行的皇后摆在第 4 列。这样我们只需要检查  $8! = 40320$  种布局。从第一行开始逐一摆放皇后。第一个皇后有 8 种摆法(八列中的某一列)。摆放第二行的皇后时,由于可能和第一个皇后相互攻击,需要避开某些列。对于第  $i$  行的皇后,找到 8 列中不被前  $i - 1$  个皇后攻击的位置。如果 8 个位置都不能摆放,我们就回退调整以前  $i - 1$  个皇后。全部 8 个皇后都成功放入棋盘后,就找到了一个解。为了找出所有解,我们记录下这一布局,然后继续检查其它可能的列并进行必要的回溯。我们用一个栈和一个列表启动搜索: `solve [[]] []`

$$\begin{aligned} \text{solve } [] \ s &= s \\ \text{solve } (c:cs) \ s &= \begin{cases} |c| = 8: & \text{solve } cs \ (c:s) \\ \text{否则}: & \text{solve } ([x:c|x \leftarrow [1..8], x \notin c, \text{safe } x \ c] \# cs) \ s \end{cases} \end{aligned} \quad (15.6)$$

若栈为空,我们已经尝试所有可能,  $s$  记录了所有找到的解;若栈顶布局  $c$  长度为 8,我们找到了一个解。将其记录到  $s$  中,然后继续查找;若  $|c| < 8$ ,我们从 8 列中找出尚未被占的列 ( $x \notin c$ ),同时不能被斜线上的其它皇后攻击(通过 `safe x c`)。可行的布局入栈用于此后的搜索。

$$\text{safe } x \ c = \forall (i, j) \leftarrow \text{zip } (\text{reverse } c) \ [1, 2, \dots] \text{ 有: } |x - i| \neq |y - j|, \text{ 其中: } y = 1 + |c| \quad (15.7)$$

*safe* 检查第  $y = 1 + |c|$  行,  $x$  列的皇后是否和  $c$  中任何皇后成对角线。若  $c = [i_{y-1}, i_{y-2}, \dots, i_1]$  是前  $y - 1$  个皇后所在的列, 我们将  $c$  反转, 并和  $1, 2, \dots$  组成每个皇后的坐标:  $[(i_1, 1), (i_2, 2), \dots, (i_{y-1}, y - 1)]$ 。然后判断每个  $(i, j)$  是否和位置  $(x, y)$  构成对角线:  $|x - i| \neq |y - j|$ 。这一实现是尾递归的, 可以消除递归优化为迭代实现:

```

1: function SOLVE-QUEENS
2:    $S \leftarrow [[]]$ 
3:    $L \leftarrow []$  ▷ 保存解
4:   while  $S \neq []$  do
5:      $A \leftarrow \text{POP}(S)$  ▷  $A$  是某一中间布局
6:     if  $|A| = 8$  then
7:        $\text{ADD}(L, A)$ 
8:     else
9:       for  $i \leftarrow 1$  to 8 do
10:        if  $\text{VALID}(i, A)$  then
11:           $\text{PUSH}(S, A \# [i])$ 
12:   return  $L$ 

13: function VALID( $x, A$ )
14:    $y \leftarrow 1 + |A|$ 
15:   for  $i \leftarrow 1$  to  $|A|$  do
16:     if  $x = A[i]$  或  $|y - i| = |x - A[i]|$  then
17:       return False
18:   return True

```

虽然每个皇后有 8 列选择, 但只尝试尚未被占的列。总共检查 15720 种情况, 小于  $8! = 40320$  种可能<sup>[89]</sup>。由于正方形棋盘水平、垂直都对称, 找到一个解后, 通过旋转、翻转可以得到其它对称解。我们可以扩展到  $n$  皇后问题, 其中  $n \geq 4$ 。但随着  $n$  增大所用时间急速增加。回溯算法仅比枚举全排列稍快(枚举全排列的时间是  $O(n!)$ )。

## 练习 15.2

15.2.1. 改进八皇后的算法, 使得它可以解决  $n$  皇后问题。

15.2.2. 八皇后问题存在 92 个不同的解。对于任何一个解, 将其旋转  $\pm 90^\circ$  也都是八皇后问题的解。并且翻转也是解。真正不同的解只有 12 个。修改八皇后的程序, 找出这 12 个不同的解。

答案: 400 页

### 15.1.3 跳棋趣题

如图 15.4, 在一排 7 块石头上有 6 只青蛙。一块石头只够一只青蛙停在上。青蛙可以跳到它前方的石头上, 或越过一只青蛙跳到更前方的石头上。青蛙只能前进或停止, 不能后退, 如图 15.5。这些青蛙如何移动、跳跃, 使得左右 3 只位置互换? 标记左侧青蛙为 -1, 右侧为 1, 没有青蛙的石头为 0, 我们要找到从  $s = [-1, -1, -1, 0, 1, 1, 1]$  转换到  $e = [1, 1, 1, 0, -1, -1, -1]$  的解。

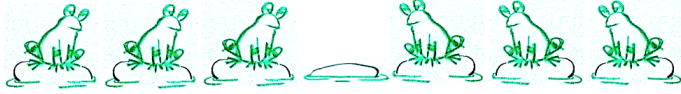


图 15.4: 跳跃的青蛙

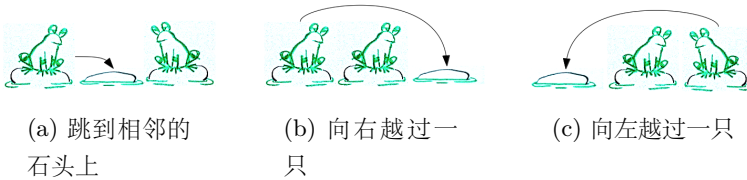


图 15.5: 移动规则

这是跳棋的一种特殊形式。并不一定限制为 6 颗棋子, 也可以是 8 或者更大的偶数。图 15.6 是这类问题的变化形式<sup>1</sup>。从左向右的石头位置为 1, 2, ..., 7。每次有 4 种移动可能。例如开始时, 第 3 块石头上的青蛙可以移动到空石头上。对称地, 第 5 块石头上的青蛙也可以左移一步。第 2 块石头上的青蛙可以向右越过一只青蛙跳到空石头上, 对称地, 第 6 块石头上的青蛙, 也可以向左越过一只。每次记录下 7 块石头和青蛙的状态, 尝试 4 种方案。如果走不下去了, 就回溯尝试其它方案。由于左侧的青蛙只能向右, 右侧的只能向左, 所有移动都是不可逆的。和走迷宫不同, 这里不存在重复情况。我们记录移动的步数用于最后输出结果。状态  $L$  是  $s$  的某种排列。  $L[i]$  的值是 0,  $\pm 1$  表示第  $i$  个石头是空的、有一只向左、向右的青蛙。令空石头的位置为  $p$ , 4 种移动方案为:

1. 向左跳跃:  $p < 6$ , 且  $L[p+2] > 0$ , 交换  $L[p] \leftrightarrow L[p+2]$ ;
2. 向左移动:  $p < 7$ , 且  $L[p+1] > 0$ , 交换  $L[p] \leftrightarrow L[p+1]$ ;
3. 向右跳跃:  $p > 2$ , 且  $L[p-2] < 0$ , 交换  $L[p-2] \leftrightarrow L[p]$ ;

<sup>1</sup>图片来源: <http://www.robspuzzlepage.com/jumping.htm>

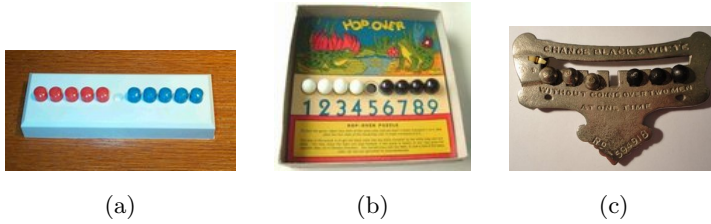


图 15.6: 三种跳棋趣题

4. 向右移动:  $p > 1$ , 且  $L[p-1] < 0$ , 交换  $L[p-1] \leftrightarrow L[p]$ 。

定义 4 个函数  $leap_l$ 、 $hop_l$ 、 $leap_r$ 、 $hop_r$ , 变化状态  $L \mapsto L'$ 。若不能移动, 则返回同样的  $L$  不变。使用栈  $S$  记录已做过的尝试。开始的时候, 栈中只有一个列表, 列表中只有开始状态。列表  $M$  记录所有找到的解。我们不断取出栈顶。如果状态  $L = e$  则找到了一个解。我们将移动步骤记录到  $M$  中。否则我们在  $L$  上尝试 4 种移动, 如果可行就入栈以备继续搜索。

$$\text{solve } [[-1, -1, -1, 0, 1, 1, 1]] [] \quad (15.8)$$

其中:

$$\begin{aligned} \text{solve } [] s &= s \\ \text{solve } (c:cs) s &= \begin{cases} L = e : \text{solve } cs(\text{reverse } c : s), \text{ 其中 } L = \text{head } c \\ \text{否则} : \text{solve } ((\text{map } (:c) (\text{moves } L)) \# cs) s \end{cases} \end{aligned} \quad (15.9)$$

$\text{moves}$  在状态  $L$  之上尝试 4 中可能的移动:

$$\text{moves } L = \text{filter}(\neq L) [\text{leap}_l L, \text{hop}_l L, \text{leap}_r L, \text{hop}_r L] \quad (15.10)$$

对应的迭代实现如下:

```

1: function SOLVE(s, e)
2:   S ← [[s]]
3:   M ← []
4:   while S ≠ [] do
5:     s ← POP(S)
6:     if s[1] = e then
7:       ADD(M, REVERSE(s))
8:     else
9:       for each m in MOVES(s[1]) do
10:        PUSH(S, m:s)
11:   return M

```

这一方法找出两个左右对称解(各 15 步), 下表列出了一个:

| 步骤 | -1 | -1 | -1 | 0  | 1  | 1  | 1  |
|----|----|----|----|----|----|----|----|
| 1  | -1 | -1 | 0  | -1 | 1  | 1  | 1  |
| 2  | -1 | -1 | 1  | -1 | 0  | 1  | 1  |
| 3  | -1 | -1 | 1  | -1 | 1  | 0  | 1  |
| 4  | -1 | -1 | 1  | 0  | 1  | -1 | 1  |
| 5  | -1 | 0  | 1  | -1 | 1  | -1 | 1  |
| 6  | 0  | -1 | 1  | -1 | 1  | -1 | 1  |
| 7  | 1  | -1 | 0  | -1 | 1  | -1 | 1  |
| 8  | 1  | -1 | 1  | -1 | 0  | -1 | 1  |
| 9  | 1  | -1 | 1  | -1 | 1  | -1 | 0  |
| 10 | 1  | -1 | 1  | -1 | 1  | 0  | -1 |
| 11 | 1  | -1 | 1  | 0  | 1  | -1 | -1 |
| 12 | 1  | 0  | 1  | -1 | 1  | -1 | -1 |
| 13 | 1  | 1  | 0  | -1 | 1  | -1 | -1 |
| 14 | 1  | 1  | 1  | -1 | 0  | -1 | -1 |
| 15 | 1  | 1  | 1  | 0  | -1 | -1 | -1 |

每侧有 3 只青蛙时需要 15 步左右互换。扩展上述解法可以得到步数和青蛙数目的一个关系:

| 每侧青蛙数 $n$ | 1 | 2 | 3  | 4  | 5  | ... |
|-----------|---|---|----|----|----|-----|
| 解法步数      | 3 | 8 | 15 | 24 | 35 | ... |

步数恰好是完全平方数减一:  $(n+1)^2 - 1$ 。我们可以证明这一结论:

**证明.** 比较最终和起始状态, 每只青蛙都向另一侧移动了  $n+1$  块石头。  $2n$  只青蛙总共移动了  $2n(n+1)$  块石头。左侧的每只青蛙必然和右侧的所有青蛙相遇一次。一旦相遇, 必然发生跳跃。由于一共有  $n^2$  次相遇, 因此共导致了所有青蛙前进了  $2n^2$  块石头。剩下的移动不是跳跃, 而是跳到相邻的石头上, 总共有  $2n(n+1) - 2n^2 = 2n$  次。将  $n^2$  次跳跃, 和  $2n$  次跳到相邻石头上相加。得到最终解的步数为:  $n^2 + 2n = (n+1)^2 - 1$ 。 □

观察上面 3 个趣题, 它们的解法有着类似的结构。它们都从某种状态开始。走迷宫从入口开始, 八皇后问题从空棋盘开始, 跳跃青蛙问题从  $[-1, -1, -1, 0, 1, 1, 1]$  开始。解的过程是一种搜索。每次尝试都有若干种可能的选项。走迷宫时, 每步有上下左右四个方向选择; 八皇后问题中, 每次摆放都有八列选择; 跳跃青蛙趣题中, 每次有 4 种不同的跳跃方式选择。虽然每次选择都不知道能继续走多远。但我们始终清楚地知道最终状态是什么。走迷宫的最终状态是出口; 八皇后问题的最终状态是八个皇后都摆在棋盘上; 跳跃青蛙趣题的最终状态是左右青蛙位置互换。

我们使用相同的策略来解决这些问题: 不断尝试可能的选项, 记录已经达到的状态, 如果无法继续就回溯并尝试其它选项。通过这样的方法, 我们或者找到解, 或者穷尽所有可能而发现问题无解。当然, 这类解法存在一些变化, 当找到一个解后, 我们可以停下结束或者继续寻找所有可能的解。如果以起始状态为根, 画出一棵树, 每个树枝代表不同的选择。搜索过程是一个不断深入的过程。只要能继续, 我们先不考虑同一深度上的其它选项。直到失败后回溯到树的上一层。如图 15.7 中的搜索顺序。箭头描述了如何先向下, 再向上回溯的过程。节点上的数字是访问顺序。

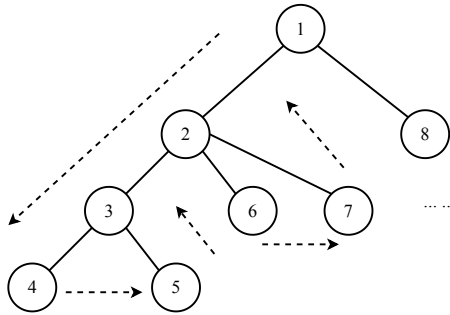


图 15.7: 深度优先搜索的顺序

这样的搜索策略称为深度优先搜索。现实中我们在不经意间广泛使用它。某些编程环境, 例如 Prolog, 使用深度优先作为默认的求值模型。例如一个迷宫可以被一组规则描述:

```
c(a, b). c(a, e).
c(b, c). c(b, f).
c(e, d), c(e, f).
c(f, c).
c(g, d). c(g, h).
c(h, f).
```

其中断言  $c(X, Y)$  表示位置  $X$  和  $Y$  连通。断言是有方向性的。如果要让  $Y$  和  $X$  连通, 我们可以增加一条对称的断言, 或者建立一条无方向性的断言。图 15.8 给出了一个有向图。任给位置  $X$  和  $Y$ , Prolog 可以通过下面的程序判定它们之间是否有通路。

```
go(X, X).
go(X, Y) :- c(X, Z), go(Z, Y)
```

这一程序说: 一个位置  $X$  和自己相通。任意两个不同位置  $X, Y$ , 若  $X$  和  $Z$  相连, 且  $Z$  和  $Y$  之间有通路, 则  $X$  和  $Y$  之间存在通路。显然,  $Z$  的选择可能不唯一。Prolog 选择一个, 然后继续搜索。只有当递归搜索失败时才会尝试其它选择。回溯并更换到下一个选项上。这恰好就是深度优先的搜索策略。当只需要找到解, 而并不关心步数时, 深度优先搜索是很有效的方法。例如, 走迷宫中找出的第一个解并不一定是最短路径。

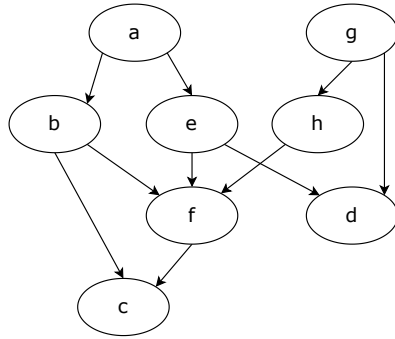


图 15.8: 一个有向图

### 练习 15.3

15.3.1. 修改跳跃青蛙问题的函数式解法,使得它可以解决每侧  $n$  只青蛙的情况。

答案: 401 页

#### 15.1.4 狼、羊、白菜过河问题

这是一道传统趣题。农夫带着一只狼、一只羊、一筐白菜过河。有一条小船,只有农夫会划船。小船只能装下农夫和另外一样东西。农夫每次只能在狼、羊、白菜中任选一样和他一起过河。但是如果农夫不在,狼会吃掉羊,而羊会吃掉白菜。如何用最快的方法让所有东西都渡过河?

由于狼不会吃掉白菜,农夫可以安全地将羊运到河对岸并返回。接下来无论将狼或白菜中的任何一样运过河,他必须将某一样运回以避免有东西被吃掉。为了寻找最快的渡河方法,我们并发检查所有的选项,比较哪个更快。不考虑渡河的方向,每渡过一次算做一步,往返算两步。我们检查渡河一次后的所有可能、两次后所有可能、三次后的所有可能……直到某次后,所有的东西都到达了对岸结束。并且这一渡河方法在所有可能中胜出,是最快的解法。

如何“并发”检查所有可能的解法?考虑一个抽奖游戏。参与者闭着眼睛从箱子里摸出一个球。箱子里只有一个黑球,其余的球都是白色的。摸到黑球获胜,摸到白球则放回箱子,然后等待下次摸球。为了使游戏公平,可以制定这样一个规则:必须等所有人都摸过之后才能再摸第二次。参与游戏的人站成一队。每次站在队伍前面的人摸球,如果他没有摸到黑球,就站到队尾等待下次摸球。这一队列可以保证游戏的公平。

用类似的思路来解决渡河问题。用集合  $A$ 、 $B$  代表河两岸。开始时,集合  $A = \{w, g, c, p\}$  包含狼、羊、白菜、农夫,集合  $B = \emptyset$ 。每次将农夫和另外一个元素在集合间移动。如果集合中不存在农夫,则不能含有相互冲突的东西。目标是用最少的次数交换  $A$ 、 $B$  的内容。使用一个队列  $Q$  包含起始状态  $A = \{w, g, c, p\}$ 、 $B = \emptyset$ 。只要队列不空,我们就取出头部元素,扩展所有可能的选择。然后将扩展后的状态放回队尾。如果队列头部等于  $A = \emptyset$ 、 $B = \{w, g, c, p\}$ ,我们就找到了解。图 15.10 描述了搜索顺



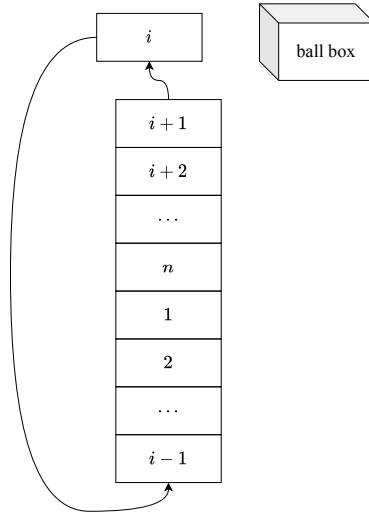


图 15.9: 第  $i$  个人出队摸球。如果没有摸到黑球就站到队尾。

序。同一深度上的所有可能都被检查了,无需进行回溯。

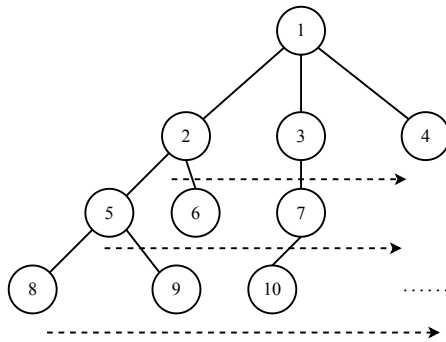


图 15.10: 从第一个状态开始, 检查第二步的所有选项 2、3、4; 然后检查第 3 层上的所有选项……

可以用 4 位二进制数来表示集合, 每一位表示一种事物, 狼  $w = 1$ 、羊  $g = 2$ 、白菜  $c = 4$ 、农夫  $p = 8$ 。0 表示空集, 15 表示包含所有事物的集合。值 3 表示只有狼和羊, 此时狼会吃掉羊。同样, 值 6 表示另一种冲突情况。每次我们将最高位(8)和另外一位(4、2、1)从一个数字移动到另一个数字上。可行的移动方法为:

$$mv A B = \begin{cases} B < 8 : [(A - 8 - i, B + 8 + i) | i \leftarrow [0, 1, 2, 4], i = 0 \text{ 或 } A \bar{\wedge} i \neq 0] \\ \text{否则} : [(A + 8 + i, B - 8 - i) | i \leftarrow [0, 1, 2, 4], i = 0 \text{ 或 } B \bar{\wedge} i \neq 0] \end{cases} \quad (15.11)$$

其中  $\bar{\wedge}$  表示按位与运算。我们用队列  $Q = \{[(15, 0)]\}$  启动搜索:  $solve Q$ 。

$$\begin{aligned}
 \text{solve } \emptyset &= \emptyset \\
 \text{solve } Q &= \begin{cases} A = 0: & \text{reverse } c, \text{ 其中 } (A, B) = c, (c, Q') = \text{pop } Q \\ \text{否则}: & \text{solve } (\text{pushAll } (\text{map } (: c) (\text{filter } (\text{valid } c) (\text{mv } A B)))) Q' \end{cases} \\
 & \qquad \qquad \qquad (15.12)
 \end{aligned}$$

其中函数 *valid c* 检查新的移动结果  $(A, B)$  是否存在冲突。不能是 3,6, 并且尚未尝试过, 不存在于  $c$  中:

$$A, B \neq 3 \text{ 或 } 6, (A, B) \notin c \qquad (15.13)$$

下面是对应的迭代实现:

```

1: function SOLVE
2:   S ← []
3:   Q ← {(15, 0)}
4:   while Q ≠ ∅ do
5:     C ← DEQ(Q)
6:     if C[1] = (0, 15) then
7:       ADD(S, REVERSE(C))
8:     else
9:       for each m in MOVES(C) do
10:        if VALID(m, C) then
11:          ENQ(Q, m:C)
12:   return S

```

下面列出了两个最优解。

| 左         | 河 | 右         |
|-----------|---|-----------|
| 狼、羊、白菜、农夫 |   |           |
| 狼、白菜      |   | 羊、农夫      |
| 狼、白菜、农夫   |   | 羊         |
| 白菜        |   | 狼、羊、农夫    |
| 羊、白菜、农夫   |   | 狼         |
| 羊         |   | 狼、白菜、农夫   |
| 羊、农夫      |   | 狼、白菜      |
|           |   | 狼、羊、白菜、农夫 |

| 左         | 河 | 右         |
|-----------|---|-----------|
| 狼、羊、白菜、农夫 |   |           |
| 狼、白菜      |   | 羊、农夫      |
| 狼、白菜、农夫   |   | 羊         |
| 狼         |   | 羊、白菜、农夫   |
| 狼、羊、农夫    |   | 白菜        |
| 羊         |   | 狼、白菜、农夫   |
| 羊、农夫      |   | 狼、白菜      |
|           |   | 狼、羊、白菜、农夫 |

### 15.1.5 倒水问题

有两个水瓶,一个 9 升、一个 4 升。如何才能从河中取出 6 升水? 这个题目可以追溯到古希腊。有一个故事说法国数学家泊松在孩童时代就解决了这个问题。在好莱坞电影《虎胆龙威 3》中也出现了这个问题。数学家波利亚在《如何解题》中用倒推法给出了一个解<sup>[90]</sup>。最终状态大瓶子中盛有 6 升水。倒数第二步时,从 9 升的瓶子中倒出 3 升水。为了达成这一点,小瓶子中需要盛有 1 升水。如图 15.11 所示。

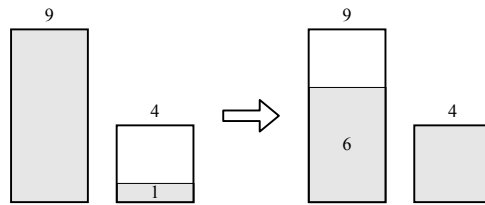


图 15.11: 最后两步

只要倒满 9 升的瓶子,然后连续两次倒入 4 升的瓶子,并将 4 升的瓶子倒空,就可以得到 1 升水。如图 15.12 所示。倒推法是一种策略,而不是具体的算法。它无法回答如何从 899 升和 1147 升的瓶子得到 2 升水。

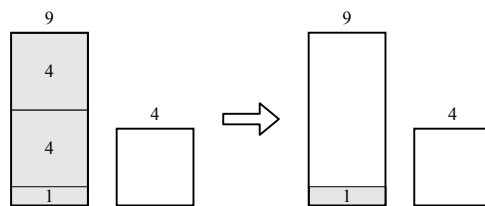


图 15.12: 将大瓶倒满,然后倒入小瓶两次

大小两个瓶子  $B, A$ , 容积分别为  $b, a$ , 每次有 6 种操作:(1)  $A$  装满水;(2)  $B$  装满水;(3) 倒空  $A$ ;(4) 倒空  $B$ ;(5) 将  $A$  中的水倒入  $B$ ;(6) 将  $B$  倒入  $A$ 。下表是一系列

倒水动作,假设容积  $a < b < 2a$ 。

| $A$       | $B$      | 操作           |
|-----------|----------|--------------|
| 0         | 0        | 开始           |
| $a$       | 0        | 倒满 $A$       |
| 0         | $a$      | 将 $A$ 倒入 $B$ |
| $a$       | $a$      | 倒满 $A$       |
| $2a - b$  | $b$      | 将 $A$ 倒入 $B$ |
| $2a - b$  | 0        | 倒光 $B$       |
| 0         | $2a - b$ | 将 $A$ 倒入 $B$ |
| $a$       | $2a - b$ | 倒满 $A$       |
| $3a - 2b$ | $b$      | 将 $A$ 倒入 $B$ |
| ...       | ...      | ...          |

无论何种操作,每个瓶中水的容量总可以表示为  $xa + yb$  的形式,其中  $x, y$  是整数。根据数论中的结论,我们可以判断是否能得到  $g$  升水:当且仅当  $g$  能够被  $a, b$  的最大公约数整除时。即:  $\gcd(a, b) | g$ 。如果  $\gcd(a, b) = 1$  ( $a, b$  互素),可以得到任意自然数  $g$  升水。虽然可以判定是否有解,但我们并不知道具体的倒水步骤。解出丢番图方程  $g = xa + yb$  中的  $x, y$ ,可以得到一组操作。设  $x > 0, y < 0$ ,我们倒满瓶  $A$  共  $x$  次、倒空瓶  $B$  共  $y$  次。例如小瓶容积  $a = 3$ 、大瓶容积  $b = 5$ ,要取得  $g = 4$  升水。因为  $4 = 3 \times 3 - 5$ ,可以设计下列步骤:

| $A$ | $B$ | 操作           |
|-----|-----|--------------|
| 0   | 0   | 开始           |
| 3   | 0   | 倒满 $A$       |
| 0   | 3   | 将 $A$ 倒入 $B$ |
| 3   | 3   | 倒满 $A$       |
| 1   | 5   | 将 $A$ 倒入 $B$ |
| 1   | 0   | 将 $B$ 倒空     |
| 0   | 1   | 将 $A$ 倒入 $B$ |
| 3   | 1   | 倒满 $A$       |
| 0   | 4   | 将 $A$ 倒入 $B$ |

共倒满  $A$  瓶 3 次、倒空  $B$  瓶 1 次。我们可以利用数论中**扩展欧几里得算法**找出整数解  $x$  和  $y$ :

$$(d, x, y) = \gcd_{ext}(a, b) \quad (15.14)$$

其中  $d = \gcd(a, b)$ ,  $ax + by = d$ 。设  $a < b$ , 商  $q$  和余数  $r$ , 满足关系  $b = aq + r$ 。公约数  $d$  整除  $a, b$ , 因此  $d$  也整除  $r$ 。由于  $r < a$ , 可以通过寻找  $a$  和  $r$  的最大公约数减小问题的规模。

$$(d, x', y') = \gcd_{ext}(r, a) \quad (15.15)$$

其中  $d = x'r + y'a$ 。将  $r = b - aq$  代入:

$$\begin{aligned} d &= x'(b - aq) + y'a \\ &= (y' - x'q)a + x'b \end{aligned} \quad (15.16)$$

与  $d = ax + by$  对比, 有下面递归关系:

$$\begin{cases} x &= y' - x'\frac{b}{a} \\ y &= x' \end{cases} \quad (15.17)$$

递归的边界条件发生在  $a = 0$  时:  $\gcd(0, b) = b = 0a + 1b$ 。这样扩展欧几里得算法定义为:

$$\begin{aligned} \gcd_{ext}(0, b) &= (b, 0, 1) \\ \gcd_{ext}(a, b) &= (d, y' - x'\frac{b}{a}, x') \end{aligned} \quad (15.18)$$

其中  $d, x', y'$  的定义如式 (15.15)。如果  $g = md$ , 则  $mx$  和  $my$  就是倒水问题的一个解; 如果  $x < 0$ , 例如  $\gcd_{ext}(4, 9) = (1, -2, 1)$ 。由于  $d = xa + yb$ , 我们不断将  $x$  加  $b$ , 同时将  $y$  减  $a$ , 直到  $x$  大于 0。这样得到的解并不一定最优。例如用 3 升、5 升的瓶子, 获取 4 升水, 扩展欧几里得算法给出 23 步:

[(0, 0), (3, 0), (0, 3), (3, 3), (1, 5), (1, 0), (0, 1), (3, 1), (0, 4), (3, 4), (2, 5), (2, 0), (0, 2), (3, 2), (0, 5), (3, 5), (3, 0), (0, 3), (3, 3), (1, 5), (1, 0), (0, 1), (3, 1), (0, 4)]

而最优解只需要 6 步:

[(0, 0), (0, 5), (3, 2), (0, 2), (2, 0), (2, 5), (3, 4)]

丢番图方程  $g = xa + by$  有无穷多个解,  $|x| + |y|$  越小, 所需步骤越少。我们可以采用类似“过河问题”的思路。在 6 种操作中(倒满 A、倒满 B、将 A 倒入 B……)“并行”尝试出最优解。使用一个队列来安排所有的尝试。队列中的元素是一系列值对  $(p, q)$ ,  $p, q$  分别是两瓶中水的体积, 记录了从开始到最后的倒水操作。开始时队列内容为:  $\{(0, 0)\}$ 。

$$solve\ a\ b\ g = bfs\{\{(0, 0)\}\} \quad (15.19)$$

只要队列不空, 我们从头部取出一操作序列。如果序列中的最后状态包含  $g$  升水, 我们就找到了一个解。我们将序列逆序输出; 否则我们扩展最后的状态, 尝试 6 种可能, 去掉重复的并入队。

$$\begin{aligned}
 bfs \ \emptyset &= [] \\
 bfs \ Q &= \begin{cases} p \text{或} q = g : \text{reverse } s, \text{ 其中 } (p, q) = \text{head } s, (s, Q') = \text{pop } Q & (15.20) \\ \text{否则} : \quad bfs \ (\text{pushAll } (\text{map } (: s) \ (\text{try } s)) \ Q') \end{cases}
 \end{aligned}$$

$$\text{try } s = \text{filter } (\notin s) [f \ (p, q) | f \leftarrow \{fl_A, fl_B, pr_A, pr_B, em_A, em_B\}] \quad (15.21)$$

其中:

$$\begin{cases} fl_A \ (p, q) = (a, q) \\ fl_B \ (p, q) = (p, b) \\ em_A \ (p, q) = (0, q) \\ em_B \ (p, q) = (p, 0) \\ pr_A \ (p, q) = (\max(0, p + q - b), \min(x + y, b)) \\ pr_B \ (p, q) = (\min(x + y, a), \max(0, x + y - a)) \end{cases} \quad (15.22)$$

这一方法总返回最快的步骤。我们无需在队列的每个元素中保存操作步骤,可以利用一个全局历史记录,如图 15.13。初始状态为 (0, 0)。只有 fill A 和 fill B 可行。接下来在记录 (3, 0) 的基础上尝试 fill B,记录新结果 (3, 5)。在 (3, 0) 的基础上尝试 empty A 将回到初始状态 (0, 0)。我们跳过这一选项。图中灰色状态是重复的。我们可以给图 15.13 中的每个节点增加一个父引用,利用它回溯到初始状态。

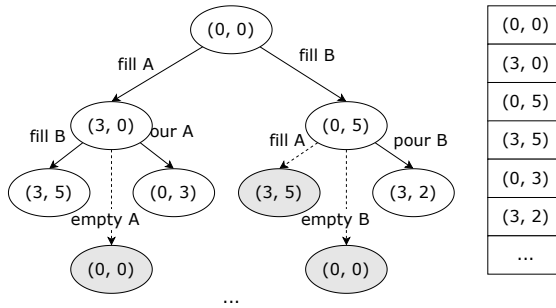


图 15.13: 全局记录

```

1: function SOLVE(a, b, g)
2:   Q ← {(0, 0, NIL)}
3:   V ← {(0, 0, NIL)}
4:   while Q ≠ ∅ do
5:     s ← POP(Q)
6:     if p(s) = g 或 q(s) = g then
7:       return BACK-TRACK(s)
8:     else
9:       for each c in EXPAND(s, a, b) do

```

▷ 全局记录集合

```

10:         if  $c \neq s$  且  $c \notin V$  then
11:             PUSH( $Q, c$ )
12:             ADD( $V, c$ )
13:     return NIL

14: function EXPAND( $s, a, b$ )
15:      $p \leftarrow p(s), q \leftarrow q(s)$ 
16:     return  $[(a, q, s), (p, b, s), (0, q, s), (p, 0, s), (\max(0, p+q-b), \min(p+q, b), s), (\min(p+q, a), \max(0, p+q-a), s)]$ 

17: function BACK-TRACK( $s$ )
18:      $r \leftarrow []$ 
19:     while  $s \neq \text{NIL}$  do
20:          $(p, q, s') = s$ 
21:          $r \leftarrow (p, q):r$ 
22:          $s \leftarrow s'$ 
23:     return  $r$ 

```

## 练习 15.4

15.4.1. 改进扩展欧几里得算法,寻找线性组合使得  $|x| + |y|$  最小,以解决倒水问题。

答案: 401 页

### 15.1.6 华容道

传统玩具华容道是滑块类游戏,如图 15.14。国外称 Klotski。滑块大小、布局会有不同。华容道有 10 个滑块,上面标有数字或者图案。最小的滑块大小为一个单位的正方形,最大的一块为  $2 \times 2$  单位。在棋盘下方的中间,有一个宽度为 2 个单位长的缺口。最大的一块代表曹操,其他的为刘备手下的五虎上将和士兵。游戏的目标是要通过滑动,将曹操移动到棋盘最下方逃走。图 15.15 是日本游戏“箱子中的女儿”,最大的一块代表女儿,剩余滑块代表其他家庭成员。

我们用  $5 \times 4$  矩阵来代表棋盘,行列从 0 开始。1 到 10 个数字代表棋子。0 代表空位置。矩阵  $M$  给出了华容道的初始状态。值为  $i$  的格子表示有棋子占据。布局用字典  $L$  代表。 $L[i]$  是棋子  $i$  覆盖的位置集合。例如  $L[4] = \{(2, 1), (2, 2)\}$  表示第 4 个棋子覆盖了位置  $(2, 1), (2, 2)$ 。我们可以把棋盘上的 20 个位置编号为 0 到 19,把行列转化为编号: $c = 4y + x$ 。这样第四个棋子占据  $L[4] = \{9, 10\}$ 。



(a) 起始布局

(b) 移动若干步后的样子

图 15.14: 华容道游戏



图 15.15: 日本的“箱子中的女儿”游戏



$$M = \begin{bmatrix} 1 & 10 & 10 & 2 \\ 1 & 10 & 10 & 2 \\ 3 & 4 & 4 & 5 \\ 3 & 7 & 8 & 5 \\ 6 & 0 & 0 & 9 \end{bmatrix} \quad L = \left\{ \begin{array}{l} 1 \mapsto \{0, 4\}, 2 \mapsto \{3, 7\}, 3 \mapsto \{8, 12\}, \\ 4 \mapsto \{9, 10\}, 5 \mapsto \{11, 15\}, \\ 6 \mapsto \{16\}, 7 \mapsto \{13\}, 8 \mapsto \{14\}, \\ 9 \mapsto \{19\}, 10 \mapsto \{1, 2, 5, 6\} \end{array} \right\}$$

我们可以定义映射  $\varphi(M) \mapsto L$  和逆映射  $\varphi^{-1}(L) \mapsto M$  在棋盘和布局间转换:

```

1: function  $\varphi(M)$ 
2:    $L \leftarrow \{\}$ 
3:   for  $y \leftarrow 0 \sim 4$  do
4:     for  $x \leftarrow 0 \sim 3$  do
5:        $k \leftarrow M[y][x]$ 
6:        $L[k] \leftarrow \text{ADD}(L[k], 4y + x)$ 
7:   return  $L$ 

8: function  $\varphi^{-1}(L)$ 
9:    $M \leftarrow [[0] \times 4] \times 5$ 
10:  for each ( $k \mapsto S$ ) in  $L$  do
11:    for each  $c$  in  $S$  do
12:       $x \leftarrow c \bmod 4, y \leftarrow \lfloor c/4 \rfloor$ 
13:       $M[y][x] \leftarrow k$ 
14:  return  $M$ 

```

我们检查全部 10 个棋子,看看能否在上下左右 4 个方向移动 1 格。在矩阵棋盘上,移动表示为  $(\Delta y, \Delta x) = (0, \pm 1), (\pm 1, 0)$ , 在布局上,四个方向分别表示为位移  $d = \pm 1, \pm 4$ ,例如棋子  $L[i] = \{c_1, c_2\}$  向左移动变为  $\{c_1 - 1, c_2 - 1\}$ 。这里要排除两个特殊情况:  $d = 1, c \bmod 4 = 3$  和  $d = -1, c \bmod 4 = 0$ ,防止棋子从一侧边界跳到另一侧。考虑两个空位四周最多有 8 种移动可能。例如第一步只有 4 种可能:第 6 块向右;第 7 或第 8 块向下;第 9 块向左。图 15.16 描述列如何检查移动可行。

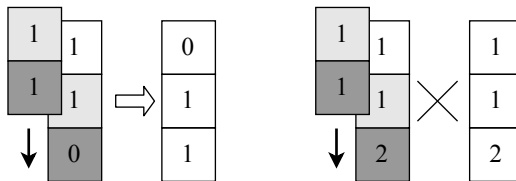


图 15.16: 左侧:两个标 1 的格子都可移动;右侧:下方标 1 的格子 and 标 2 的格子冲突。

为了确定可否移动,我们检查棋子  $k$  移动到的目标格子,如果为 0 或者  $k$  则可移动:

valid  $L[k] d$  :

$$\forall c \in L[k] \Rightarrow y = \lfloor c/4 \rfloor + \lfloor d/4 \rfloor, x = (c \bmod 4) + (d \bmod 4), \quad (15.23)$$

$$(0, 0) \leq (y, x) \leq (4, 3), M[y][x] \in \{k, 0\}$$

经过一系列移动可能回到某个布局。仅避免棋盘矩阵相同是不够的,虽然下面的  $M_1 \neq M_2$ ,但它们本质上是相同的布局。

$$M_1 = \begin{bmatrix} 1 & 10 & 10 & 2 \\ 1 & 10 & 10 & 2 \\ 3 & 4 & 4 & 5 \\ 3 & 7 & 8 & 5 \\ 6 & 0 & 0 & 9 \end{bmatrix} \quad M_2 = \begin{bmatrix} 2 & 10 & 10 & 1 \\ 2 & 10 & 10 & 1 \\ 3 & 4 & 4 & 5 \\ 3 & 7 & 6 & 5 \\ 8 & 0 & 0 & 9 \end{bmatrix}$$

我们需要比较布局来避免重复。忽略具体棋子,定义归一化布局为集合:  $\|L\| = \{p | (k \mapsto p) \in L\}$ ,也就是字典  $L$  中所有值的集合。上面两个矩阵的归一化布局都等于  $\{\{1, 2, 5, 6\}, \{0, 4\}, \{3, 7\}, \{8, 12\}, \{9, 10\}, \{11, 15\}, \{16\}, \{13\}, \{14\}, \{19\}\}$ 。左右对称的布局也是一种重复,需要避免。例如下面的  $M_1$  和  $M_2$  对称。

$$M_1 = \begin{bmatrix} 10 & 10 & 1 & 2 \\ 10 & 10 & 1 & 2 \\ 3 & 5 & 4 & 4 \\ 3 & 5 & 8 & 9 \\ 6 & 7 & 0 & 0 \end{bmatrix} \quad M_2 = \begin{bmatrix} 3 & 1 & 10 & 10 \\ 3 & 1 & 10 & 10 \\ 4 & 4 & 2 & 5 \\ 7 & 6 & 2 & 5 \\ 0 & 0 & 9 & 8 \end{bmatrix}$$

它们的归一化布局也是对称的。我们定义归一化布局的对称变换:

$$\text{mirror}(\|L\|) = \{\{f(c) | c \in s\} | s \in \|L\|\} \quad (15.24)$$

其中  $f(c) = 4y' + x'$ ,  $y' = \lfloor c/4 \rfloor$ ,  $x' = 3 - (c \bmod 4)$ 。我们使用一个队列进行搜索,队列中每个元素包含两部分:一系列移动和这些移动导致的布局。每次移动为  $(k, d)$ ,表示在棋盘上移动棋子  $k$ ,方向为  $d$ (值为  $\pm 1, \pm 4$ )。队列初始化为  $Q = \{(s, [])\}$ ,其中  $s$  是起始布局。只要  $Q \neq \emptyset$ ,我们就从头部取出一个元素,检查最大的棋子(编号 10)是否到达目标位置  $t = \{13, 14, 17, 18\}$ ,即  $L[10] = t$ 。如到达则结束;否则,我们尝试上下左右移动每块棋子,把所有可行的、不重复布的局移动方案  $(k, d)$  入队。在搜索过程中,我们用集合  $H$  记录归一化布局以避免重复。

$$\begin{aligned} \text{solve } \emptyset H &= [] \\ \text{solve } Q H &= \begin{cases} L[10] = t : & \text{reverse } ms, \text{ 其中 } ((L, ms), Q') = \text{pop } Q \\ \text{否则} : & \text{solve } (\text{pushAll } cs \ Q') \ H' \end{cases} \quad (15.25) \end{aligned}$$

其中  $cs = [(move \ L \ e, e:ms) | e \leftarrow \text{expand } L]$  是新产生的移动方案。

$$\begin{aligned} \text{expand } L &= \{(k, d) | k \leftarrow [1, 2, \dots, 10], d \leftarrow [\pm 1, \pm 4], \\ &\quad \text{valid } k \ d, \text{ unique } k \ d\} \quad (15.26) \end{aligned}$$

*move* 把棋子  $L[k]$  移动  $d$  成为:  $move\ L(k, d) = map(+d)\ L[k]$ 。 *unique* 判断新归一化布局  $\|L'\| \notin H$ , 镜像  $mirror(\|L'\|) \notin H$  不在历史记录中。如果没有重复, 更新它们到新记录  $H'$  用于后继搜索。下面是对应的迭代式实现。“横刀立马”布局的最少步数解共 116 步(每步移动 1 格), 最后 3 步如下:

```

1: function SOLVE( $s, e$ )
2:    $H \leftarrow \{\|s\|\}$ 
3:    $Q \leftarrow \{(s, \emptyset)\}$ 
4:   while  $Q \neq \emptyset$  do
5:      $(L, p) \leftarrow POP(Q)$ 
6:     if  $L[10] = e$  then
7:       return  $(L, p)$ 
8:     else
9:       for each  $L'$  in EXPAND( $L, H$ ) do
10:        PUSH( $Q, (L', L)$ )
11:        ADD( $H, \|L'\|\}$ )
12:   return  $\emptyset$ 

```

```

['5', '3', '2', '1']
['5', '3', '2', '1']
['7', '9', '4', '4']
['A', 'A', '6', '0']
['A', 'A', '0', '8']

```

```

['5', '3', '2', '1']
['5', '3', '2', '1']
['7', '9', '4', '4']
['A', 'A', '0', '6']
['A', 'A', '0', '8']

```

```

['5', '3', '2', '1']
['5', '3', '2', '1']
['7', '9', '4', '4']
['0', 'A', 'A', '6']
['0', 'A', 'A', '8']

```

过河问题、倒水问题、华容道游戏的解法有共同的结构。和深度优先类似, 它们都有起始、终止状态。在过河问题中, 起始状态是农夫、狼、羊、白菜在河的一岸, 对岸为空; 终止状态是全部渡河到了对岸。倒水问题的起始状态是两个空瓶子, 而终止状态是某个瓶子盛有指定容量的水。华容道问题的起始状态是某种布局(“横刀立马”), 终止状态是另外一个布局, 其中最大的棋子移动到了指定的位置。每个问题都有相应的规则, 可以从一个状态转移到另外一个状态。我们“并行”尝试可能选项。在同一步内的选项未尝试完之前, 我们不会进一步深入搜索。这就保证了最小步骤解在其它解之前出现。由于向水平方向扩展, 这种方法称为广度优先搜索。图 15.17 给出了广度、深度优先搜索的对比。

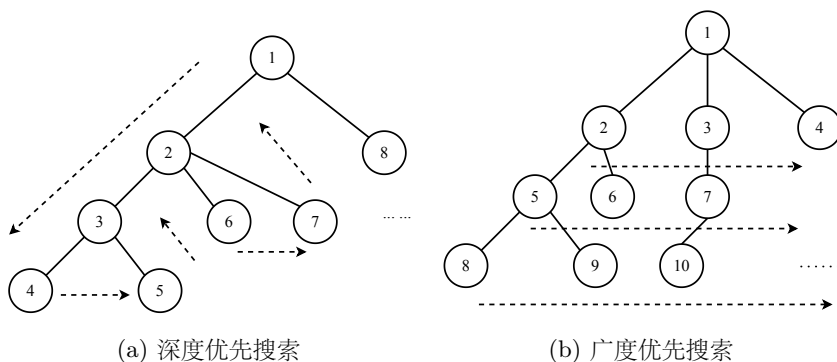


图 15.17: 深度和广度优先搜索的顺序

由于我们无法真正的“并行”搜索,广度优先搜索使用队列来记录尝试。从队列头部不断取出步骤较少的候选项,从队列尾部不断加入步骤较多的新候选项。广度优先搜索提供了一种简单的方法寻找最少步骤解,但它不能直接搜索其它最优解。考虑如图 15.18 的有向图,每段路径长度不同,我们无法用广度优先搜索找出两个城市之间的最短路径。注意从城市  $a$  到城市  $c$  之间的最短路径并非经过最少城市的  $a \rightarrow b \rightarrow c$ 。这条路径的总长度为 22;而是经过更多城市的路径  $a \rightarrow e \rightarrow f \rightarrow c$ ,它的总长度只有 20。

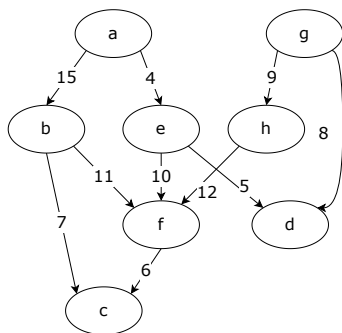


图 15.18: 带权重的有向图

## 练习 15.5

15.5.1. 按层遍历一棵二叉树。

15.5.2. 康威<sup>2</sup>提出了一种滑动趣题。图 15.19 给出的是一种简化的版本。8 个圆圈中的 7 个已经放入了棋子,每个棋子上标有编号 1 到 7。如果和棋子相邻的圆圈是空的,则棋子可以滑动过去。圆圈间如果有连线,则表示它们是相连的。目标是将棋子从顺序 1、2、3、4、5、6、7 通过滑动反转成 7、6、5、4、3、2、1。编写一个程序解决康威滑动问题。

<sup>2</sup>约翰·康威(1937-2020),英国数学家。

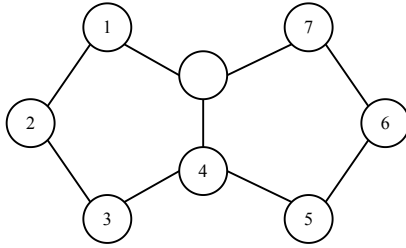


图 15.19: 康威滑动趣题

答案: 402 页

## 15.2 贪心算法

人们需要“最好”的解来节省时间、空间、成本、能量。使用有限的资源搜索最优解并不容易。有很多问题不存在多项式时间的解法。仅对一些特定问题存在简单方法找到最优解。

### 15.2.1 哈夫曼编码

哈夫曼编码是一种用最小长度对信息编码的方法。早期的 ASCII 码使用 7 位二进制数对字母、数字、符号编码,可以表达  $2^7 = 128$  个字符。只用 0、1,我们至少需要  $\log_2 n$  位分辨  $n$  个不同字符。下面是大写英文字母的码表,将字母 A 到 Z 映射为 0 到 25。每个编码 5 位。零被扩充到 5 位 00000 而非 0。这样的编码方式称为定长编码。

| 字符 | 编码    | 字符 | 编码    |
|----|-------|----|-------|
| A  | 00000 | N  | 01101 |
| B  | 00001 | O  | 01110 |
| C  | 00010 | P  | 01111 |
| D  | 00011 | Q  | 10000 |
| E  | 00100 | R  | 10001 |
| F  | 00101 | S  | 10010 |
| G  | 00110 | T  | 10011 |
| H  | 00111 | U  | 10100 |
| I  | 01000 | V  | 10101 |
| J  | 01001 | W  | 10110 |
| K  | 01010 | X  | 10111 |
| L  | 01011 | Y  | 11000 |
| M  | 01100 | Z  | 11001 |

文本“INTERNATIONAL”可以编码为 65 位的二进制数:

00010101101100100100100011011000000110010001001110101100000011010

另一种方式是变长编码。用一位二进制数 0 代表 A, 用两位二进制数 10 代表 C, 用 5 位二进制数 11001 代表 Z。虽然可以显著缩短编码长度。但在解码时会造成歧义。例如二进制数 1101, 我们不知道它是 1, 后面跟着 101 (表示“BF”), 还是 110, 后面跟着 1 (表示“GB”), 或是 1101 (表示 N)。摩尔斯电码是变长编码。最常用的字符 E 编码为“.”, 字符 Z 编码为“- -..”。使用特殊终止符分割编码, 所以不会发生歧义。下面是一个特殊的无歧义码表:

| 字符 | 编码  | 字符 | 编码   |
|----|-----|----|------|
| A  | 110 | E  | 1110 |
| I  | 101 | L  | 1111 |
| N  | 01  | O  | 000  |
| R  | 001 | T  | 100  |

文本“INTERNATIONAL”编码为 38 位的二进制数:

10101100111000101110100101000011101111

按上表解码不会遇到有歧义的字符。这是因为没有任何编码是其它编码的前缀。这样的编码称为**前缀码**<sup>3</sup>。前缀码不需要分隔符, 编码长度可进一步缩短。这自然引发了一个问题: 给定文本, 能否找到码表使得编码长度最短? 1951 年, 麻省理工学院的老师罗伯特·费诺在课上宣布: 谁解出了这个问题就不用参加期末考试了。哈夫曼<sup>[91]</sup>尝试了很久, 他几乎放弃了, 开始准备考试。恰在此时, 他灵光乍现找到了解法。哈夫曼根据文本中字符出现的频率构造码表。最“常见”的字符编码最短。首先处理文本获得每个字符出现的次数。定义字符的权重是它出现的次数或概率。哈夫曼用二叉树产生前缀码。字符保存于叶子节点。从根节点遍历产生编码。向左前进时添加 0, 向右前进时添加 1, 如图 15.20。例如, 从根节点遍历到 N 的路径是, 向左然后向右到达 N。因此 N 编码为 01; 字符 A 的路径是: 右, 右, 左, 编码是 110。

这棵树还可以解码。扫描二进制数, 0 向左、1 向右。到达叶子时, 节点上的字符就是解码内容。然后重新返回根节点继续扫描。哈夫曼用自底向上的方法构造树。开始时, 每个字符都放入一个叶子节点中。每次选出两个权重最小的节点合并成一个分支。分支权重为两个子树的权重和。不断选择权重最小的两棵树合并, 最后得到一棵树, 如图 15.21。

我们重用二叉树的定义实现哈夫曼树。节点中增加了权重, 只有叶子节点保存字符。分枝节点表示为  $(w, l, r)$ , 其中  $w$  是权重,  $l, r$  是左右子树。叶子节点表示为  $(w, c)$  其中  $c$  是字符。合并子树时, 权重相加  $merge\ a\ b = (weight\ a + weight\ b, a, b)$ , 其中:

<sup>3</sup>英文叫做 prefix-code, 而不是无前缀码。

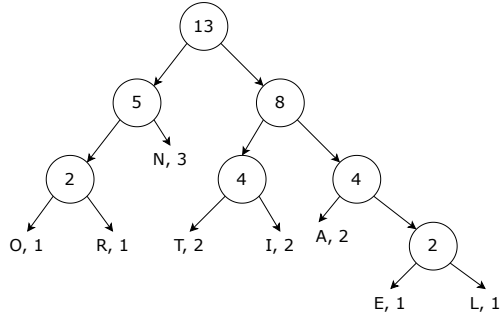


图 15.20: 哈夫曼树

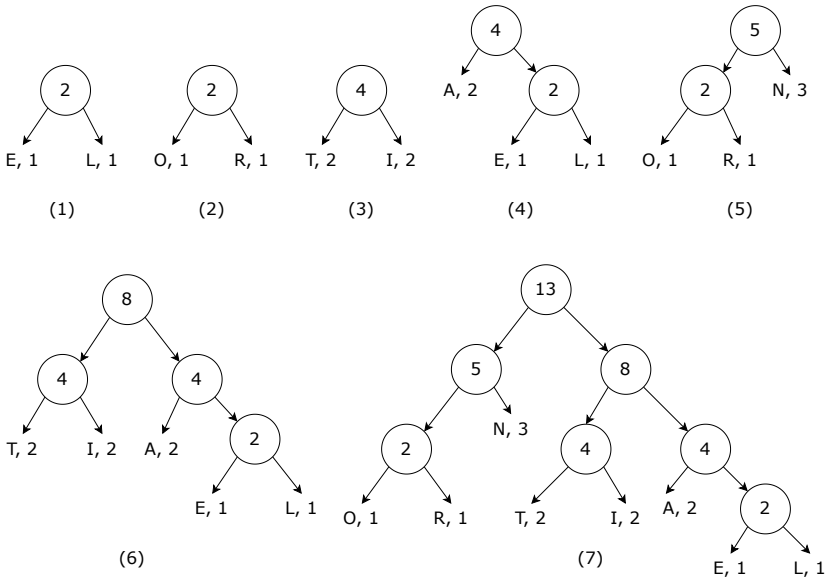


图 15.21: 构造哈夫曼树

$$\begin{aligned} \text{weight}(w, a) &= w \\ \text{weight}(w, l, r) &= w \end{aligned} \quad (15.27)$$

下面算法不断选出两棵权重最小的子树合并:

$$\begin{aligned} \text{build } [t] &= t \\ \text{build } ts &= \text{build } (\text{merge } t_1 \ t_2) \ ts', \text{ 其中 } (t_1, t_2, ts') = \text{extract } ts \end{aligned} \quad (15.28)$$

函数 *extract* 从子树列表中选出权重最小的两棵树, 若  $\text{weight } t_1 < \text{weight } t_2$ , 定义  $t_1 < t_2$ 。

$$\text{extract}(t_1:t_2:ts) = \text{foldr } \text{min}_2 \ (\text{min } t_1 \ t_2, \text{max } t_1 \ t_2, [ ]) \ ts \quad (15.29)$$

其中:

$$\text{min}_2 \ t \ (t_1, t_2, ts) = \begin{cases} t < t_2 : & (\text{min } t \ t_1, \text{max } t \ t_1, t_2:ts) \\ \text{否则} : & (t_1, t_2, t:ts) \end{cases} \quad (15.30)$$

为了迭代构造哈夫曼树, 我们用数组  $A$  存储  $n$  棵子树。从右向左扫描  $A$ , 如果  $A[i]$  的权重小于  $A[n-1]$  或  $A[n]$ , 就交换  $A[i]$  和  $\text{MAX}(A[n-1], A[n])$ 。扫描后合并  $A[n]$  到  $A[n-1]$ 。这样数组缩小一。重复此过程得到最终的哈夫曼树:

```

1: function HUFFMAN( $A$ )
2:   while  $|A| > 1$  do
3:      $n \leftarrow |A|$ 
4:     for  $i \leftarrow n - 2$  down to 1 do
5:        $T \leftarrow \text{MAX}(A[n], A[n - 1])$ 
6:       if  $A[i] < T$  then
7:         EXCHANGE  $A[i] \leftrightarrow T$ 
8:        $A[n - 1] \leftarrow \text{MERGE}(A[n], A[n - 1])$ 
9:       DROP( $A[n]$ )
10:  return  $A[1]$ 

```

我们可以从哈夫曼树  $T$  构造码表。令  $p = []$ 。从根节点遍历树。左转时  $p \leftarrow 0:p$ , 右转时  $p \leftarrow 1:p$ 。到达叶节点字符  $c$  时, 记录  $c \mapsto \text{reverse } p$  到码表。定义(柯里化)  $\text{code} = \text{traverse } [ ]$ , 其中:

$$\begin{aligned} \text{traverse } p \ (w, c) &= [c \mapsto \text{reverse } p] \\ \text{traverse } p \ (w, l, r) &= \text{traverse } (0:p) \ l \ \# \ \text{traverse } (1:p) \ r \end{aligned} \quad (15.31)$$

编码过程一边扫描文本  $w$  一边查询码表  $\text{dict}$  产生二进制序列:

$$\text{encode } \text{dict } w = \text{concatMap } (c \mapsto \text{dict}[c]) \ w, \text{ 其中 } \text{dict} = \text{code } T \quad (15.32)$$



解码过程相反,一边扫描二进制序列  $bs$  一边查询哈夫曼树。从根节点开始,0 向左、1 向右,到达叶子节点时输出字符  $c$ ,然后回到根节点继续解码。 $decode\ T\ bs = lookup\ T\ bs$ ,其中:

$$\begin{aligned} lookup(w, c) [] &= [c] \\ lookup(w, c) bs &= c : lookup\ T\ bs \\ lookup(w, l, r) (b:bs) &= lookup\ (if\ b = 0\ then\ l\ else\ r)\ bs \end{aligned} \quad (15.33)$$

哈夫曼树的构造过程采取了一种特殊策略:每次合并有若干选项,总是选取权重最小的两棵树合并。这一系列局部最优选择,产生了全局最优的前缀编码。局部最优一般不产生全局最优解。哈夫曼编码是个例外。我们称这种每次选择局部最优选项的方法为贪心策略。贪心算法可以解决、简化很多问题。但判断贪心策略能否产生全局最优解并不容易。通用的形式化证明仍然是一个活跃的研究领域<sup>[4]</sup>。

## 练习 15.6

15.6.1. 实现命令式的 Huffman 码表生成算法。

答案: 403 页

### 15.2.2 换零钱问题

如果用现金购物,如何用最少的硬皮找零钱? 假设有 5 种面额的硬币:1 分、5 分、2.5 角、5 角、1 元。1 角合 10 分,1 元合 10 角,记为:  $C = \{1, 5, 25, 50, 100\}$ 。需要找零  $x$  分。使用贪心策略,每次总挑选最大面值硬币:

$$\begin{aligned} change\ 0 &= [] \\ change\ x &= c_m : change\ (x - c_m), \text{其中: } c_m = \max\ \{c \in C, c \leq x\} \end{aligned} \quad (15.34)$$

例如兑换 1.42 元,函数  $change$  生成硬币列表:  $[100, 25, 5, 5, 5, 1, 1]$ 。我们可以将其转换为  $[(100, 1), (25, 1), (5, 3), (1, 2)]$ ,表示一枚 1 元、一枚 2.5 角、三枚 5 分、两枚 1 分硬币。对于例子  $C$  这些面额,贪心策略可找出最优解。贪心算法对大多数国家的硬币系统都有效。但也有例外:如  $C = \{1, 3, 4\}$ 。兑换  $x = 6$  分钱,最优解是 2 枚 3 分硬币。但贪心策略给出是  $6 = 4 + 1 + 1$ ,共 3 枚硬币。

尽管不是最优解,贪心策略通常实现简单,效果差强人意,因而应用广泛。例如英文折行功能,如果文本  $T$  长度超过页面宽度  $W$ ,需要拆成若干行。令单词间隔为  $s$ ,贪心算法可以给出接近最优的折行方案:在一行中尽可能多放入单词。

```

1:  $L \leftarrow W$ 
2: for  $w \in T$  do
3:   if  $|w| + s > L$  then
4:     Insert line break
5:      $L \leftarrow W - |w|$ 
```

```

6:   else
7:      $L \leftarrow L - |w| - s$ 

```

### 练习 15.7

- 15.7.1. 使用堆来构造哈夫曼树:不断从堆顶取出两颗子树,合并后放回堆中。
- 15.7.2. 如果字符已按权重排序成列表  $A$ , 存在一个线性时间的构造哈夫曼树的方法: 用队列  $Q$  保存合并结果。不断从  $Q$  和  $A$  头部取出较小的树, 合并后入队。处理完列表中所有树后, 队列中将只剩下一棵树。即最终的哈夫曼树。请实现这一方法。
- 15.7.3. 给定哈夫曼树  $T$ , 用左侧叠加实现哈夫曼解码。

答案: 404 页

## 15.3 动态规划

考虑在任何币值系统下寻找换零钱的最优解(使用最少的硬币)。假设已找到兑换零钱  $x$  的最优解: 硬币列表  $C_m$ 。将  $C_m$  中的硬币分成两组:  $C_1, C_2$ 。分别等价于  $x_1, x_2$ 。即  $C_m = C_1 + C_2, x = x_1 + x_2$ 。则  $C_1$  是兑换  $x_1$  的最优解, 且  $C_2$  是兑换  $x_2$  的最优解。

证明. 用反证法。对  $x_1$ , 假设存在另一个更好的兑换方法  $C'_1$ , 比  $C_1$  的硬币更少。则兑换方法  $C'_1 + C_2$  的硬币要少于  $C_m$ 。这和  $C_m$  是兑换  $x$  的最优解矛盾。同样, 我们也可以证明  $C_2$  是兑换  $x_2$  的最优解。□

注意相反命题不成立。任选整数  $y < x$ , 将原问题分解为两个子问题: 分别兑换  $y$  和  $x - y$  的最优解。将这两个最优解合并不一定是兑换  $x$  的最优解。作为反例: 用三种硬币  $C = \{1, 2, 4\}$  兑换  $x = 6$ 。最优解需要两枚硬币:  $2 + 4$ 。但由  $6 = 3 + 3$  分解为兑换 3 的两个子问题, 每个子问题的最优解为  $3 = 1 + 2$ , 但组合方案  $(1 + 2) + (1 + 2)$  共需 4 枚硬币。如果最优化问题可分解为若干最优化子问题, 我们称它具备“最优化子结构”。换零钱问题须根据币值分解出最优化子结构:

$$\begin{aligned} \text{change } 0 &= [] \\ \text{change } x &= \min [c : \text{change } (x - c) | c \in C, c < x] \end{aligned} \quad (15.35)$$

其中  $\min$  选出长度最短的列表。但这一定义不能直接转化为可用的程序。自顶向下递归中有大量重复计算。若  $C = \{1, 2, 25, 50, 100\}$ , 兑换  $\text{change}(142)$  时, 需要计算  $\text{change}(141)$ 、 $\text{change}(137)$ 、 $\text{change}(117)$ 、 $\text{change}(92)$ 、 $\text{change}(42)$ 。在计算  $\text{change}(141)$  时, 将 141 分别减去 1、2、25、50、100。这样就会再次计算 137、117、92、42。搜索空间以  $5^n$  指数膨胀。参考计算斐波那契数列的方法, 我们用表格  $T$  记录最优兑换子问题的解。 $T$  初始时空白。兑换  $y$  时先查询  $T[y]$  获取子问题最优解, 如果  $T[y] = \emptyset$ , 则递归计算子问题。子问题计算好后存入  $T[y]$ 。

```

1:  $T \leftarrow [[ ], \emptyset, \emptyset, \dots]$ 
2: function CHANGE( $x$ )
3:   if  $x > 0$  且  $T[x] = \emptyset$  then
4:     for each  $c$  in  $C$  且  $c \leq x$  do
5:        $C_m \leftarrow c$ : CHANGE( $x - c$ )
6:       if  $T[x] = \emptyset$  或  $|C_m| < |T[x]|$  then
7:          $T[x] \leftarrow C_m$ 
8:   return  $T[x]$ 

```

▷  $T[0] = [ ]$

我们还可以自底向上产生、记录兑换子问题的最优解。从  $T[0] = [ ]$  开始, 依次产生  $T[1] = [1]$ ,  $T[2] = [1, 1]$ ,  $T[3] = [1, 1, 1]$ ,  $T[4] = [1, 1, 1, 1]$ , 如表 15.1(a) 所示。兑换  $T[5]$  有两个选择: 5 个 1 分或 1 枚 5 分硬币。显然后者更优。最优解表格更新为表 15.1(b),  $T[5] = [5]$ 。接下来兑换  $x = 6$ , 1 分、5 分都小于 6, 有两个选择: (1) 1 分加  $T[5]$  得  $[1, 5]$ ; (2) 5 分加  $T[1]$  得  $[5, 1]$ 。这两个解等价, 任选  $T[6] = [1, 5]$ 。每次迭代产生  $T[i]$ , 其中  $i \leq x$  时, 逐一检查所有  $c \leq i$  的币值。查询表格获取子问题  $T[i - c]$  的最优解, 再加上  $c$  得到一个兑换方案。选择硬币最少的一个作为  $T[i]$ 。

| $x$ | 0   | 1   | 2      | 3         | 4            |
|-----|-----|-----|--------|-----------|--------------|
| 最优解 | [ ] | [1] | [1, 1] | [1, 1, 1] | [1, 1, 1, 1] |

(a) 兑换 4 分以内的最优解列表

| $x$ | 0   | 1   | 2      | 3         | 4            | 5   |
|-----|-----|-----|--------|-----------|--------------|-----|
| 最优解 | [ ] | [1] | [1, 1] | [1, 1, 1] | [1, 1, 1, 1] | [5] |

(b) 兑换 5 分以内的最优解列表

表 15.1: 兑换零钱的最优解列表

```

1: function CHANGE( $x$ )
2:    $T \leftarrow [[ ], \emptyset, \dots]$ 
3:   for  $i \leftarrow 1$  to  $x$  do
4:     for each  $c$  in  $C$  且  $c \leq i$  do
5:       if  $T[i] = \emptyset$  或  $1 + |T[i - c]| < |T[i]|$  then
6:          $T[i] \leftarrow c$ :  $T[i - c]$ 
7:   return  $T[x]$ 

```

最优解表格(如下)中有大量重复内容。父问题的解包含着子问题的解。可以只记录相对子问题的变化部分: 兑换  $T[i]$  所选择的硬币  $c$  以及需要的硬币数量  $n$ 。即  $T[i] = (n, c)$ 。为了获得兑换  $x$  的硬币列表, 我们从表格  $T[x]$  中找到  $c$ , 然后再从  $T[x - c]$  中找到  $c'$ ……直到  $T[0]$ 。

|     |        |           |              |                 |        |     |
|-----|--------|-----------|--------------|-----------------|--------|-----|
| 价值  | 6      | 7         | 8            | 9               | 10     | ... |
| 最优解 | [1, 5] | [1, 1, 5] | [1, 1, 1, 5] | [1, 1, 1, 1, 5] | [5, 5] | ... |

```

1: function CHANGE( $x$ )
2:    $T \leftarrow [(0, \emptyset), (\infty, \emptyset), (\infty, \emptyset), \dots]$ 
3:   for  $i \leftarrow 1$  to  $x$  do
4:     for each  $c$  in  $C$  且  $c \leq i$  do
5:        $(n, \_ ) \leftarrow T[i - c], (m, \_ ) \leftarrow T[i]$ 
6:       if  $1 + n < m$  then
7:          $T[i] \leftarrow (1 + n, c)$ 
8:    $s \leftarrow []$ 
9:   while  $x > 0$  do
10:     $(\_, c) \leftarrow T[x]$ 
11:     $s \leftarrow c : s$ 
12:     $x \leftarrow x - c$ 
13:   return  $s$ 

```

最优解表格  $T$  也可以用左侧叠加构建:  $foldl\ fill\ [(0, 0)]\ [1, 2, \dots]$ , 其中:

$$fill\ T\ x = T \triangleright \min \{(fst\ T[x - c], c) | c \in C, c \leq x\} \quad (15.36)$$

其中  $s \triangleright a$  把元素  $a$  加入到序列  $s$  右侧(见第 12 章手指树)。然后利用  $T$  反演出最优解列表:

$$\begin{aligned} change\ 0\ T &= [] \\ change\ x\ T &= c : change\ (x - c)\ T, \text{ 其中 } : c = snd\ T[x] \end{aligned} \quad (15.37)$$

若兑换值  $x = n$ , 算法循环  $n$  次。每次最多检查  $k = |C|$  个币值。复杂度为  $\Theta(nk)$ <sup>4</sup>, 并需要  $O(n)$  的空间保存  $T$ 。不管是自底向上还是自顶向下的方法, 都需要记录最优子问题的解。在计算整体的最优解时, 反复多次使用子问题的结果。这一特性称为重叠子问题。解决换零钱问题的方法叫做动态规划, 是贝尔曼于 1940 年提出的。动态规划问题具备两个性质:

1. 最优化子结构。问题可以被分解为若干规模较小的子问题, 最优解可以从子问题解中构造出;
2. 重叠子问题。子问题的解可被反复使用以寻找整体上的解。

### 15.3.1 最长公共子序列

和子串不同, 子序列无需连续。例如, Mississippi 和 Missunderstanding 的最长公共子串为 Miss, 而最长公共子序列为 Misssi, 如图 15.22 所示。如果这是两段代码, 旋

<sup>4</sup>上界

转 90 度就是代码对比 (diff) 的结果。搜索最长公共子序列是版本控制工具的常见功能。两个字符串  $xs$ 、 $ys$  的最长公共子序列定义如下：

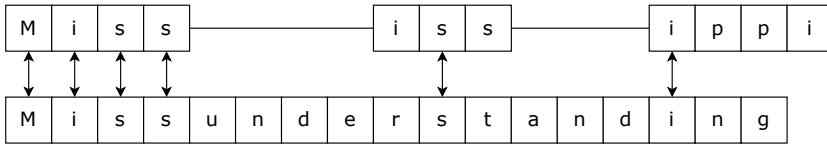


图 15.22: 最长公共子序列

$$\begin{aligned}
 LCS([], ys) &= [] \\
 LCS(xs, []) &= [] \\
 LCS(x:xs, y:ys) &= \begin{cases} x = y : x : LCS(xs, ys) \\ \text{否则} : \max LCS(x:xs, ys) \quad LCS(xs, y:ys) \end{cases} \quad (15.38)
 \end{aligned}$$

其中  $\max$  选出较长的序列。 $LCS$  的定义含有最优化子结构, 可以分解为规模较小的子问题。序列长度每次至少有一个减 1。这一定义也含有重叠子问题。子串间的最长公共子序列多次被用于搜索全局最优解。我们使用二维表格  $T$  记录子问题最优解。行、列分别代表  $xs$ 、 $ys$ 。序列元素的索引从 0 开始, 第 0 行、0 列代表空序列。 $T[i][j]$  表示  $LCS(xs[0..j], ys[0..i])$  的长度, 最终从  $T$  反演出最长公共子序列。因为  $LCS([], ys) = LCS(xs, []) = []$ , 表格第 0 行、0 列都是 0。以 antenna 和 banana 为例。我们从  $T[1][1]$  开始填充第 1 行。b 和 antenna 中的任何一个都不同, 所以第一行都为 0。考虑  $T[2][1]$ , 行列都对应 a,  $T[2][1] = T[1][0] + 1 = 1$ , 即  $LCS(a, ba) = a$ 。接下来移动到  $T[2][2]$ ,  $a \neq n$ , 选择上方 ( $LCS(an, b)$ ) 和左侧 ( $LCS(a, ba)$ ) 的较大值填充  $T[2][2]$ , 结果是 1, 即  $LCS(ba, an) = a$ 。这样我们逐步将表格填充完毕。填充规则归纳为: 对  $T[i][j]$ , 如果  $xs[i-1] = ys[i-1]$ , 则  $T[i][j] = T[i-1][j-1] + 1$ , 否则从上方  $T[i-1][j]$  和左侧  $T[i][j-1]$  中选择较大的。

|   |    |    |   |   |   |   |   |   |   |
|---|----|----|---|---|---|---|---|---|---|
|   |    | 0  | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|   |    | [] | a | n | t | e | n | n | a |
| 0 | [] | 0  | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | b  | 0  | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | a  | 0  | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | n  | 0  | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| 4 | a  | 0  | 1 | 2 | 2 | 2 | 2 | 2 | 3 |
| 5 | n  | 0  | 1 | 2 | 2 | 2 | 3 | 3 | 3 |
| 6 | a  | 0  | 1 | 2 | 2 | 2 | 3 | 3 | 4 |

1: **function** LCS( $xs, ys$ )

```

2:   $m \leftarrow |xs|, n \leftarrow |ys|$ 
3:   $T \leftarrow [[0, 0, \dots], [0, 0, \dots], \dots]$  ▷  $(m + 1) \times (n + 1)$ 
4:  for  $i \leftarrow 1$  to  $m$  do
5:    for  $j \leftarrow 1$  to  $n$  do
6:      if  $xs[i] = ys[j]$  then
7:         $T[i + 1][j + 1] \leftarrow T[i][j] + 1$ 
8:      else
9:         $T[i + 1][j + 1] \leftarrow \text{MAX}(T[i][j + 1], T[i + 1][j])$ 
10: return  $\text{FETCH}(T, xs, ys)$  ▷ 反演结果

```

接下来从  $T$  反演最长公共子序列。从右下方开始, 如果末尾  $xs[m] = ys[n]$ , 则  $xs[m]$  就是 LCS 的末尾, 接下来检查  $xs[m - 1], ys[n - 1]$ , 否则我们选择  $T[m - 1][n]$  和  $T[m][n - 1]$  较大的继续反演。

```

1: function  $\text{FETCH}(T, xs, ys)$ 
2:    $m \leftarrow |xs|, n \leftarrow |ys|$ 
3:    $r \leftarrow []$ 
4:   while  $m > 0$  且  $n > 0$  do
5:     if  $xs[m - 1] = ys[n - 1]$  then
6:        $r \leftarrow xs[m - 1] : r$ 
7:        $m \leftarrow m - 1$ 
8:        $n \leftarrow n - 1$ 
9:     else if  $T[m - 1][n] > T[m][n - 1]$  then
10:       $m \leftarrow m - 1$ 
11:     else
12:       $n \leftarrow n - 1$ 
13:   return  $r$ 

```

## 练习 15.8

15.8.1. 使用叠加操作构造最长公共子序列的最优解表格

答案: 405 页

### 15.3.2 子集和问题

给定整数集合  $X$ , 如何找到所有子集  $S \subseteq X$ , 使得  $S$  中整数相加之和为  $s$ , 即:  $\sum_{i \in S} i = s$ ? 例如  $X = \{11, 64, -82, -68, 86, 55, -88, -21, 51\}$ ,  $s = 0$  有三个解:  $S = \emptyset, \{64, -82, 55, -88, 51\}, \{64, -82, -68, 86\}$ 。如果穷举, 总共检查  $2^n$  个子集之和, 其中  $n = |X|$ 。复杂度为  $O(n2^n)$ 。

$$\begin{aligned} \text{sets } s \ \emptyset &= [\emptyset] \\ \text{sets } s \ (x:xs) &= \begin{cases} s = x : \{x\} : \text{sets } s \ xs \\ \text{否则} : (\text{sets } s \ xs) \# [x:S | S \in \text{sets } (s - x) \ xs] \end{cases} \end{aligned} \quad (15.39)$$

穷举定义中有子结构和重叠子问题,可以用动态规划求解。我们自底向上构建表格  $T$ ,记录子问题解,并最终得出子集之和的解。首先考虑解的存在性问题:判断是否存在某个子集  $S$ ,使得  $\sum S = s$ 。通过一轮遍历可以确定子集之和的上下限  $l \leq s \leq u$ 。如果  $s < l$  或  $s > u$  则无解。

$$l = \sum_X \{x < 0\}, u = \sum_X \{x > 0\} \quad (15.40)$$

集合元素是整数。表格  $T$  有  $m = u - l + 1$  列,每列代表一个值:  $l \leq j \leq u$ 。表格有  $n = |X| + 1$  行,每行对应集合中某个元素  $x_i$ ,  $T[i][j]$  表示是否存在子集  $S \subseteq \{x_1, x_2, \dots, x_i\}$  使得  $\sum S = j$ 。第 0 行特殊,代表空集  $\emptyset$  的子集之和。 $T$  中所有项初始化为假 F,只有  $T[0][0] = \text{T}$ ,表示  $\sum \emptyset = 0$ 。我们从  $x_1$  开始,构造表格第 1 行。除  $\sum \emptyset = 0$  外,显然  $\sum \{x_1\} = x_1$ ,因此  $T[1][0] = \text{T}, T[1][x_1] = \text{T}$ 。

|             |     |       |     |   |     |       |     |     |
|-------------|-----|-------|-----|---|-----|-------|-----|-----|
|             | $l$ | $l+1$ | ... | 0 | ... | $x_1$ | ... | $u$ |
| $\emptyset$ | F   | F     | ... | T | ... | F     | ... | F   |
| $x_1$       | F   | F     | ... | T | ... | T     | ... | F   |
| ...         | F   | F     | ... | T | ... | T     | ... | F   |

加入  $x_2$  可以得到 4 种可能的子集之和:  $\sum \emptyset = 0$ 、 $\sum \{x_1\} = x_1$ 、 $\sum \{x_2\} = x_2$ 、 $\sum \{x_1, x_2\} = x_1 + x_2$ 。

|             |     |       |     |   |     |       |     |       |     |             |     |     |
|-------------|-----|-------|-----|---|-----|-------|-----|-------|-----|-------------|-----|-----|
|             | $l$ | $l+1$ | ... | 0 | ... | $x_1$ | ... | $x_2$ | ... | $x_1 + x_2$ | ... | $u$ |
| $\emptyset$ | F   | F     | ... | T | ... | T     | ... | F     | ... | F           | ... | F   |
| $x_1$       | F   | F     | ... | T | ... | T     | ... | F     | ... | F           | ... | F   |
| $x_2$       | F   | F     | ... | T | ... | T     | ... | T     | ... | T           | ... | F   |
| ...         | F   | F     | ... | T | ... | T     | ... | T     | ... | T           | ... | F   |

构造第  $i$  行时,加入元素  $x_i$ 。仍可得到此前由  $\{x_1, x_2, \dots, x_{i-1}\}$  获得的所有子集之和。上一行中为真的项仍然为真。因为  $\sum \{x_i\} = x_i$ ,所以  $T[i][x_i] = \text{T}$ 。我们将  $x_i$  加到已知的所有和之上得到一些新值,它们对应的项也为真。处理完全部  $n$  个元素后,表格中  $T[n][s]$  的真假值表示子集和  $s$  是否存在。

1: **function** SUBSET-SUM( $X, s$ )

2:  $l \leftarrow \sum \{x \in X, x < 0\}, u \leftarrow \sum \{x \in X, x > 0\}$

```

3:   $n \leftarrow |X|$ 
4:   $T \leftarrow \{\{F, F, \dots\}, \{F, F, \dots\}, \dots\}$   $\triangleright (n+1) \times (u-l+1)$ 
5:   $T[0][0] \leftarrow T$   $\triangleright \sum \emptyset = 0$ 
6:  for  $i \leftarrow 1$  to  $n$  do
7:       $T[i][X[i]] \leftarrow T$ 
8:      for  $j \leftarrow l$  to  $u$  do
9:           $T[i][j] \leftarrow T[i][j] \vee T[i-1][j]$ 
10:          $j' \leftarrow j - X[i]$ 
11:         if  $l \leq j' \leq u$  then
12:              $T[i][j] \leftarrow T[i][j] \vee T[i-1][j']$ 
13:  return  $T[n][s]$ 

```

表格的列索引  $j$  不从 0 开始, 而是从  $l$  到  $u$ 。在实际编程环境中可以用  $j-l$  转换。接下来要利用  $T$  反演出所有  $S$  满足  $\sum S = s$ 。若  $T[n][s] = F$  则无解; 否则有两种情况: (1) 如果  $x_n = s$ , 则子集  $\{x_n\}$  是一个解。我们接下来检查  $T[n-1][s]$  的真假, 如果为  $T$ , 就递归从  $\{x_1, x_2, x_3, \dots, x_{n-1}\}$  中反演和为  $s$  的所有子集。(2) 令  $s' = s - x_n$ , 如果  $l \leq s' \leq u$ , 并且  $T[n-1][s']$  为真, 我们递归从  $\{x_1, x_2, x_3, \dots, x_{n-1}\}$  反演和  $s'$ , 然后将  $x_n$  加入到每个子集中。

```

1: function GET( $X, s, T, n$ )
2:    $r \leftarrow []$ 
3:   if  $X[n] = s$  then
4:        $r \leftarrow \{X[n]\} : r$ 
5:   if  $n > 1$  then
6:       if  $T[n-1][s]$  then
7:            $r \leftarrow r \# \text{GET}(X, s, T, n-1)$ 
8:            $s' \leftarrow s - X[n]$ 
9:           if  $l \leq s' \leq u$  且  $T[n-1][s']$  then
10:               $r \leftarrow r \# [(X[n]:r') | r' \leftarrow \text{GET}(X, s', T, n-1)]$ 
11:  return  $r$ 

```

动态规划法循环了  $O(n(u-l+1))$  次以构建表格  $T$ , 然后递归  $O(n)$  层反演解。二维表格需要  $O(n(u-l+1))$  的空间。我们可以用一维  $u-l+1$  元列表  $V$  代替二维表格。每一项  $V[j] = \{S_1, S_2, \dots\}$  存储不同的子集, 满足  $\sum S_1 = \sum S_2 = \dots = j$ 。  $V$  中所有项初始为空。对每个  $x_i$  更新一轮  $V$ , 记录加入  $x_i$  后能得到的子集之和。最终答案在  $V[s]$  中。

```

1: function SUBSET-SUM( $X, s$ )
2:    $l \leftarrow \sum \{x \in X, x < 0\}, u \leftarrow \sum \{x \in X, x > 0\}$ 
3:    $V \leftarrow [\emptyset, \emptyset, \dots]$   $\triangleright u-l+1$ 
4:   for each  $x$  in  $X$  do

```



```

5:      U ← COPY(V)
6:      for j ← l to u do
7:          if x = j then
8:              U[j] ← {{x}} ∪ U[j]
9:              j' ← j - x
10:             if l ≤ j' ≤ u 且 V[j'] ≠ ∅ then
11:                 U[j] ← U[j] ∪ {{x} ∪ S | S ∈ V[j']}
12:     V ← U
13:     return V[s]

```

我们可以用左侧叠加构建解列表  $V = \text{foldl bld } (\text{replicate } (u - l + 1) \ \emptyset) \ X$ , 其中  $\text{replicate } n \ a$  产生长度为  $n$  的列表:  $[a, a, \dots, a]$ 。  $\text{bld}$  用  $X$  中每个元素更新  $V$ 。

$$\text{bld } V \ x = \text{foldl } f \ V \ [l, l + 1, \dots, u] \quad (15.41)$$

其中:

$$f \ V \ j = \begin{cases} j = x : & V[j] \cup \{x\} \\ l \leq j' \leq u \text{ 且 } T[j'] \neq \emptyset : & V[j] \cup \{x\} \cup S \mid S \in T[j'] \}, \text{ 其中 } j' = j - x \\ \text{否则} : & V \end{cases} \quad (15.42)$$

### 练习 15.9

15.9.1. 对最长公共子序列问题, 另一种自底向上的解法是子表格中记录子序列的长度和“方向”。‘N’代表向北, ‘W’代表向西, ‘NW’代表向西北。这些方向指示我们如何构建最终的结果。我们从表格的右下角开始, 如果值为‘NW’, 我们就沿着对角线移动到左上方的格子; 如果值为‘N’, 就垂直移动到上方的格子; 如果为‘W’, 就水平移动到左侧的格子。编程实现这一算法。

15.9.2. 对于子集和的上下限, 一定有  $l \leq 0 \leq u$  成立么? 上下限能进一步缩小么?

15.9.3. 编辑距离 (又称莱文斯坦编辑距离) 是一种度量字符串相似程度的量。它定义为从字符串  $s$  转换到  $t$  所需成本。广泛用于拼写检查、手写识别中。编辑距离包含三种字符操作: 增加、删除、替换, 每次一个字符, 如下:  $\text{kitten} \mapsto \text{sitting}$ , 编辑距离为 3。

1.  $\text{kitten} \rightarrow \text{sitten}$  ( $k \mapsto s$ );
2.  $\text{sitten} \rightarrow \text{sittin}$  ( $e \mapsto i$ );
3.  $\text{sittin} \rightarrow \text{sitting}$  ( $+ g$ )。

使用动态规划, 求两个字符串间编辑距离。

## 15.4 附录:例子程序

深度优先走迷宫:

```
dfsSolve m from to = solve [[from]] where
  solve [] = []
  solve (c@(p:path):cs)
    | p == to = reverse c
    | otherwise = let os = filter (`notElem` path) (adj p) in
      if os == [] then solve cs
      else solve ((map (:c) os) # cs)
  adj (x, y) = [(x', y') | (x', y') ← [(x-1, y), (x+1, y), (x, y-1), (x, y+1)],
    inRange (bounds m) (x', y'), m ! (x', y') == 0]
```

八皇后问题:

```
solve = dfsSolve [[]] [] where
  dfsSolve [] s = s
  dfsSolve (c:cs) s
    | length c == 8 = dfsSolve cs (c:s)
    | otherwise = dfsSolve (([x:c | x ← [1..8]] \\ c,
      not $ attack x c) # cs) s
  attack x c = let y = 1 + length c in
    any ( $\lambda(i, j) \rightarrow \mathbf{abs}(x - i) == \mathbf{abs}(y - j)$ ) $
      zip (reverse c) [1..]
```

跳跃的青蛙:

```
solve = dfsSolve [[[-1, -1, -1, 0, 1, 1, 1]]] [] where
  dfsSolve [] s = s
  dfsSolve (c:cs) s
    | head c == [1, 1, 1, 0, -1, -1, -1] = dfsSolve cs (reverse c:s)
    | otherwise = dfsSolve ((map (:c) $ moves $ head c) # cs) s
  moves s = filter ( $\neq$  s) [leapLeft s, hopLeft s, leapRight s, hopRight s] where
    leapLeft [] = []
    leapLeft (0:y:1:ys) = 1:y:0:ys
    leapLeft (y:ys) = y:leapLeft ys
    hopLeft [] = []
    hopLeft (0:1:ys) = 1:0:ys
    hopLeft (y:ys) = y:hopLeft ys
    leapRight [] = []
    leapRight (-1:y:0:ys) = 0:y:(-1):ys
    leapRight (y:ys) = y:leapRight ys
    hopRight [] = []
    hopRight (-1:0:ys) = 0:(-1):ys
    hopRight (y:ys) = y:hopRight ys
```

跳跃青蛙问题的命令式实现:

```
[Int] solve([Int] start, [Int] end) {
  stack = [[start]]
  s = []
  while stack  $\neq$  [] {
```

```

    c = pop(stack)
    if c[0] == end {
      s += reverse(c)
    } else {
      for [Int] m in moves(c[0]) {
        stack += (m:c)
      }
    }
  }
}
return s
}

[[Int]] moves([[Int] s) {
  [[Int]] ms = []
  n = length(s)
  p = find(s, 0)
  if p < n - 2 and s[p+2] > 0 then ms += swap(s, p, p+2)
  if p < n - 1 and s[p+1] > 0 then ms += swap(s, p, p+1)
  if p > 1 and s[p-2] < 0 then ms += swap(s, p, p-2)
  if p > 0 and s[p-1] < 0 then ms += swap(s, p, p-1)
  return ms
}

[Int] swap([Int] s, Int i, Int j) {
  a = copy(s)
  (a[i], a[j]) = (a[j], a[i])
  return a
}

```

过河问题:

```

import Data.Bits
import qualified Data.Sequence as Queue
import Data.Sequence (Seq((:<|)), (×))

solve = bfsSolve $ Queue.singleton [(15, 0)] where
  bfsSolve Queue.Empty = [] — no solution
  bfsSolve (c@(p:_) :<| cs)
    | fst p == 0 = reverse c
    | otherwise = bfsSolve (cs × (Queue.fromList $ map (:c)
      (filter (`valid` c) $ moves p)))

valid (a, b) r = not $ or [ a `elem` [3, 6], b `elem` [3, 6], (a, b) `elem` r ]

moves (a, b) = if b < 8 then trans a b else map swap (trans b a) where
  trans x y = [(x - 8 - i, y + 8 + i)
    | i ← [0, 1, 2, 4], i == 0 || (x .&. i) ≠ 0]
  swap (x, y) = (y, x)

```

利用扩展欧几里得算法求解倒水问题:

```

extGcd 0 b = (b, 0, 1)
extGcd a b = let (d, x', y') = extGcd (b `mod` a) a in
  (d, y' - x' * (b `div` a), x')

```

```

solve a b g | g `mod` d ≠ 0 = []
      | otherwise = solve' (x * g `div` d)
  where
    (d, x, y) = extGcd a b
    solve' x | x < 0 = solve' (x + b)
      | otherwise = pour x [(0, 0)]
    pour 0 ps = reverse ((0, g):ps)
    pour x ps@((a', b'):_ ) | a' == 0 = pour (x - 1) ((a, b'):ps)
      | b' == b = pour x ((a', 0):ps)
      | otherwise = pour x ((max 0 (a' + b' - b),
                              min (a' + b') b):ps)

```

倒水问题的广度优先解法:

```

import qualified Data.Sequence as Queue
import Data.Sequence (Seq(:<|), (×))

solve' a b g = bfs $ Queue.singleton [(0, 0)] where
  bfs Queue.Empty = []
  bfs (c@(p:_ ) :<| cs)
    | fst p == g || snd p == g = reverse c
    | otherwise = bfs (cs × (Queue.fromList $ map (:c) $ expand c))
  expand ((x, y):ps) = filter (`notElem` ps) $ map (λf → f x y)
    [fillA, fillB, pourA, pourB, emptyA, emptyB]
  fillA _ y = (a, y)
  fillB x _ = (x, b)
  emptyA _ y = (0, y)
  emptyB x _ = (x, 0)
  pourA x y = (max 0 (x + y - b), min (x + y) b)
  pourB x y = (min (x + y) a, max 0 (x + y - a))

```

倒水问题的迭代式广度优先实现:

```

data Step {
  Pair<Int> (p, q)
  Step parent
  Step(Pair<Int>(x, y), Step p = null) {
    (p, q) = (x, y), parent = p
  }
}

Bool (==) (Step a, Step b) = {a.(p, q) == b.(p, q)}
Bool (≠) (Step a, Step b) = not o (==)

[Step] expand(Step s, Int a, Int b) {
  var (p, q) = s.(p, q)
  return [Step(a, q, s), /*fill A*/
    Step(p, b, s), /*fill B*/
    Step(0, q, s), /*empty A*/
    Step(p, 0, s), /*empty B*/
    Step(max(0, p + q - b), min(p + q, b), s), /*pour A into B*/
    Step(min(p + q, a), max(0, p + q - a), s)] /*pour B into A*/
}

```

```

Optional<[Step]> solve(Int a, Int b, Int g) {
  q = Queue<Step>(Step(0, 0))
  Set<Step> visited = {head(q)}
  while not empty(q) {
    var cur = pop(q)
    if cur.p == g || cur.q == g {
      return Optional.of(backtrack(cur))
    } else {
      for s in expand(cur, a, b) {
        if cur ≠ s and s not in visited {
          push(q, s)
          visited += s
        }
      }
    }
  }
  return Optional.Nothing
}

[Step] backtrack(Step s) {
  [Step] seq
  while s ≠ null {
    seq = s : seq
    s = s.parent
  }
  return seq
}

```

华容道:

```

import qualified Data.Map as Map
import qualified Data.Set as Set
import qualified Data.Sequence as Queue
import Data.Sequence (Seq(:<|), (×))

cellOf (y, x) = y * 4 + x
posOf c = (c `div` 4, c `mod` 4)

cellSet = Set.fromList o (map cellOf)

type Layout = Map.Map Integer (Set.Set Integer)
type NormLayout = Set.Set (Set.Set Integer)
type Move = (Integer, Integer)

start = Map.map cellSet $ Map.fromList
  [(1, [(0, 0), (1, 0)]),
   (2, [(0, 3), (1, 3)]),
   (3, [(2, 0), (3, 0)]),
   (4, [(2, 1), (2, 2)]),
   (5, [(2, 3), (3, 3)]),
   (6, [(4, 0)]), (7, [(3, 1)]), (8, [(3, 2)]), (9, [(4, 3)]),
   (10, [(0, 1), (0, 2), (1, 1), (1, 2)])]

```

```

end = cellSet [(3, 1), (3, 2), (4, 1), (4, 2)]

normalize = Set.fromList ◦ Map.elems

mirror = Map.map (Set.map f) where
  f c = let (y, x) = posOf c in cellOf (y, 3 - x)

klotski = solve q visited where
  q = Queue.singleton (start, [])
  visited = Set.singleton (normalize start)

solve Queue.Empty _ = []
solve ((x, ms) :<| cs) visited | Map.lookup 10 x == Just end = reverse ms
  | otherwise = solve q visited'

where
  q = cs × (Queue.fromList [(move x op, op:ms) | op ← ops ])
  visited' = foldr Set.insert visited (map (normalize ◦ move x) ops)
  ops = expand x visited

expand x visited = [(i, d) | i ← [1..10], d ← [-1, 1, -4, 4],
  valid i d, unique i d]

where
  valid i d = let p = trans d (maybe Set.empty id $ Map.lookup i x) in
    (not $ any (outside d) p) &&
    (Map.keySet $ Map.filter (overlapped p) x)
    `Set.isSubsetOf` Set.singleton i
  outside d c = c < 0 || c ≥ 20 ||
    (d == 1 && c `mod` 4 == 0) || (d == -1 && c `mod` 4 == 3)
  unique i d = let ly = move x (i, d) in all (`Set.notMember` visited)
    [normalize ly, normalize (mirror ly)]

move x (i, d) = Map.update (Just ◦ trans d) i x

trans d = Set.map (d+)

overlapped :: (Set.Set Integer) → (Set.Set Integer) → Bool
overlapped a b = (not ◦ Set.null) $ Set.intersection a b

```

华容道的迭代式解:

```

type Layout = [Set<Int>]

Layout START = [{0, 4}, {3, 7}, {8, 12}, {9, 10},
  {11, 15},{16},{13}, {14}, {19}, {1, 2, 5, 6}]

Set<Int> END = {13, 14, 17, 18}

(Int, Int) pos(Int c) = (y = c / 4, x = c mod 4)

[[Int]] matrix(Layout layout) {
  [[Int]] m = replicate(replicate(0, 4), 5)
  for Int i, var p in (zip([1, 2, ...], layout)) {

```

```

        for var c in p {
            y, x = pos(c)
            m[y][x] = i
        }
    }
    return m
}

data Node {
    Node parent
    Layout layout

    Node(Layout l, Node p = null) {
        layout = l, parent = p
    }
}

//usage: solve(START, END)
Optional<Node> solve(Layout start, Set<Int> end) {
    var visit = {Set(start)}
    var queue = Queue.of(Node(start))
    while not empty(queue) {
        cur = pop(queue)
        if last(cur.layout) == end {
            return Optional.of(cur)
        } else {
            for ly in expand(cur.layout, visit) {
                push(queue, Node(ly, cur))
                add(visit, Set(ly))
            }
        }
    }
    return Optional.None
}

[Layout] expand(Layout layout, Set<Set<Layout>> visit):
    Bool bound(Set<Int> piece, Int d) {
        for c in piece {
            if c + d < 0 or c + d ≥ 20 then return False
            if d == 1 and c mod 4 == 3 then return False
            if d == -1 and c mod 4 == 0 then return False
        }
        return True
    }

    var m = matrix(layout)
    Bool valid(Set<Int> piece, Int d, Int i) {
        for c in piece {
            y, x = pos(c + d)
            if m[y][x] not in [0, i] then return False
        }
        return True
    }
}

```

```

Bool unique(Layout ly) {
  n = Set(ly)
  Set<Set<Int>> m = map(map(c → 4 * (c / 4) + 3 - (c mod 4), p), n)
  return (n not in visit) and (m not in visit)
}

[Layout] s = []
for i, p in zip([1, 2, ...], layout) {
  for d in [-1, 1, -4, 4] {
    if bound(p, d) and valid(p, d, i) {
      ly = move(layout, i - 1, d)
      if unique(ly) then s.append(ly)
    }
  }
}
return
}

Layout move(Layout layout, Int i, Int d) {
  ly = clone(layout)
  ly[i] = map((d+), layout[i])
  return ly
}

```

从哈夫曼树构造码表,编解码:

```

code = Map.fromList ◦ (traverse []) where
  traverse bits (Leaf _ c) = [(c, reverse bits)]
  traverse bits (Branch _ l r) = traverse (0:bits) l # traverse (1:bits) r

encode dict = concatMap (dict !)

decode tr cs = find tr cs where
  find (Leaf _ c) [] = [c]
  find (Leaf _ c) bs = c : find tr bs
  find (Branch _ l r) (b:bs) = find (if b == 0 then l else r) bs

```

使用贪心策略兑换硬币:

```

import qualified Data.Set as Set
import Data.List (group)

solve x = assoc ◦ change x where
  change 0 _ = []
  change x cs = let c = Set.findMax $ Set.filter (≤ x) cs in c : change (x - c) cs
  assoc = (map (λcs → (head cs, length cs))) ◦ group

example = solve 142 $ Set.fromList [1, 5, 25, 50, 100]

```

动态规划法求换零钱问题最优解:

```

[Int] change(Int x, Set<Int> cs) {
  t = [(0, None)] ++ [(x + 1, None)] * x

```



```

for i = 1 to x {
  for c in cs {
    if c ≤ i {
      (n, _) = t[i - c]
      (m, _) = t[i]
      if 1 + n < m then t[i] = (1 + n, c)
    }
  }
  s = []
  while x > 0:
    (_, c) = t[x]
    s += c
    x = x - c
  return s
}

```

叠加构建换零钱问题最优解:

```

import qualified Data.Set as Set
import Data.Sequence ((|>), singleton, index)

changemk x cs = makeChange x $ foldl fill (singleton (0, 0)) [1..x] where
  fill tab i = tab |> (n, c) where
    (n, c) = minimum $ Set.map lookup $ Set.filter (≤ i) cs
    lookup c = (1 + fst (tab `index` (i - c)), c)
  makeChange 0 _ = []
  makeChange x tab = let c = snd $ tab `index` x in c : makeChange (x - c) tab

```

最长公共子序列:

```

[K] lcs([K] xs, [K] ys) {
  Int m = length(xs), n = length(ys)
  [[Int]] c = [[0]*(n + 1)]*(m + 1)
  for i = 1 to m {
    for j = 1 to n {
      if xs[i-1] == ys[j-1] {
        c[i][j] = c[i-1][j-1] + 1
      } else {
        c[i][j] = max(c[i-1][j], c[i][j-1])
      }
    }
  }
  return fetch(c, xs, ys)
}

[K] fetch([[Int]] c, [K] xs, [K] ys) {
  [K] r = []
  var m = length(xs), n = length(ys)
  while m > 0 and n > 0 {
    if xs[m - 1] == ys[n - 1] {
      r += xs[m - 1]
      m = m - 1
      n = n - 1
    } else if c[m - 1][n] > c[m][n - 1] {
      m = m - 1
    } else {

```

```

        n = n - 1
    }
}
return reverse(r)
}

```

子集和判定问题:

```

Bool subsetsum([Int] xs, Int s) {
    Int l = 0, u = 0, n = length(xs)
    for x in xs {
        if x > 0 then u++ else l++
    }
    tab = [[False]*(u - l + 1)] * (n + 1)
    tab = [0][0 - l] = True
    for i, x in zip([1, 2, ..., n], xs) {
        tab[i][x - l] = True
        for j = l to u {
            tab[i][j - l] or = tab[i-1][j - l]
            j1 = j - x
            if l ≤ j1 ≤ u then tab[i][j - l] or = tab[i-1][j1 - l]
        }
    }
    return tab[n][s - l]
}

```

用一维列表求子集和问题的解:

```

{{Int}} subsetsum(xs, s) {
    Int l = 0, u = 0, n = length(xs)
    for x in xs {
        if x > 0 then u++ else l++
    }
    tab = {} * (u - l + 1)
    for x in xs {
        tab1 = copy(tab)
        for j = low to up {
            if x == j then add(tab1[j], {x})
            j1 = j - x
            if low ≤ j1 ≤ up and tab[j1] {
                tab1[j] |= {add(ys, x) for ys in tab[j1]}
            }
        }
        tab = tab1
    }
    return tab[s]
}

```

## 附录 A 红黑树的命令式删除算法

和插入相比,红黑树的命令式删除算法需要处理更多的情况。在普通二叉搜索树的删除算法之上,我们通过旋转和重新染色恢复平衡性。删除黑色节点会破坏红黑树的第五条性质,使得某一路径上的黑色节点数目减少。为此,我们引入“双重黑色”,来保持黑色节点数目不变。下面的例子程序增加了双重黑色的定义:

```
data Color {RED, BLACK, DOUBLY_BLACK}
```

我们首先复用二叉搜索树的删除算法,并记录被删除节点的父节点。如果被删除节点的颜色是黑色,我们通过双重黑色保持性质 5,然后再做进一步修复。

```
1: function DELETE( $T, x$ )
2:    $p \leftarrow$  PARENT( $x$ )
3:    $q \leftarrow$  NIL
4:   if LEFT( $x$ ) = NIL then
5:      $q \leftarrow$  RIGHT( $x$ )
6:     REPLACE( $x$ , RIGHT( $x$ ))           ▷ 用右子树替换  $x$ 
7:   else if RIGHT( $x$ ) = NIL then
8:      $q \leftarrow$  LEFT( $x$ )
9:     REPLACE( $x$ , LEFT( $x$ ))           ▷ 用左子树替换  $x$ 
10:  else
11:     $y \leftarrow$  MIN(RIGHT( $x$ ))
12:     $p \leftarrow$  PARENT( $y$ )
13:     $q \leftarrow$  RIGHT( $y$ )
14:    KEY( $x$ )  $\leftarrow$  KEY( $y$ )
15:    copy data from  $y$  to  $x$ 
16:    REPLACE( $y$ , RIGHT( $y$ ))           ▷ 用右子树替换  $y$ 
17:     $x \leftarrow y$ 
18:  if COLOR( $x$ ) = BLACK then
19:     $T \leftarrow$  DELETE-FIX( $T$ , MAKE-BLACK( $p, q$ ),  $q =$  NIL?)
20:  release  $x$ 
21:  return  $T$ 
```

删除算法接受两个输入:根节点  $T$  和待删除节点  $x$ 。 $x$  可以通过查找定位到。如果  $x$  含有为空的子树,我们将  $x$ “切下”,并用另一子树  $q$  来替代  $x$ 。否则,我们在  $x$  的右子树中找到最小元素  $y$ ,用  $y$  替换  $x$ 。然后递归地将  $y$ “切下”。如果  $x$  是黑色的,我们调用  $\text{MAKE-BLACK}(p, q)$  保持黑色属性,以便进行下一步的修复。

```

1: function MAKE-BLACK( $p, q$ )
2:   if  $p = \text{NIL}$  and  $q = \text{NIL}$  then
3:     return NIL
4:   else if  $q = \text{NIL}$  then
5:      $n \leftarrow \text{Doubly Black NIL}$ 
6:      $\text{PARENT}(n) \leftarrow p$ 
7:     return  $n$ 
8:   else
9:     return BLACKEN( $q$ )

```

▷ 树中只有一个叶子节点

如果  $p$  和  $q$  都为空,我们在删除只有一个叶子节点的树,树变为空。如果父节点  $p$  不为空,而  $q$  为空,说明删除了一个黑色叶子节点。我们用  $\text{NIL}$  替换掉它。根据红黑树性质 3,  $\text{NIL}$  是黑色的。我们把这一  $\text{NIL}$  变成“双重黑色”的  $\text{NIL}$  来保持性质 5 仍然成立。如果  $p, q$  都不为空,我们调用  $\text{BLACKEN}(q)$  检查  $q$  的颜色,如果是红色的,将它染成黑色,如果  $q$  已经是黑色的,将它染成双重黑色。接下来,我们通过旋转和重新染色,最终去掉“双重黑色”。这里有三种情况需要处理([4], 292 页)。每种情况中,双重黑色的节点即可以是普通节点,也可以是双重黑色  $\text{NIL}$ 。

**情况 1:** 双重黑色节点的兄弟为黑色,并且该兄弟节点有一个红色子节点。我们可以通过旋转来修复。共有四种细分情况,它们全部可以变换到一种统一形式。如图 A.1 所示。

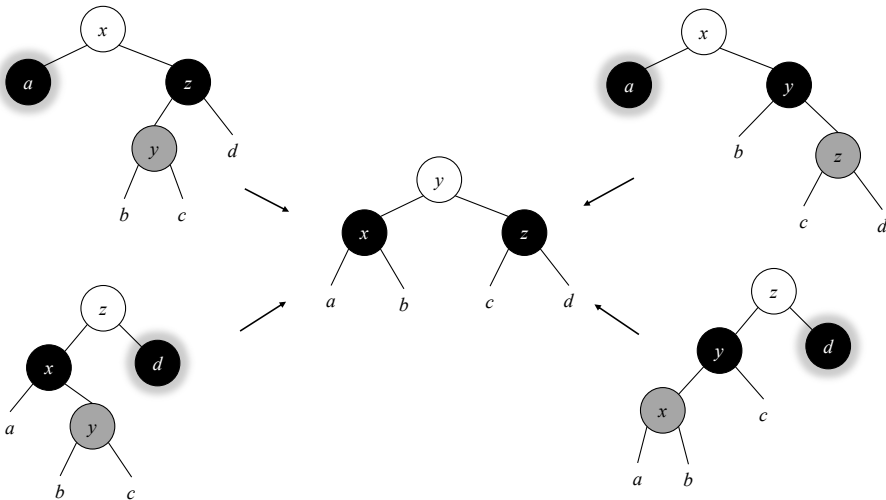


图 A.1: 双重黑色节点的兄弟为黑色,并且该兄弟节点有一个红色子节点。通过一次旋转操作修复。

```

1: function DELETE-FIX( $T, x, f$ )
2:    $n \leftarrow \text{NIL}$ 
3:   if  $f = \text{True}$  then                                ▷  $x$  是双重黑色 NIL
4:      $n \leftarrow x$ 
5:   if  $x = \text{NIL}$  then                                ▷ 删除唯一的叶子
6:     return NIL
7:   while  $x \neq T$  and  $\text{COLOR}(x) = \mathcal{B}^2$  do      ▷  $x$  是双重黑色,但不是根节点
8:     if  $\text{SIBLING}(x) \neq \text{NIL}$  then                ▷ 兄弟节点不为空
9:        $s \leftarrow \text{SIBLING}(x)$ 
10:    ...
11:    if  $s$  is black and  $\text{LEFT}(s)$  is red then
12:      if  $x = \text{LEFT}(\text{PARENT}(x))$  then              ▷  $x$  在左侧
13:        set  $x, \text{PARENT}(x)$ , and  $\text{LEFT}(s)$  all black
14:         $T \leftarrow \text{ROTATE-RIGHT}(T, s)$ 
15:         $T \leftarrow \text{ROTATE-LEFT}(T, \text{PARENT}(x))$ 
16:      else   ▷  $x$  在右侧
17:        set  $x, \text{PARENT}(x), s$ , and  $\text{LEFT}(s)$  all black
18:         $T \leftarrow \text{ROTATE-RIGHT}(T, \text{PARENT}(x))$ 
19:      else if  $s$  is black and  $\text{RIGHT}(s)$  is red then
20:        if  $x = \text{LEFT}(\text{PARENT}(x))$  then            ▷  $x$  在左侧
21:          set  $x, \text{PARENT}(x), s$ , and  $\text{RIGHT}(s)$  all black
22:           $T \leftarrow \text{ROTATE-LEFT}(T, \text{PARENT}(x))$ 
23:        else   ▷  $x$  在右侧
24:          set  $x, \text{PARENT}(x)$ , and  $\text{RIGHT}(s)$  all black
25:           $T \leftarrow \text{ROTATE-LEFT}(T, s)$ 
26:           $T \leftarrow \text{ROTATE-RIGHT}(T, \text{PARENT}(x))$ 
27:    ...

```

**情况 2:** 双重黑色节点的兄弟节点为红色。可以通过旋转,将双重黑色恢复为普通黑色。如图图 A.2 所示, $a$  或  $c$  恢复为黑色。

我们在此前给出算法上增加这一处理。

```

1: function DELETE-FIX( $T, x, f$ )
2:    $n \leftarrow \text{NIL}$ 
3:   if  $f = \text{True}$  then                                ▷  $x$  是双重黑色 NIL
4:      $n \leftarrow x$ 
5:   if  $x = \text{NIL}$  then                                ▷ 删除唯一的叶子
6:     return NIL
7:   while  $x \neq T$  and  $\text{COLOR}(x) = \mathcal{B}^2$  do

```

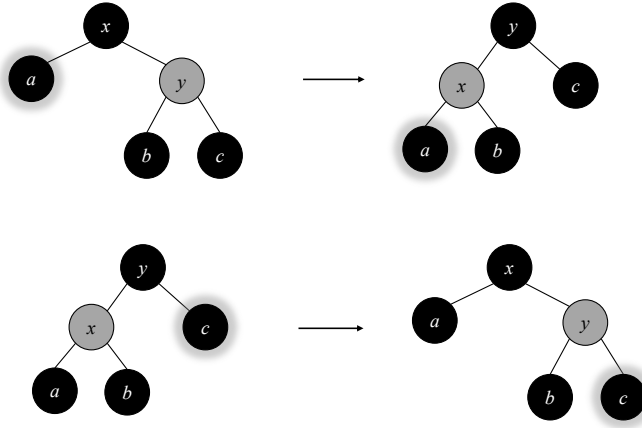


图 A.2: 双重黑色节点的兄弟节点为红色

```

8:   if SIBLING( $x$ )  $\neq$  NIL then
9:        $s \leftarrow$  SIBLING( $x$ )
10:      if  $s$  is red then                                     ▷ 兄弟节点为红色
11:          set PARENT( $x$ ) red
12:          set  $s$  black
13:          if  $x =$  LEFT(PARENT( $x$ )) then                       ▷  $x$  在左侧
14:               $T \leftarrow$  ROTATE-LEFT $T$ , PARENT( $x$ )
15:          else   ▷  $x$  在右侧
16:               $T \leftarrow$  ROTATE-RIGHT $T$ , PARENT( $x$ )
17:          else if  $s$  is black and LEFT( $s$ ) is red then
18:              ...

```

**情况 3:** 双重黑色节点的兄弟节点为黑色, 该兄弟节点的两个子节点也全是黑色。可以将兄弟节点染成红色, 将双重黑色变回黑色, 然后将黑色向上传递。如图图 A.3 所示, 有两种对称的情况。

上述三种情况中, 双重黑色节点的兄弟节点都不为空。否则, 我们直接将双重黑色恢复为普通黑色, 然后将黑色向上传递。如果最终到达根节点, 我们将根节点变为普通黑色, 并结束修复过程。另外, 如果在双重黑色在修复过程中被消除, 也可以终止。最后, 针对双重黑色 NIL 的情况, 我们将其恢复为普通 NIL。

```

1: function DELETE-FIX( $T$ ,  $x$ ,  $f$ )
2:      $n \leftarrow$  NIL
3:     if  $f =$  True then                                       ▷  $x$  是双重黑色 NIL
4:          $n \leftarrow x$ 
5:     if  $x =$  NIL then   ▷ 删除唯一的叶子
6:         return NIL
7:     while  $x \neq T$  and COLOR( $x$ ) =  $\mathcal{B}^2$  do

```

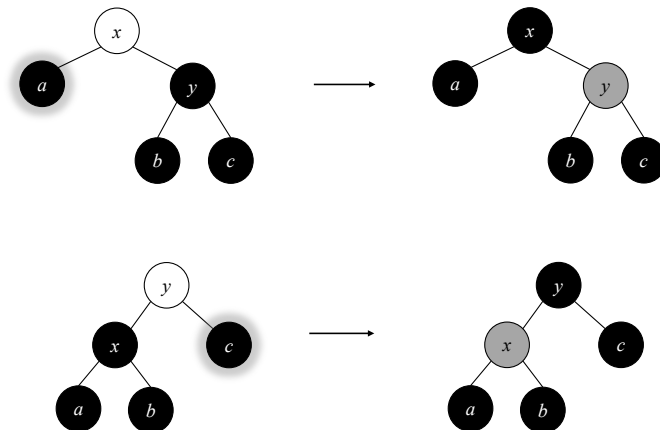


图 A.3: 向上传递黑色

```

8:   if SIBLING( $x$ )  $\neq$  NIL then                                ▷ 兄弟节点不为空
9:        $s \leftarrow$  SIBLING( $x$ )
10:   if  $s$  is red then  ▷ 兄弟节点为红色
11:       set PARENT( $x$ ) red
12:       set  $s$  black
13:       if  $x =$  LEFT(PARENT( $x$ )) then                            ▷  $x$  在左侧
14:            $T \leftarrow$  ROTATE-LEFT( $T$ , PARENT( $x$ ))
15:       else  ▷  $x$  在右侧
16:            $T \leftarrow$  ROTATE-RIGHT( $T$ , PARENT( $x$ ))
17:   else if  $s$  is black and LEFT( $s$ ) is red then
18:       if  $x =$  LEFT(PARENT( $x$ )) then                            ▷  $x$  在左侧
19:           set  $x$ , PARENT( $x$ ), and LEFT( $s$ ) all black
20:            $T \leftarrow$  ROTATE-RIGHT( $T$ ,  $s$ )
21:            $T \leftarrow$  ROTATE-LEFT( $T$ , PARENT( $x$ ))
22:       else  ▷  $x$  在右侧
23:           set  $x$ , PARENT( $x$ ),  $s$ , and LEFT( $s$ ) all black
24:            $T \leftarrow$  ROTATE-RIGHT( $T$ , PARENT( $x$ ))
25:   else if  $s$  is black and RIGHT( $s$ ) is red then
26:       if  $x =$  LEFT(PARENT( $x$ )) then                            ▷  $x$  在左侧
27:           set  $x$ , PARENT( $x$ ),  $s$ , and RIGHT( $s$ ) all black
28:            $T \leftarrow$  ROTATE-LEFT( $T$ , PARENT( $x$ ))
29:       else  ▷  $x$  在右侧
30:           set  $x$ , PARENT( $x$ ), and RIGHT( $s$ ) all black
31:            $T \leftarrow$  ROTATE-LEFT( $T$ ,  $s$ )
32:            $T \leftarrow$  ROTATE-RIGHT( $T$ , PARENT( $x$ ))

```

```

33:         else if  $s$ , LEFT( $s$ ), and RIGHT( $s$ ) are all black then
34:             set  $x$  black
35:             set  $s$  red
36:             BLACKEN(PARENT( $x$ ))
37:              $x \leftarrow$  PARENT( $x$ )
38:         else ▷ 向上传递黑色
39:             set  $x$  black
40:             BLACKEN(PARENT( $x$ ))
41:              $x \leftarrow$  PARENT( $x$ )
42:     set  $T$  black
43:     if  $n \neq$  NIL then
44:         replace  $n$  with NIL
45:     return  $T$ 

```

修复时,我们传入三个参数:根节点  $T$ 、待修复节点  $x$ (可能是双重黑色)、标记  $f$ 。如果  $x$  是双重黑色 NIL,则  $f$  为真。此时我们用  $n$  来记录它,并在最终修复完毕后,用普通 NIL 替换  $n$ 。

下面的例子程序实现了红黑树删除算法。

```

Node del(Node t, Node x) {
    if  $x ==$  null then return t
    var parent = x.parent;
    Node db = null; //doubly black

    if x.left == null {
        db = x.right
        x.replaceWith(db)
    } else if x.right == null {
        db = x.left
        x.replaceWith(db)
    } else {
        var y = min(x.right)
        parent = y.parent
        db = y.right
        x.key = y.key
        y.replaceWith(db)
        x = y
    }
    if x.color == Color.BLACK {
        t = deleteFix(t, makeBlack(parent, db), db == null);
    }
    remove(x)
    return t
}

```

其中 `makeBlack` 检查删除后节点是否变为双重黑色,并处理双重黑色 NIL 的特殊情况。



```

Node makeBlack(Node parent, Node x) {
    if parent == null and x == null then return null
    return if x == null
        then replace(parent, x, Node(0, Color.DOUBLY_BLACK))
        else blacken(x)
}

```

其中函数 `replace(parent, x, y)` 将 `parent` 的子节点 `x`, 用 `y` 替换。

```

Node replace(Node parent, Node x, Node y) {
    if parent == null {
        if y != null then y.parent = null
    } else if parent.left == x {
        parent.setLeft(y)
    } else {
        parent.setRight(y)
    }
    if x != null then x.parent = null
    return y
}

```

函数 `blacken(node)` 将红色节点染为黑色, 将黑色节点染为双重黑色。

```

Node blacken(Node x) {
    x.color = if isRed(x) then Color.BLACK else Color.DOUBLY_BLACK
    return x
}

```

下面的例子程序实现了修复过程:

```

Node deleteFix(Node t, Node db, Bool isDBEmpty) {
    var dbEmpty = if isDBEmpty then db else null
    if db == null then return null // delete the root
    while (db != t and db.color == Color.DOUBLY_BLACK) {
        var s = db.sibling()
        var p = db.parent
        if (s != null) {
            if isRed(s) {
                // the sibling is red
                p.color = Color.RED
                s.color = Color.BLACK
                t = if db == p.left then leftRotate(t, p)
                    else rightRotate(t, p)
            } else if isBlack(s) and isRed(s.left) {
                // the sibling is black, and one sub-tree is red
                if db == p.left {
                    db.color = Color.BLACK
                    p.color = Color.BLACK
                    s.left.color = p.color
                    t = rightRotate(t, s)
                    t = leftRotate(t, p)
                } else {
                    db.color = Color.BLACK

```

```

        p.color = Color.BLACK
        s.color = p.color
        s.left.color = Color.BLACK
        t = rightRotate(t, p)
    }
} else if isBlack(s) and isRed(s.right) {
    if (db == p.left) {
        db.color = Color.BLACK
        p.color = Color.BLACK
        s.color = p.color
        s.right.color = Color.BLACK
        t = leftRotate(t, p)
    } else {
        db.color = Color.BLACK
        p.color = Color.BLACK
        s.right.color = p.color
        t = leftRotate(t, s)
        t = rightRotate(t, p)
    }
} else if isBlack(s) and isBlack(s.left) and
    isBlack(s.right) {
    // the sibling and both sub-trees are black.
    // move blackness up
    db.color = Color.BLACK
    s.color = Color.RED
    blacken(p)
    db = p
}
} else { // no sibling, move blackness up
    db.color = Color.BLACK
    blacken(p)
    db = p
}
}
t.color = Color.BLACK
if (dbEmpty ≠ null) { // change the doubly black nil to nil
    dbEmpty.replaceWith(null)
    delete dbEmpty
}
return t
}
}

```

其中 `isBlack(node)` 判断一个节点是否为黑色, 根据红黑树的性质, `NIL` 也是黑色的。

```
Bool isBlack(Node x) = (x == null or x.color == Color.BLACK)
```

```
Bool isRed(Node x) = (x ≠ null and x.color == Color.RED)
```

算法在结束前修复双重黑色 `NIL`, 调用 `Node` 中的 `replaceWith` 进行替换。

```
data Node<T> {
    //...
```

```
void replaceWith(Node y) = replace(parent, this, y)  
}
```

考虑红黑树的平衡性, 删除算法或者到达根节点终止, 或者在双重黑色消失时终止。对于含有  $n$  个节点的红黑树, 其复杂度为  $O(\lg n)$ 。



# 附录 B AVL 树——证明和删除算法

## B.1 插入后的高度变化

向 AVL 树插入元素后,高度变化存在四种情况:

$$\begin{aligned}\Delta H &= |T'| - |T| \\ &= 1 + \max(|r'|, |l'|) - (1 + \max(|r|, |l|)) \\ &= \max(|r'|, |l'|) - \max(|r|, |l|) \\ &= \begin{cases} \delta \geq 0, \delta' \geq 0: & \Delta r \\ \delta \leq 0, \delta' \geq 0: & \delta + \Delta r \\ \delta \geq 0, \delta' \leq 0: & \Delta l - \delta \\ \text{否则:} & \Delta l \end{cases} \quad (\text{B.1})\end{aligned}$$

**证明.** 一次插入不可能同时增加左右分支的高度。平衡因子等于右子树的高度减去左子树的高度。根据其前后变化,共有四种情况:

1. 如果  $\delta \geq 0$  并且  $\delta' \geq 0$ , 在插入前后, 右子树的高度都不小于左子树的高度。高度的增加全部“贡献”自右子树的变化  $\Delta r$ ;
2. 如果  $\delta \leq 0$ , 在插入前左子树的高度不小于右子树, 但是插入后  $\delta' \geq 0$ , 说明右子树的高度由于插入增加了, 而左子树保持不变 ( $|l'| = |l|$ )。所以高度的增加为:

$$\begin{aligned}\Delta H &= \max(|r'|, |l'|) - \max(|r|, |l|) \quad \{\delta \leq 0 \text{ 且 } \delta' \geq 0\} \\ &= |r'| - |l| \quad \{|l| = |l'|\} \\ &= |r| + \Delta r - |l| \\ &= \delta + \Delta r\end{aligned}$$

3. 如果  $\delta \geq 0$  且  $\delta' \leq 0$ , 和情况二类似, 我们有:

$$\begin{aligned}\Delta H &= \max(|r'|, |l'|) - \max(|r|, |l|) \quad \{\delta \geq 0 \text{ 且 } \delta' \leq 0\} \\ &= |l'| - |r| \\ &= |l| + \Delta l - |r| \\ &= \Delta l - \delta\end{aligned}$$

4. 否则  $\delta$  和  $\delta'$  都不大于 0, 说明插入前后左子树的高度都不小于右子树。高度的增加全部“贡献”自左子树的变化  $\Delta l$ 。

□

## B.2 插入后的平衡调整

如图图 B.1 所示, 所有需要修复的四种情况中, 平衡因子都是  $\pm 2$ 。调整后, 平衡因子  $\delta(y)$  恢复为 0。左右子树具有相同的高度。

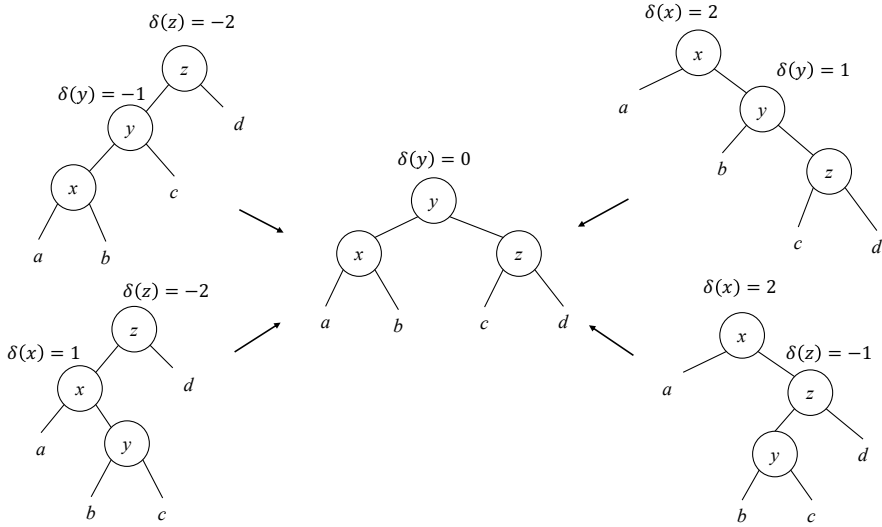


图 B.1: 插入后需要恢复平衡的 4 种情况

这 4 种情况分别是: 左-左、右-右、右-左、左-右。记修复前的平衡因子分别为  $\delta(x)$ 、 $\delta(y)$ 、 $\delta(z)$ , 修复后分别为  $\delta'(x)$ 、 $\delta'(y)$ 、 $\delta'(z)$ 。我们接下来证明, 调整后所有 4 种情况的平衡因子都变成  $\delta(y) = 0$ 。并且将给出  $\delta'(x)$  和  $\delta'(z)$  的结果。

证明. 我们分别针对四种情况证明。

### 左-左

子树  $x$  在调整前后不变, 因此  $\delta'(x) = \delta(x)$ 。因为  $\delta(y) = -1$  且  $\delta(z) = -2$ , 所以:

$$\begin{aligned}\delta(y) = |c| - |x| = -1 &\Rightarrow |c| = |x| - 1 \\ \delta(z) = |d| - |y| = -2 &\Rightarrow |d| = |y| - 2\end{aligned}\quad (\text{B.2})$$

调整后:

$$\begin{aligned}\delta'(z) &= |d| - |c| && \{\text{式 (式 (B.2))}\} \\ &= |y| - 2 - (|x| - 1) \\ &= |y| - |x| - 1 && \{x \text{ 是 } y \text{ 的子节点} \Rightarrow |y| - |x| = 1\} \\ &= 0\end{aligned}\quad (\text{B.3})$$

对于  $\delta'(y)$ , 有如下结果:

$$\begin{aligned}
 \delta'(y) &= |z| - |x| \\
 &= 1 + \max(|c|, |d|) - |x| \quad \{\text{式 (式 (B.3)), } |c| = |d|\} \\
 &= 1 + |c| - |x| \quad \{\text{式 (式 (B.2))}\} \\
 &= 1 + |x| - 1 - |x| \\
 &= 0
 \end{aligned} \tag{B.4}$$

汇总上述结果, 对于左-左情况, 新的平衡因子如下:

$$\begin{aligned}
 \delta'(x) &= \delta(x) \\
 \delta'(y) &= 0 \\
 \delta'(z) &= 0
 \end{aligned} \tag{B.5}$$

### 右-右

右-右和左-左对称, 易知新的平衡因子结果如下:

$$\begin{aligned}
 \delta'(x) &= 0 \\
 \delta'(y) &= 0 \\
 \delta'(z) &= \delta(z)
 \end{aligned} \tag{B.6}$$

### 右-左

首先考虑  $\delta'(x)$ 。调整平衡后, 我们有:

$$\delta'(x) = |b| - |a| \tag{B.7}$$

调整平衡前,  $z$  的高度为:

$$\begin{aligned}
 |z| &= 1 + \max(|y|, |d|) \quad \{\delta(z) = -1 \Rightarrow |y| > |d|\} \\
 &= 1 + |y| \\
 &= 2 + \max(|b|, |c|)
 \end{aligned} \tag{B.8}$$

因为  $\delta(x) = 2$ , 所以:

$$\begin{aligned}
 \delta(x) = 2 &\Rightarrow |z| - |a| = 2 \quad \{\text{式 (式 (B.8))}\} \\
 &\Rightarrow 2 + \max(|b|, |c|) - |a| = 2 \\
 &\Rightarrow \max(|b|, |c|) - |a| = 0
 \end{aligned} \tag{B.9}$$

如果  $\delta(y) = |c| - |b| = 1$ , 则:

$$\max(|b|, |c|) = |c| = |b| + 1 \tag{B.10}$$

将其代入式 (式 (B.9)) 得到:

$$\begin{aligned}
 |b| + 1 - |a| = 0 &\Rightarrow |b| - |a| = -1 \quad \{\text{式 (式 (B.7))}\} \\
 &\Rightarrow \delta'(x) = -1
 \end{aligned} \tag{B.11}$$

反之如果  $\delta(y) \neq 1$ , 则  $\max(|b|, |c|) = |b|$ , 代入式 (式 (B.9)) 得到:

$$\begin{aligned} |b| - |a| &= 0 \quad \{\text{式 (式 (B.7))}\} \\ \Rightarrow \delta'(x) &= 0 \end{aligned} \quad (\text{B.12})$$

合并上述两种情况, 我们得到  $\delta'(x)$  和  $\delta(y)$  的关系:

$$\delta'(x) = \begin{cases} \delta(y) = 1 : & -1 \\ \text{否则} : & 0 \end{cases} \quad (\text{B.13})$$

对于  $\delta'(z)$ , 根据定义它等于:

$$\begin{aligned} \delta'(z) &= |d| - |c| && \{\delta(z) = -1 = |d| - |y|\} \\ &= |y| - |c| - 1 && \{|y| = 1 + \max(|b|, |c|)\} \\ &= \max(|b|, |c|) - |c| \end{aligned} \quad (\text{B.14})$$

如果  $\delta(y) = |c| - |b| = -1$ , 则  $\max(|b|, |c|) = |b| = |c| + 1$ . 将其代入式 (式 (B.14)) 中得到:  $\delta'(z) = 1$ . 反之如果  $\delta(y) \neq -1$ , 则  $\max(|b|, |c|) = |c|$ , 有  $\delta'(z) = 0$ . 合并上述两种情况,  $\delta'(z)$  和  $\delta(y)$  的关系如下:

$$\delta'(z) = \begin{cases} \delta(y) = -1 : & 1 \\ \text{否则} : & 0 \end{cases} \quad (\text{B.15})$$

最后, 对于  $\delta'(y)$ , 我们可以推导出下面的关系:

$$\begin{aligned} \delta'(y) &= |z| - |x| \\ &= \max(|c|, |d|) - \max(|a|, |b|) \end{aligned} \quad (\text{B.16})$$

这里又分为三种情况:

1. 若  $\delta(y) = 0$ , 说明  $|b| = |c|$ , 根据式 (式 (B.13)) 和式 (式 (B.15)), 有:  $\delta'(x) = 0 \Rightarrow |a| = |b|$  以及  $\delta'(z) = 0 \Rightarrow |c| = |d|$ . 因此  $\delta'(y) = 0$ .
2. 若  $\delta(y) = 1$ , 根据式 (式 (B.15)), 我们有  $\delta'(z) = 0 \Rightarrow |c| = |d|$ .

$$\begin{aligned} \delta'(y) &= \max(|c|, |d|) - \max(|a|, |b|) \quad \{|c| = |d|\} \\ &= |c| - \max(|a|, |b|) \quad \{\text{式 (式 (B.13))}: \delta'(x) = -1 \Rightarrow |b| - |a| = -1\} \\ &= |c| - (|b| + 1) \quad \{\delta(y) = 1 \Rightarrow |c| - |b| = 1\} \\ &= 0 \end{aligned}$$

3. 若  $\delta(y) = -1$ , 根据式 (式 (B.13)), 我们有  $\delta'(x) = 0 \Rightarrow |a| = |b|$ .

$$\begin{aligned} \delta'(y) &= \max(|c|, |d|) - \max(|a|, |b|) \quad \{|a| = |b|\} \\ &= \max(|c|, |d|) - |b| \quad \{\text{式 (式 (B.15))}: |d| - |c| = 1\} \\ &= |c| + 1 - |b| \quad \{\delta(y) = -1 \Rightarrow |c| - |b| = -1\} \\ &= 0 \end{aligned}$$



全部三种情况的结果都是  $\delta'(y) = 0$ 。将上述结果归纳起来,可以得到新的平衡因子如下:

$$\begin{aligned} \delta'(x) &= \begin{cases} \delta(y) = 1: & -1 \\ \text{否则}: & 0 \end{cases} \\ \delta'(y) &= 0 \\ \delta'(z) &= \begin{cases} \delta(y) = -1: & 1 \\ \text{否则}: & 0 \end{cases} \end{aligned} \quad (\text{B.17})$$

### 左-右

左-右和右-左对称。使用类似的推导,我们可以得到和式(式(B.17))完全相同的结果。

□

## B.3 删除算法

删除后会引起子树高度的降低。如果平衡因子超出了  $[-1, 1]$  的范围,就需要修复以保持 AVL 树的性质。

### B.3.1 函数式删除

我们先复用二叉搜索树删除算法,然后检查平衡因子并进行修复。删除的结果为一对值  $(T', \Delta H)$ , 其中  $T'$  是删除后的新树、 $\Delta H$  是高度的减少量。删除算法定义如下:

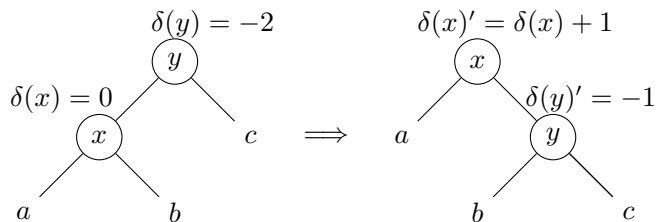
$$\text{delete} = \text{fst} \circ \text{del} \quad (\text{B.18})$$

其中  $\text{del}(T, k)$  从树  $T$  中将元素  $k$  删除:

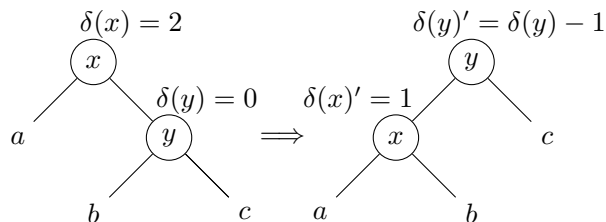
$$\text{del } \emptyset k = (\emptyset, 0)$$

$$\text{del}(l, k', r, \delta) = \begin{cases} k < k': & \text{tree}(\text{del } l k) k' (r, 0) \delta \\ k > k': & \text{tree}(l, 0) k' (\text{del } r k) \delta \\ k = k': & \begin{cases} l = \emptyset: & (r, -1) \\ r = \emptyset: & (l, -1) \\ \text{否则}: & \text{tree}(l, 0) k'' (\text{del } r k'') \delta \\ & \text{其中 } k'' = \min(r) \end{cases} \end{cases} \quad (\text{B.19})$$

如果树为空,结果为  $(\emptyset, 0)$ ; 否则令树为  $T = (l, k', r, \delta)$ 。我们比较  $k$  和  $k'$  的关系,并沿着子树递归查找和删除。当  $k = k'$  时,我们定位到了要删除的节点。如果它的任一子树为空,可以将其切下并用另一棵子树替代。否则,将右子树中的最小值  $k''$



(a) 情况 A



(b) 情况 B

图 B.2: 删除修复

切下,并替换  $k'$ 。我们可以复用  $tree$  函数和  $\Delta H$  的结果。和插入算法相比,需要增加两种删除中特有的情况:

$$\begin{aligned} \dots \\ \text{balance}((a, x, b, \delta(x)), y, c, -2) \Delta H &= (a, x, (b, y, c, -1), \delta(x) + 1, \Delta H) \\ \text{balance}(a, x, (b, y, c, \delta(y)), 2) \Delta H &= ((a, x, b, 1), y, c, \delta(y) - 1, \Delta H) \\ \dots \end{aligned} \quad (\text{B.20})$$

相应的例子程序如下:

```
delete t x = fst $ del t x where
del Empty _ = (Empty, 0)
del (Br l k r d) x
  | x < k = node (del l x) k (r, 0) d
  | x > k = node (l, 0) k (del r x) d
  | isEmpty l = (r, -1)
  | isEmpty r = (l, -1)
  | otherwise = node (l, 0) k' (del r k') d where k' = min r
```

其中  $\text{min}$  和  $\text{isEmpty}$  定义为:

```
min (Br Empty x _ _) = x
min (Br l _ _ _) = min l

isEmpty Empty = True
isEmpty _ = False
```

这样总共有 7 种情况需要在  $\text{balance}$  中实现:

```
balance (Br (Br (Br a x b dx) y c (-1)) z d (-2), dH) =
```

```

        (Br (Br a x b dx) y (Br c z d 0) 0, dH-1)
balance (Br a x (Br b y (Br c z d dz) 1) 2, dH) =
        (Br (Br a x b 0) y (Br c z d dz) 0, dH-1)
balance (Br (Br a x (Br b y c dy) 1) z d (-2), dH) =
        (Br (Br a x b dx') y (Br c z d dz') 0, dH-1) where
    dx' = if dy == 1 then -1 else 0
    dz' = if dy == -1 then 1 else 0
balance (Br a x (Br (Br b y c dy) z d (-1)) 2, dH) =
        (Br (Br a x b dx') y (Br c z d dz') 0, dH-1) where
    dx' = if dy == 1 then -1 else 0
    dz' = if dy == -1 then 1 else 0
— Delete specific
balance (Br (Br a x b dx) y c (-2), dH) =
        (Br a x (Br b y c (-1)) (dx+1), dH)
balance (Br a x (Br b y c dy) 2, dH) =
        (Br (Br a x b 1) y c (dy-1), dH)
balance (t, d) = (t, d)

```

### B.3.2 命令式删除

命令式删除使用树旋作来修复平衡。和插入相比,删除需要处理更多的情况。我们首先复用二叉搜索树删除,然后再修复子树高度变化引起的平衡问题。

```

1: function DELETE( $T, x$ )
2:   if  $x = \text{NIL}$  then
3:     return  $T$ 
4:    $p \leftarrow \text{PARENT}(x)$ 
5:   if LEFT( $x$ ) = NIL then
6:      $y \leftarrow \text{RIGHT}(x)$ 
7:     replace  $x$  with  $y$ 
8:   else if RIGHT( $x$ ) = NIL then
9:      $y \leftarrow \text{LEFT}(x)$ 
10:    replace  $x$  with  $y$ 
11:  else
12:     $z \leftarrow \text{MIN}(\text{RIGHT}(x))$ 
13:    copy key and satellite data from  $z$  to  $x$ 
14:     $p \leftarrow \text{PARENT}(z)$ 
15:     $y \leftarrow \text{RIGHT}(z)$ 
16:    replace  $z$  with  $y$ 
17:  return AVL-DELETE-FIX( $T, p, y$ )

```

删除节点  $x$  时,记  $x$  的父节点为  $p$ 。如果任一子树为空,我们将  $x$  切下,取代为另一子树。否则,如果两棵子树都不为空,我们在右子树中找到最小值节点  $z$ ,将其中的数据复制到  $x$ ,然后将  $z$  切下。最后,我们调用 AVL-DELETE-FIX,并传入根节点  $T$ 、

父节点  $p$ 、和替换节点  $y$ 。记父节点  $p$  的平衡因子为  $\delta(p)$ ，删除后的平衡因子为  $\delta(p)'$ 。它们之间的关系有三种不同的情况：

1.  $|\delta(p)| = 0, |\delta(p)'| = 1$ 。虽然删除后子树的高度降低了，但是父节点仍然满足 AVL 树的性质。修复中止。
2.  $|\delta(p)| = 1, |\delta(p)'| = 0$ 。删除前左右子树的高度差为 1，删除后原来较高的树减小了 1。左右子树现在高度相等。结果是父节点的高度也减小了 1。我们需要继续自底向上更新树的高度。
3.  $|\delta(p)| = 1, |\delta(p)'| = 2$ 。这说明删除后子树的高度差违反了 AVL 树的性质，我们需要通过树旋转来修复平衡。

对于情况 3，大部分修复和插入算法相同。我们需要针对图图 B.2 中所示的两种情况进行额外的处理。

```

1: function AVL-DELETE-FIX( $T, p, x$ )
2:   while  $p \neq \text{NIL}$  do
3:      $l \leftarrow \text{LEFT}(p), r \leftarrow \text{RIGHT}(p)$ 
4:      $\delta \leftarrow \delta(p), \delta' \leftarrow \delta$ 
5:     if  $x = l$  then
6:        $\delta' \leftarrow \delta' + 1$ 
7:     else
8:        $\delta' \leftarrow \delta' - 1$ 
9:     if  $p$  is leaf then ▷  $l = r = \text{NIL}$ 
10:       $\delta' \leftarrow 0$ 
11:    if  $|\delta| = 1 \wedge |\delta'| = 0$  then
12:       $x \leftarrow p$ 
13:       $p \leftarrow \text{PARENT}(x)$ 
14:    else if  $|\delta| = 0 \wedge |\delta'| = 1$  then
15:      return  $T$ 
16:    else if  $|\delta| = 1 \wedge |\delta'| = 2$  then
17:      if  $\delta' = 2$  then
18:        if  $\delta(r) = 1$  then ▷ 右-右
19:           $\delta(p) \leftarrow 0$ 
20:           $\delta(r) \leftarrow 0$ 
21:           $p \leftarrow r$ 
22:           $T \leftarrow \text{LEFT-ROTATE}(T, p)$ 
23:        else if  $\delta(r) = -1$  then ▷ 右-左
24:           $\delta_y \leftarrow \delta(\text{LEFT}(r))$ 
25:          if  $\delta_y = 1$  then

```

```

26:            $\delta(p) \leftarrow -1$ 
27:       else
28:            $\delta(p) \leftarrow 0$ 
29:            $\delta(\text{LEFT}(r)) \leftarrow 0$ 
30:       if  $\delta_y = -1$  then
31:            $\delta(r) \leftarrow 1$ 
32:       else
33:            $\delta(r) \leftarrow 0$ 
34:       else ▷ 删除特有, 右-右
35:            $\delta(p) \leftarrow 1$ 
36:            $\delta(r) \leftarrow \delta(r) - 1$ 
37:            $T \leftarrow \text{LEFT-ROTATE}(T, p)$ 
38:           break ▷ 高度不再变化
39:       else if  $\delta' = -2$  then
40:           if  $\delta(l) = -1$  then ▷ 左-左
41:                $\delta(p) \leftarrow 0$ 
42:                $\delta(l) \leftarrow 0$ 
43:                $p \leftarrow l$ 
44:                $T \leftarrow \text{RIGHT-ROTATE}(T, p)$ 
45:           else if  $\delta(l) = 1$  then ▷ 左-右
46:                $\delta_y \leftarrow \delta(\text{RIGHT}(l))$ 
47:               if  $\delta_y = -1$  then
48:                    $\delta(p) \leftarrow 1$ 
49:               else
50:                    $\delta(p) \leftarrow 0$ 
51:                    $\delta(\text{RIGHT}(l)) \leftarrow 0$ 
52:               if  $\delta_y = 1$  then
53:                    $\delta(l) \leftarrow -1$ 
54:               else
55:                    $\delta(l) \leftarrow 0$ 
56:           else ▷ 删除特有, 左-左
57:                $\delta(p) \leftarrow -1$ 
58:                $\delta(l) \leftarrow \delta(l) + 1$ 
59:                $T \leftarrow \text{RIGHT-ROTATE}(T, p)$ 
60:               break ▷ 高度不再变化
▷ 高度减小, 继续自底向上更新
61:        $x \leftarrow p$ 
62:        $p \leftarrow \text{PARENT}(x)$ 

```

```

63:   if  $p = \text{NIL}$  then                                     ▷ 删除根节点
64:       return  $x$ 
65:   return  $T$ 

```

## B.4 例子程序

下面的例子程序实现了 AVL 树的删除算法：

```

Node del(Node t, Node x) {
    if  $x == \text{null}$  then return t
    Node y
    var parent = x.parent
    if  $x.\text{left} == \text{null}$  {
         $y = x.\text{replaceWith}(x.\text{right})$ 
    } else if  $x.\text{right} == \text{null}$  {
         $y = x.\text{replaceWith}(x.\text{left})$ 
    } else {
         $y = \text{min}(x.\text{right})$ 
         $x.\text{key} = y.\text{key}$ 
        parent = y.parent
         $x = y$ 
         $y = y.\text{replaceWith}(y.\text{right})$ 
    }
    t = deleteFix(t, parent, y)
    release(x)
    return t
}

```

其中 `replaceWith` 的定义参见红黑树的部分。`release(x)` 释放节点  $x$  的空间。修复函数的实现如下：

```

Node deleteFix(Node t, Node parent, Node x) {
    int d1, d2, dy
    Node p, l, r
    while parent  $\neq \text{null}$  {
         $d2 = d1 = \text{parent}.\text{delta}$ 
         $d2 = d2 + \text{if } x == \text{parent}.\text{left} \text{ then } 1 \text{ else } -1$ 
        if isLeaf(parent) then  $d2 = 0$ 
        parent.delta = d2
        p = parent
        l = parent.left
        r = parent.right
        if  $\text{abs}(d1) == 1$  and  $\text{abs}(d2) == 0$  {
             $x = \text{parent}$ 
            parent = x.parent
        } else if  $\text{abs}(d1) == 0$  and  $\text{abs}(d2) == 1$  {
            return t
        } else if  $\text{abs}(d1) == 1$  and  $\text{abs}(d2) == 2$  {
            if  $d2 == 2$  {
                if  $r.\text{delta} == 1$  { // 右-右

```

```

        p.delta = 0
        r.delta = 0
        parent = r
        t = leftRotate(t, p)
    } else if r.delta == -1 { // 右-左
        dy = r.left.delta
        p.delta = if dy == 1 then -1 else 0
        r.left.delta = 0
        r.delta = if dy == -1 then 1 else 0
        parent = r.left
        t = rightRotate(t, r)
        t = leftRotate(t, p)
    } else { // 删除特有, 右-右
        p.delta = 1
        r.delta = r.delta - 1
        t = leftRotate(t, p)
        break // 高度不再继续变化
    }
} else if d2 == -2 {
    if (l.delta == -1) { // 左-左
        p.delta = 0
        l.delta = 0
        parent = l
        t = rightRotate(t, p)
    } else if l.delta == 1 { // 左-右
        dy = l.right.delta
        l.delta = if dy == 1 then -1 else 0
        l.right.delta = 0
        p.delta = if dy == -1 then 1 else 0
        parent = l.right;
        t = leftRotate(t, l)
        t = rightRotate(t, p)
    } else { // 删除特有, 左-左
        p.delta = -1
        l.delta = l.delta + 1
        t = rightRotate(t, p)
        break // 高度不再继续变化
    }
}
// 高度减小, 继续自底向上更新
x = parent
parent = x.parent
}
}
if parent == null then return x // 删除根节点
return t
}

```





## 附录 C 参考答案

### 答案 1

- 1.1. 最小可用数趣题中,所有数都是非负整数。我们可以利用正负号来标记一个数字是否存在。首先扫描一遍列表,若  $|x| < n$ , 其中  $n$  为列表长度,将位置  $|x|$  上的数字置为负数。之后再扫描一遍列表,找到第一个正数所在的位置就是答案。编程实现这一算法。

```
Int minFree([Int] nums) {
    var n = length(nums)
    for Int i = 0 to n - 1 {
        var k = abs(nums[i])
        if k < n then nums[k] = -abs(nums[k])
    }
    for Int i = 0 to n - 1 {
        if nums[i] > 0 then return i
    }
    return n
}
```

- 1.2.  $n$  个数字  $1, 2, \dots, n$ , 经过某一处理后,它们的顺序被打乱了,并且某一个数  $x$  被改成了  $y$ 。假设  $1 \leq y \leq n$ , 设计一个方法能够在线性时间、常数空间内找出  $x$  和  $y$ 。

例如  $X = [3, 1, 3, 5, 4]$  中,丢失的是  $x = 2$ , 重复的是  $y = 3$ 。我们给出 4 种解法:(1)分治、(2)鸽笼排序、(3)符号编码、(4)解方程。

**分治方法。**使用 midpoint  $m = \lfloor \frac{1+n}{2} \rfloor$  划分序列为两部分,左侧:  $as = [a \leq m, a \leftarrow X]$ 、右侧:  $bs = [b > m, b \leftarrow X]$ 。如果左侧长度  $|as| < m$ , 可知丢失的数字在左侧,记  $s = 1 + 2 + \dots + m = \frac{m(m+1)}{2}$ , 则  $x = s - \text{sum}(as)$ 。同时可计算出重复的数字在右侧,记  $s' = (m+1) + (m+2) + \dots + n = \frac{(n+m+1)(n-m)}{2}$ , 则  $y = \text{sum}(bs) - s'$ ; 若左侧的长度  $|as| > m$ , 可知重复的数字在左侧。使用同样的方法,可以算出丢失的数字  $x = s' - \text{sum}(bs)$ , 重复的数字  $y = \text{sum}(as) - s$ ; 否则,如果左侧的长度  $|as| = m$ , 说明有  $m$  个不大于  $m$  的数字,但我们不知道它们是否是  $1, 2, \dots, m$  的某个排列。为此,我们可以计算并比较  $\text{sum}(as)$  和  $s$ 。

如果相等,说明左侧没有问题,可以丢弃左侧的所有数字,然后递归地在右侧寻找  $x$  和  $y$ ; 否则,我们丢弃右侧,在左侧递归寻找答案。在递归查找时,需要用序列的下限  $l$  替换上述步骤中的 1。由于每次去掉一半的列表,根据主定理,总时间复杂度为  $O(n)$ 。

```
missDup xs = solve xs 1 (length xs) where
  solve xs@(l:_:_) l u | k < m - l + 1 = (sl - sl', sr' - sr)
                       | k > m - l + 1 = (sr - sr', sl' - sl)
                       | sl == sl' = solve bs (m + 1) u
                       | otherwise = solve as l m

  where
    m = (l + u) `div` 2
    (as, bs) = partition (<=m) xs
    k = length as
    sl = (l + m) * (m - l + 1) `div` 2
    sr = (m + 1 + u) * (u - m) `div` 2
    (sl', sr') = (sum as, sum bs)
```

**鸽笼排序。**由于所有的数字都在 1 到  $n$  之间,我们可以使用鸽笼排序来重新排列数字。我们自左向右扫描,对于每个位置  $i$  上的数字  $x$ ,如果  $x \neq i$ ,我们就将它和位置  $x$  上的数字  $y$  交换。如果  $x = y$ ,我们就找到了重复的数字,同时,我们得知丢失的数字就是  $i$ ,重复这一交换过程,直到  $x$  等于  $i$  或者找到重复的数字。由于每个数字仅被交换一次以到达正确的位置,因此总时间复杂度为  $O(n)$ 。

```
(Int, Int) missDup([Int] xs) {
  Int miss = -1, dup = -1
  for Int i = 0 to length(xs) - 1 {
    while xs[i] ≠ i {
      Int j = xs[i]
      if xs[j] == xs[i] {
        dup = xs[j]
        miss = i
        break
      } else {
        j = xs[i]
        (xs[i], xs[j]) = (xs[j], xs[i])
      }
    }
  }
  return (miss, dup)
}
```

**符号编码。**假设存在一个长度为  $n$  的标记数组,对于序列中的每个数字  $x$ ,我们都将标记数组中的第  $x$  个位置做上标记。当我们遇到重复元素时,我们会发现这个位置上的标记已经做过了。记重复的数字为  $d$ ,我们知道  $s = 1 + 2 + \dots + n = \frac{n(n+1)}{2}$ ,以及序列中所有的数字和  $s'$ 。我们可以计算出丢失的数字  $m = d + s - s'$ 。但是这一方法需要额外长度为  $n$  的空间用作标记数组。由于数字的存在与否是一种二值化的信息(有、无),我们可以将其编码为数字的正

负号,从而复用待查找的数字序列。对于序列中的每个数字  $x$ ,我们将序列中第  $|x|$  位置上的元素标记为负数,其中  $|x|$  表示绝对值。如果发现某一位置已经为负了,我们就找到了重复的元素,接下来我们就可以计算出丢失的数字。

```
(Int, Int) missDup([Int] xs) {
  Int miss = -1, dup = -1
  Int n = length(xs)
  Int s = sum(xs)
  for i = 0 to n - 1 {
    Int j = abs(xs[i]) - 1
    if xs[j] < 0 {
      dup = j
      miss = dup + n * (n + 1) / 2 - s
      break
    }
    xs[j] = -abs(xs[j])
  }
  return (miss, dup)
}
```

**解方程。**考虑一个简化的问题:给定一个从 1 到  $n$  的列表,去掉一个元素,然后打乱序列的顺序,怎样能够快速找出去掉的元素呢? 我们可以将列表中所有的元素相加,然后从  $\frac{n(n+1)}{2}$  减去这一结果就得出了答案。这一思路可以表示为如下方程:

$$m = s - s'$$

其中  $m$  表示丢失的数字,  $s$  是从 1 到  $n$  的累加和,  $s'$  是列表中所有元素的和。但是同时存在丢失的元素和重复的元素,无法仅用一个方程求出两个未知数。

$$\sum(x[i] - i) = d - m \quad (3)$$

其中左侧是列表中第  $i$  个元素减去  $i$  后的累加和。能否找出第二个独立的方程呢? 思路是使用平方。如果我们将第  $i$  个元素的平方减去  $i$  的平方,然后将结果累加起来。就可以得到下面的结果:

$$\sum(x[i]^2 - i^2) = d^2 - m^2 = (d + m)(d - m) \quad (4)$$

由于  $d - m$  不等于 0,我们可以用式 (3) 除以式 (4) 两侧,得到另一个方程:

$$\sum(x[i]^2 - i^2) / \sum(x[i] - i) = d + m \quad (5)$$

比较式 (3) 和式 (5),两个方程、两个未知数,这样就可以得到结果:

$$\begin{cases} m = \frac{1}{2} \left( \frac{\sum(x[i]^2 - i^2)}{\sum(x[i] - i)} - \sum(x[i] - i) \right) \\ d = \frac{1}{2} \left( \frac{\sum(x[i]^2 - i^2)}{\sum(x[i] - i)} + \sum(x[i] - i) \right) \end{cases}$$

```
missDup xs = ((b `div` a - a) `div` 2, (b `div` a + a) `div` 2) where
  ys = zip xs [1..]
  a = sum [x - y | (x, y) <- ys]
  b = sum [x^2 - y^2 | (x, y) <- ys]
```

1.3. 是的,它是一种等价的队列解法。

### 答案 1.2

1.2.1. 对于元素类型为  $A$  的列表,如果能够判断任何两个元素  $x, y \in A$  是否相等,定义一个算法来判断两个列表是否相等。  
我们用类型限制比较同一类型的列表。

```
(==) :: [a] -> [a] -> Bool
[] == [] = True
(x:xs) == (y:ys) = x == y && (xs == ys)
xs == ys = False
```

### 答案 1.3

1.3.1. 在  $\text{GET-AT}(i, X)$  的迭代实现中, $X$  为空会怎样?  $i$  越界时会怎样?  
此时的行为是未定义的。可以利用  $\text{Optional}\langle T \rangle$  处理此种情况:

```
Optional<T> getAt(List<T> xs, Int i) {
  while i ≠ 0 and xs ≠ null {
    xs = xs.next
    i--
  }
  return if xs ≠ null then Optional.of(xs.key) else Optional.Nothing
}
```

### 答案 1.4

1.4.1. 在  $\text{INIT}$  中,可以用  $\text{APPEND}(X', \text{FIRST}(X))$  来替换  $\text{CONS}$  么?  
 $\text{APPEND}$  每次需要前进到末尾添加元素,这样使得复杂度从  $O(n)$  下降到  $O(n^2)$ ,其中  $n$  是列表长度。

1.4.2. 在  $\text{LAST-AT}$  中,如何处理空列表和越界的情况?  
使用  $\text{Optional}\langle T \rangle$ :

```
Optional<T> lastAt(List<T> xs, Int i) {
  List<T> p = xs
  while i ≠ 0 and xs ≠ null {
    xs = xs.next
    i--
  }
  if xs == null then return Optional.Nothing
```

```

while xs.next ≠ null {
    xs = xs.next
    p = p.next
}
return Optional.of(p.key)
}

```

### 答案 1.5

- 1.5.1. 在列表的定义中增加一个尾部变量 `tail`, 将添加算法优化为常数时间。我们需要两层定义, 内层是基本的单向链表, 外层是带有首尾变量的列表。

```

data List<A> {
    data Node<A> {
        A key
        Node<A> next
    }

    Node<A> head = null
    Node<A> tail = null
    Int length = 0
}

List<A> append(List<A> xs, A x) {
    List.Node<A> tl = xs.tail
    xs.tail = List.Node<A>(x, null)
    if tl == null {
        xs.head = xs.tail
    } else {
        tl.next = xs.tail
    }
    xs.length++
    return xs
}

```

- 1.5.2. 何时应该更新 `tail` 变量? 对性能有何影响?  
 在尾部添加、删除时, 向空列表加入节点时, 从单一元素列表删除元素时, 从中间位置分割列表时。除最后一个操作外, 其它都是常数时间操作。从中间位置分割时  $as = bs \# cs$ , 需要线性时间确定  $bs$  的尾部。
- 1.5.3. 在 `setAt` 中, 如何处理空列表和越界的情况?  
`setAt 0 x []` 的意义等同于 `x : []`。如果  $|xs| = n$ , 可以认为 `setAt n x xs` 等价于 `xs # [x]`。其它空列表和越界情况可做异常处理。

### 答案 1.6

- 1.6.1. 当插入位置越界时, 将其按照添加来处理。

```
insert n x [] = [x]
```

```
insert n x (y:ys) | n ≤ 0 = x : y : ys
                | otherwise y : insert (n - 1) x ys
```

1.6.2. 针对数组实现插入算法,插入位置  $i$  后的所有元素需要向后移动一个位置。

```
[K] insert([K] xs, Int i, K x) {
  append(xs, x)
  for Int j = length(xs) - 1, j > i, j-- {
    swap(xs[j], xs[j-1])
  }
  return xs
}
```

### 答案 1.7

1.7.1. 设计算法将等于给定值的所有元素删除。

```
delAll x [] = []
delAll x (y:ys) | x == y = delAll x ys
                | otherwise = y : delAll x ys
```

或使用叠加操作:

```
delAll x = foldr f [] where
  f y ys = if x == y then ys else y : ys
```

1.7.2. 设计数组的删除算法,被删除位置后的所有元素需要向前移动一个位置。

```
[K] delAt(Int i, [K] xs) {
  Int n = length(xs)
  if 0 ≤ i < n {
    while i < n - 1 {
      xs[i] = xs[i + 1]
      i++
    }
    popLast(xs)
  }
  return xs
}
```

### 答案 1.8

1.8.1. 使用尾递归实现 *length*

```
length = len 0 where
  len n [] = n
  len n (x:xs) = len (n + 1) xs
```

1.8.2. 使用  $n$  的二进制形式,实现幂  $b^n$  的快速计算,使得复杂度为  $O(\lg n)$

我们并不要求  $b$  一定是整数,甚至是实数。 $b$  可以是任何定义了单位  $e$  (例如 1) 和二元运算 (例如乘法) 的集合。数学上我们称这样的集合为么半群 (Monoid)。

```

Monoid pow(Monoid b, Int n) {
  Monoid a = Monoid.e
  while n ≠ 0 {
    if n & 1 == 1 then a = a * b
    b = b * b
    n = n >> 1
  }
  return a
}

```

### 答案 1.9

1.9.1. 使用尾递归求  $[(k, v)]$  列表中  $v$  值最大的元素。

```

maxValue ((k, v):kvs) = maxV v kvs where
  maxV m [] = m
  maxV m ((_, v):kvs) = maxV (max m v) kvs

```

### 答案 1.10

1.10.1. 修改 *take/drop*, 当  $n$  是负数时 *take* 返回  $[], drop$  返回全部列表。

```

safeTake n [] = []
safeTake n (x:xs) | n ≤ 0 = []
                  | otherwise = x : safeTake (n - 1) xs

safeDrop n [] = []
safeDrop n (x:xs) | n ≤ 0 = x:xs
                  | otherwise = safeDrop (n - 1) xs

```

1.10.2. 实现就地修改的 *take* 和 *drop* 算法。

需要在数组尾部加入、删除元素。在头部或中间则需要 *shift* 操作。令数组  $xs$  长度为  $n$ ,  $take(m, xs)$  可以通过连续进行  $n - m$  次尾部删除实现。为了实现  $drop(m, xs)$ , 从  $xs[m]$  起, 把每个元素  $xs[i]$  向前移动到  $xs[i - m]$ 。最后从尾部删除  $m$  个元素。

```

[K] take(Int m, [K] xs) {
  Int d = length(xs) - m
  while d > 0 and xs ≠ [] {
    pop(xs)
    d--
  }
  return xs
}

[K] drop(Int m, xs) {
  if m ≤ 0 then return xs
  for Int i = m to length(xs) { xs[i - m] = xs[i] }
  while m > 0 and xs ≠ [] {

```

```

    pop(xs)
    m--
  }
  return xs
}

```

1.10.3. 将 *sublist* 和 *slice* 改写为柯里化形式, 无需 *X* 作为参数。

```

sublist from cnt = take cnt o drop (from - 1)

slice from to = drop (from - 1) o take to

```

1.10.4. 考虑下面 *span* 的实现:

$$\begin{aligned}
 \text{span } p [] &= ([], []) \\
 \text{span } p (x:xs) &= \begin{cases} p(x) : (x : as, bs), \text{ 其中 } : (as, bs) = \text{span}(p, xs) \\ \text{否则} : (as, x : bs) \end{cases}
 \end{aligned}$$

它和我们本节给出的实现有何不同?

这一实现把列表 *xs* 分成了两部分: 满足 *p* 的所有元素和不满足 *p* 的所有元素:  $as = [p(a), a \leftarrow xs]$  和  $bs = [not\ p(b), b \leftarrow xs]$ 。 *as* 不一定是满足 *p* 的最长前缀, 而是满足 *p* 的最长子序列。

### 答案 1.11

1.11.1. 为了用 *foldr* 定义插入排序, 我们将插入函数设计成 *insert x X*, 这样排序可以表示为:  $sort = foldr\ insert\ []$ 。 *foldr* 的类型为:

$$foldr :: (A \rightarrow B \rightarrow B) \rightarrow B \rightarrow [A] \rightarrow B$$

其中第一个参数 *f* 的类型是:  $A \rightarrow B \rightarrow B$ , 初始值 *z* 的类型为 *B*。它对元素类型为 *A* 的列表进行叠加, 最终结果的类型为 *B*。如何用 *foldl* 定义插入排序? *foldl* 的类型是什么?

*foldl* 的类型为:  $foldl :: (B \rightarrow A \rightarrow B) \rightarrow B \rightarrow [A] \rightarrow B$ 。使用 *foldl*, 我们需要对调参数的位置定义插入排序:

$$sort = foldl (xs\ x \mapsto insert\ x\ xs)\ []$$

或者定义一个“对调”函数:

$$flip\ f\ x\ y = f\ y\ x \tag{1.58}$$

这样插入排序可定义为:  $sort = foldl\ (flip\ insert)\ []$



1.11.2. *concat* 的时间复杂度是什么? 设计一个线性时间的 *concat* 算法。

使用 *foldl* 和 *foldr* 定义 *concat* 的结果大相径庭。 $as \# bs$  的复杂度是  $O(m)$ , 其中  $m = |as|$ , 使用 *foldl* 实现  $xs_1 \# xs_2 \# \dots \# xs_n$  时,  $as$  的长度越来越长, 复杂度为:  $O(m_1 + (m_1 + m_2) + \dots + \sum_1^{n-1} m_i)$ 。使用 *foldr* 实现  $xs_1 \# (xs_2 \# (\dots \# xs_n) \dots)$  时,  $as$  的长度不会越来越长, 而是每个  $xs_i$  的长度, 其复杂度为  $O(m_{n-1} + \dots + m_2 + m_1)$  是线性时间的。我们可以展开定义这一算法:

```
concat [] = []
concat ([]:xss) = concat xss
concat ((x:xs):xss) = x : concat (xs:xss)
```

1.11.3. 使用 *foldr* 来定义 *map*。

$$\text{map } f = \text{foldr } (x \text{ xs} \mapsto (f \ x):\text{xs}) \ [] \quad (1.59)$$

### 答案 1.12

1.12.1. 使用 *reverse* 实现线性时间的过滤算法。

```
List<K> filterL((K → Bool) p, List<K> xs) {
  List<K> ys = null
  while xs ≠ null {
    if p(xs.key) then ys = cons(xs.key, ys)
    xs = xs.next
  }
  return reverse(ys)
}
```

1.12.2. 给定一个列表, 枚举出它的所有后缀。

有多种方法。从列表  $xs$  开始, 不断去掉头部并记录下结果。

$$[xs, \text{tail } xs, \text{tail } (\text{tail } xs), \dots, []]$$

也可以利用 pattern matching 不断去掉头部, 实现 *drop* 的效果:

```
tails [] = [[]]
tails xs@(_:ys) = xs : tails ys
```

我们还可以利用右侧叠加从  $[[[]]]$  开始, 不断从右侧加入元素构造后缀:

```
tails = foldr f [[]] where
  f x xss@(xs:_) = (x:xs) : xss
```

### 答案 1.13

1.13.1. 设计 *iota* 算法(希腊字母  $I$ )其用法如下:

- $iota(\dots, n) = [1, 2, 3, \dots, n]$ ;

- $iota(m, n) = [m, m + 1, m + 2, \dots, n]$ , 其中  $m \leq n$ ;
- $iota(m, m + a, \dots, n) = [m, m + a, m + 2a, \dots, m + ka]$ ;  $k$  是使得  $m + ka \leq n$  的最大整数。
- $iota(m, m, \dots) = repeat(m) = [m, m, m, \dots]$ ;
- $iota(m, \dots) = [m, m + 1, m + 2, \dots]$ .

首先是尾递归的  $[1, 2, \dots, n]$ :

```
iota = iota' [] where
  iota' ns n | n < 1 = ns
            | otherwise = iota' (n : ns) (n - 1)
```

其次规定起始值:  $[m, m + 1, \dots, n]$ , 我们只需要把下限从 1 改为  $m$  即可:

```
iota m n = iota' [] n where
  iota' ns n | n < m = ns
            | otherwise = iota' (n : ns) (n - 1)
```

其次规定步长:  $[m, m + a, m + 2a, \dots, m + ka]$ , 其中  $k$  是使  $m + ka \leq n$  的最大整数。

```
iota m n a | m ≤ n = m : iota (m + a) n a
           | otherwise = []
```

为了产生  $n = \infty$  的开放序列:  $[m, m + 1, \dots]$ , 可以去掉上面的递归终止条件:

```
iota m = m : iota (m + 1)
```

例如 `take 10 (iota 1)` 产生前 10 个自然数

把 +1 去掉就可以得到 `repeat`

```
repeat m = m : repeat m
```

综上, 我们可以定义一个 `iterate` 函数, 包含全部 `iota` 功能:

```
iterate f x = x : iterate f (f x)

iota1 n = take n $ iterate (+1) 1
iota2 m n = takeWhile (<=n) $ iterate (+1) m
iota3 m n a = takeWhile (<=n) $ iterate (+a) m
repeat m = iterate id m
iota5 m = iterate (+1) m
```

### 1.13.2. 实现线性时间的命令式 `zip` 算法。

```
[(A, B)] zip([A] xs, [B] ys) {
  [(A, B)] zs = null
  while xs ≠ null and ys ≠ null {
    zs = cons((xs.key, ys.key), zs)
    xs = xs.next
    ys = ys.next
  }
}
```

```

    return reverse(zs)
}

```

- 1.13.3. 用叠加定义 *zip*(提示:定义两个列表的叠加 *foldr2 f z xs ys*)。我们需要定义针对两个列表的叠加操作:

$$\begin{aligned}
 \text{foldr2 } f \ z \ [] \ ys &= z \\
 \text{foldr2 } f \ z \ xs \ [] &= z \\
 \text{foldr2 } f \ z \ (x:xs) \ (y:ys) &= \text{foldr2 } f \ (f \ x \ y \ z) \ xs \ ys
 \end{aligned}
 \tag{1.75}$$

然后用 *foldr2* 定义 *zip*(柯里化):

$$\text{zip} = \text{foldr2 } f \ [], \text{其中 } : f \ x \ y \ xys = (x, y):xys$$

- 1.13.4. 使用 *zip* 实现 *lastAt*。

为了实现右侧索引第 *k* 个元素,我们将  $xs = [x_0, x_1, \dots, x_{n-1}]$  去掉前 *k* 个元素后得到  $ys = \text{drop } k \ xs$ 。然后把 *xs*, *ys* 关联起来 *zip xs ys*, 这样最后一对元素就是  $(x_{n-k-1}, x_{n-1})$ 。

$$\text{lastAt } k \ xs = (\text{fst} \circ \text{last}) (\text{zip } xs \ (\text{drop } k \ xs))$$

- 1.13.5. 编写一个程序从列表中去重重复的元素。在命令式环境中,请用就地修改的方式删除这些重复元素。在纯函数环境中,构建一个只含有不同元素的新列表。结果列表中的元素顺序应保持和原列表中的一致。这一算法的复杂度是怎样的? 如果允许使用额外的数据结构,可以如何简化实现?

```

dedup [] = []
dedup (x : xs) = x : dedup (filter (x ≠) xs)

```

可以用叠加来定义 *dedup*:

```

dedup = foldr f [] where
  f x xs = x : filter (x ≠) xs

```

由于对每个元素进行 *filter*, 复杂度为  $O(n^2)$ 。使集合(见第三、四章)可以实现  $O(n \lg n)$  的去重操作:

```

dedup = Set.toList ∘ Set.fromList

```

对应的命令式就地修改实现为:

```

[K] dedup([K] xs) {
  Int n = length(xs)
  for Int i = 0, i < n, i++ {
    Int j = i + 1
    while j < n {
      if xs[i] == xs[j] {
        swap(xs[j], xs[n - 1])
        n--
      }
    }
  }
}

```

```

        } else {
            j++
        }
    }
}
Int m = length(xs) - n
loop m { pop(xs) }
return xs
}

```

- 1.13.6. 可以用列表来表示十进制的非负整数。例如 1024 可以表示为：“ $4 \rightarrow 2 \rightarrow 0 \rightarrow 1$ ”。一般来说， $n = d_m \dots d_2 d_1$  可以表示为：“ $d_1 \rightarrow d_2 \rightarrow \dots \rightarrow d_m$ ”。任给两个用列表表示的数  $a$  和  $b$ 。实现它们的算数运算(加减乘除)。

把一个十进制自然数的各位上的数字放入列表，个位在左，高位在右。 $n = (d_m \dots d_2 d_1)_{10}$  表示为  $[d_1, d_2, \dots, d_m]$ 。例如 1024 表示为  $[4, 2, 0, 1]$ 。我们可以从这样的列表，通过  $foldr (c \ d \mapsto 10d + c) \ 0$  转换回自然数。反之，可以利用下面的函数将任意自然数转换成列表：

```

toList n | n < 10 = [n]
         | otherwise = (n `mod` 10) : toList (n `div` 10)

```

首先实现加法。 $0$  是加法的单位元，即  $0 + as = as + 0 = as$ ，其中  $0$  可以表示为  $[]$  或  $[0]$ ，所以：

$$\begin{aligned}
 [] + bs &= bs \\
 [0] + bs &= bs \\
 as + [] &= as \\
 as + [0] &= as
 \end{aligned}$$

将各位相加  $d = (a + b) \bmod 10$ ，进位为  $c = \lfloor \frac{a+b}{10} \rfloor$ ，加到后续位上：

$$(a:as) + (b:bs) = d:(as + bs + [c])$$

下面是对应的例子程序：

```

add [] bs = bs
add [0] bs = bs
add as [] = as
add as [0] = as
add (a:as) (b:bs) = ((a + b) `mod` 10) : add as (add bs [(a + b) `div` 10])

```

其次实现减法。 $as - 0 = as$ 。否则从个位起按位对  $as, bs$  相减。如果某一位上  $a < b$ ，则需要借位： $d = 10 + a - b$ ，而剩余位  $as' = as - [1]$ 。

```

minus as [] = as
minus as [0] = as
minus (a:as) (b:bs) | a < b = (10 + a - b) : minus (minus as [1]) bs
                    | otherwise = (a - b) : minus as bs

```

接下来实现乘法  $as \times bs$ , 我们将  $bs$  中的每一位  $b$  乘以  $as$ , 将结果乘以 10 后累加起来  $cs' = 10 \times cs + (b \times as)$ 。计算  $b \times as$  时, 如果  $b = 0$ , 结果为 0; 其它情况下, 用  $b$  乘以首位  $a$  得到  $d = ab \bmod 10$ , 进位  $c = \lfloor \frac{ab}{10} \rfloor$  加到后继位乘法的结果上:

$$b \times (a:as) = d : ([c] + (b \times as))$$

对应的例子程序为:

```
mul as = foldr (\b cs → add (mul1 b as) (0:cs)) []

mul1 0 _ = []
mul1 b [] = []
mul1 b (a:as) = (b * a `mod` 10) : add [b * a `div` 10] (mul1 b as)
```

最后实现带余数的除法。  $as$  除以  $bs$  的过程相当于寻找商  $q$  和余数  $r$  使得:  $as = q \cdot bs + r$ , 其中余数必须小于除数, 即:  $0 \leq r < bs$ , 余数等于 0 则整除。所以首先需要定义如何判断一个数等于 0, 以及自然数之间的比较关系。如果列表为空或者所有位都是 0 则这个列表代表的数等于 0:

```
isZero = all (== 0)
```

比较  $as$ 、 $bs$  大小时, 0 小于任何非零自然数, 否则从高位向低位进行比较 (EQ 表示相等, LT 表示小于, GT 表示大于):

```
cmp [] [] = EQ
cmp [] bs = if isZero bs then EQ else LT
cmp as [] = if isZero as then EQ else GT
cmp (a:as) (b:bs) = case cmp as bs of EQ → compare a b
                    r → r
```

这样, 小于判断、等于判断都可以用 `cmp` 定义:

```
eq as bs = EQ == cmp as bs
lt as bs = LT == cmp as bs
```

除法可以用穷举的方式实现(设除数  $bs \neq 0$ ):

$$\lfloor as/bs \rfloor = \begin{cases} as < bs : 0 \\ \text{否则} : 1 + \lfloor (as - bs)/bs \rfloor \end{cases}$$

但是我们可以做得更好。模仿课堂上长除法的过程: 考虑  $as = q \cdot bs + r$ , 令  $as = (a_m \dots a_2 a_1)_{10} = [a_1, a_2, \dots, a_m]$ , 我们先用穷举法求  $a_m$  除以  $bs$  的商  $q_m$  和余数  $r_m = a_m - q_m \cdot bs$ 。然后附加上下一位  $a_{m-1}$ , 求  $10r_m + a_{m-1}$  除以  $bs$  的商  $q_{m-1}$  和余数  $r_{m-1}$ 。最后把商的各位连接起来。这恰好是 `foldr` 的过程:

```
ldiv as bs | isZero bs = error "divide by 0"
            | otherwise = foldr f ([], []) as where
  f a (qs, rs) = (q:qs, (a:rs) `minus` (q `mull` bs)) where
    q = ndec (a:rs)
  ndec as = if as `lt` bs then 0 else 1 + ndec (as `minus` bs)
```

- 1.13.7. 在命令式环境中,循环列表是一种有缺陷的列表:某个节点指向了以前的位置,如图 1.6 所示。当遍历的时候,会陷入无限循环。设计一个算法检查某个列表是否含有循环。在此基础上,设计算法找到循环开始的节点(被两个不同祖先指向的节点)。

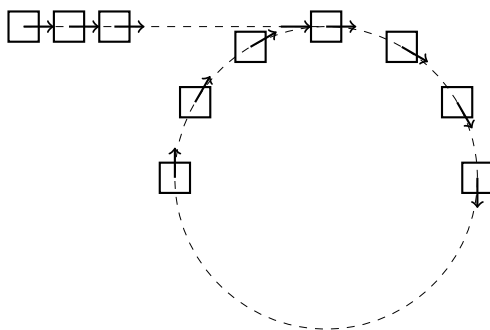


图 1.6: 带有循环的列表

如果列表带有环,从表头开始遍历会陷入循环。如果在环形跑道上有两名选手,跑得快的会不断追上、遇到、超过跑得慢的选手<sup>11</sup>。我们可以利用两个指针遍历列表: $p$  每次前进 1 个节点, $q$  每次前进 2 个节点。如果它们在某个时刻相遇,则说明有环。

```

Bool hasCycle(List<K> h) {
    var p = h, q = h
    while p ≠ null and q ≠ null {
        p = p.next
        q = q.next
        if q == null then return False
        q = q.next
        if p == q return True
    }
    return False
}

```

罗伯特·弗洛伊德把这种方法叫做“龟兔赛跑”。我们还可以进一步求出循环开始的位置。考虑下图 1.7,令  $OA$  长度为  $k$  个节点,从  $A$  开始循环,周长为  $n$  个节点。当  $p$  到达  $A$  时,速度为 2 倍的  $q$  已经到达位置  $B$ 。此时两个指针开始转圈。 $p$  落后  $q$  距离  $k$ ,从环上看等价于  $q$  要追上在它前面  $n - k$  距离的  $p$ 。这需要用时:

$$t = \frac{n - k}{2v - v} = \frac{n - k}{v}$$

<sup>11</sup>严格来说,这只适用于连续情况,对于离散列表,当两个速度互素时,一定会套圈相遇。

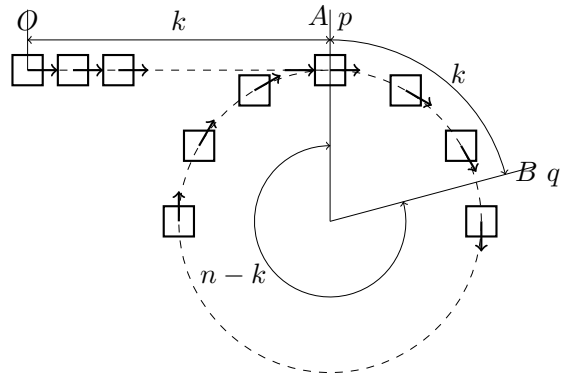


图 1.7: 龟兔赛跑方法

此时  $p$  从  $A$  前进了:  $v \frac{n-k}{v} = n-k$ 。也就是说  $p$  再前进  $k$  就将回到  $A$ 。如果这时让  $q$  回到表头再次出发, 一次前进 1 个节点。  $q$  前进  $k$  后也将到达  $A$ , 即  $p$  和  $q$  在  $A$  相遇。上述分析假设了  $k < n$ 。我们可以证明  $k \geq n$  时结论也是对的:

证明.  $p, q$  从表头出发后, 当  $p$  到达  $A$  时,  $q$  到达了圆周上  $A$  后面  $k \bmod n$  的位置。此时转换为追及问题:  $q$  要追上前面距离  $n - (k \bmod n)$  的  $p$ , 用时:

$$t = \frac{n - (k \bmod n)}{2v - v} = \frac{n - (k \bmod n)}{v}$$

此时  $p$  从  $A$  前进了:  $v \frac{n - (k \bmod n)}{v} = n - (k \bmod n)$ , 如果  $p$  再前进  $k$  将到达位置:  $(k \bmod n) + n - (k \bmod n) = n$ , 恰好回到了  $A$ 。因此如果让  $q$  在追上  $p$  后回到表头再次出发, 一次前进一个节点, 经过  $k$  后也将到达  $A$  和  $p$  再次相遇。  $\square$

根据上述分析, 下面的例子程序求出  $A$  所在的位置:

```
Optional<List<K>> findCycle(List<K> h) {
    var p = h, q = h
    while p ≠ null and q ≠ null {
        p = p.next
        q = q.next
        if q == null then return Optional.Nothing
        q = q.next
        if p == q {
            q = h
            while p ≠ q {
                p = p.next
                q = q.next
            }
            return Optional.of(p)
        }
    }
    return Optional.Nothing
}
```

## 答案 2.1

2.1.1. 给定如下前序遍历和中序遍历的结果, 请重建出二叉树, 并给出后序遍历的结果。

- 前序遍历结果: 1, 2, 4, 3, 5, 6
- 中序遍历结果: 4, 2, 1, 5, 3, 6
- 后序遍历结果: ?

[4, 2, 5, 6, 3, 1]

2.1.2. 编程实现从前序遍历和中序遍历的结果重建二叉树。

记前序遍历结果为  $P$ , 中序遍历结果为  $I$ 。如果  $P = I = []$ , 则二叉树为空  $\emptyset$ 。否则, 前序遍历是递归的“根、左、右”, 因此  $P$  中第一个元素  $m$  是根节点的值。中序遍历是递归的“左、根、右”, 我们一定可以在  $I$  中找到  $m$ , 并把  $I$  分成三部分:  $[a_1, a_2, \dots, a_{i-1}, m, a_{i+1}, a_{i+2}, \dots, a_n]$ , 令:  $I_l = I[1, i), I_r = I[i + 1, n]$ , 其中  $[l, r)$  表示左闭右开区间, 包括  $l$  但不包括  $r$ , 它们可以为空  $[]$ 。这三部分  $I_l, m, I_r$  中,  $I_l$  是左子树的中序遍历结果,  $I_r$  是右子树的中序遍历结果。令  $k = |I_l|$  表示左子树的大小, 我们可以用  $k$  分割  $P[2, n]$  为两部分:  $P_l, P_r$ , 其中  $P_l$  是前  $k$  个元素,  $P_r$  是剩余部分。这样我们就可以递归地用  $(P_l, I_l)$  重建左子树, 用  $(P_r, I_r)$  重建右子树:

$$\begin{aligned} \text{rebuild } [] [] &= \emptyset \\ \text{rebuild } (m:ps) I &= (\text{rebuild } P_l I_l, m, \text{rebuild } P_r I_r) \end{aligned}$$

其中:

$$\begin{cases} (I_l, I_r) &= \text{splitWith } m \ I \\ (P_l, P_r) &= \text{splitAt } |I_l| \ ps \end{cases}$$

下面是例子程序:

```
rebuild [] _ = Empty
rebuild [c] _ = Node Empty c Empty
rebuild (x:xs) ins = Node (rebuild prl inl) x (rebuild prr inr) where
  (inl, _:inr) = (takeWhile (\x -> x /= m) ins, dropWhile (\x -> x /= m) ins)
  (prl, prr) = splitAt (length inl) xs
```

也可以通过更新左右边界来实现:

```
Node<T> rebuild([T] pre, [T] ins, Int l = 0, Int r = length(ins)) {
  if l >= r then return null
  T c = popFront(pre)
  Int m = find(c, ins)
  var left = rebuild(pre, ins, l, m)
  var right = rebuild(pre, ins, m + 1, r)
  return Node(left, c, right)
}
```



2.1.3. 证明对二叉搜索树进行中序遍历可以将全部元素按照从小到大的顺序输出。

**证明.** 使用反证法。假设存在二叉搜索树, 其中序遍历结果无序。我们在所有中序遍历无序的树中, 选出一棵最小的  $T$ 。首先  $T$  不可能为空  $\emptyset$ , 空树的中序遍历结果为  $[\ ]$  是有序的。其次  $T$  不可能是叶子  $T = (\emptyset, k, \emptyset)$ , 叶子的中序遍历结果为  $[k]$  是有序的。因此  $T$  只能是分枝节点  $(l, k, r)$ , 中序遍历结果为  $toList\ l \# [x] \# toList\ r$ 。由于  $T$  是最小的中序遍历无序的树, 而  $l, r$  都比  $T$  小, 所以  $toList\ l$  和  $toList\ r$  都有序。根据二叉搜索树的定义, 任何  $x \in l, x < k$ , 任何  $y \in r, y > k$ 。所以中序遍历结果  $toList\ l \# [x] \# toList\ r$  有序。这与假设  $T$  的中序遍历结果无序矛盾。所以任何二叉搜索树的中序遍历结果有序。  $\square$

2.1.4. 树排序的复杂度是什么?

$(n \lg n)$ , 其中  $n$  是元素数目。

2.1.5. 使用叠加定义  $toList$

$$\begin{aligned} toList &= foldt\ id\ (as\ b\ bs \mapsto as \# b : bs)\ [\ ] \\ &= fold\ (:)\ [\ ] \end{aligned}$$

2.1.6. 使用叠加定义  $depth\ t$ , 获得一棵二叉树的高度。

$$depth = foldt\ (x \mapsto 1)\ (x\ d\ y \mapsto d + \max\ x\ y)\ 0$$

## 答案 2.2

2.2.1. 判断某个值  $k$  是否存在于二叉搜索树  $t$  中。

```
member x (Node l k r) | x == k = True
                  | x < k = member x l
                  | otherwise = member x r
```

2.2.2. 使用 PRED 和 SUCC 实现一个二叉搜索树的迭代器。用它遍历一棵含有  $n$  个元素的树的复杂度是什么?

```
data TreeIterator<T> {
  Node<T> node = null

  TreeIterator(Node<T> root) { node = min(root) }

  T get() = node.key

  Bool hasNext() = node != null

  Self next() { if hasNext() then node = succ(node) }
}
```

尽管在访问前驱和后继元素时需要寻找子树的最大(小)值,或沿着父节点回溯,但使用迭代器遍历搜索树容器却只需要线性时间。在遍历时,对于每个节点,我们仅访问一次(到达一次、离开一次)。例如:

```
for var it = TreeIterator(root), it.hasNext(), it = it.next() {
    print(it.get())
}
```

所以遍历的复杂度为  $O(n)$ 。

2.2.3. 下面程序可以遍历一个区间  $[a, b]$  内的元素:

```
for_each (m.lower_bound(12), m.upper_bound(26), f)
```

试用纯函数式的方法解决这一问题

```
mapR f a b t = map' t where
    map' Empty = Empty
    map' (Node l k r) | k < a = map' r
                    | a ≤ k && k ≤ b = Node (map' l) (f k) (map' r)
                    | k > b = map' l
```

### 答案 2.3

2.3.1. 当节点的两个分支都不为空时,存在一种对称的删除算法:用左子树的最大值替换待删除的节点,然后将此最大值的节点“切下”。编程实现这一算法。

```
delete _ Empty = Empty
delete x (Node l k r) | x < k = Node (delete x l) k r
                    | x > k = Node l k (delete x r)
                    | otherwise = del l r

where
    del Empty r = r
    del l Empty = l
    del l r = let m = max l in Node (delete m l) m r
```

2.3.2. 编程实现随机构建。

```
fromList = (foldr insert Empty) o shuffle
```

2.3.3. 如何在一棵二叉树中找到“距离最远”的两个节点?

我们先求最长距离  $m$ , 然后给出最长路径  $[s, a, b, \dots, e]$ 。路径的两个端点  $s, e$  就是所求的两个节点。首先定义节点间的距离。对于两个节点间的连通路径(无方向)  $s \rightarrow n_1 \rightarrow n_2 \rightarrow \dots \rightarrow n_m \rightarrow e$ , 令每个节点间长度为 1, 则从  $s$  到  $e$  间的总长度为节点间的距离。上述路径的距离为  $m + 1$ 。空树的最长距离为 0, 并且单个叶子节点  $(\emptyset, k, \emptyset)$  的最长路径为  $[k]$ , 最长距离也是 0 (从  $k$  到  $k$ )。考虑分支节点  $(l, k, r)$  的最长距离, 它必然是三者中的最大值: (1) 从左子树最深节点到达根节点, 再从根节点到达右子树最深节点的路径长度。等于  $depth\ l + depth\ r$ ; (2) 左子树  $l$  中的最长距离; (3) 右子树  $r$  中的最长距离。

```

maxDistance Empty = 0
maxDistance (Node Empty _ Empty) = 0
maxDistance (Node l _ r) = maximum [depth l + depth r,
                                     maxDistance l, maxDistance r]

```

其中 *depth* 的定义见练习 2.1.6。稍作改动就可以求出最长路径。空树的最长路径是 [], 单叶子节点的最长路径是 [k], 分支节点 (l, k, r) 的最长路径是三者间的最长: (1) 从根到左侧最深节点的路径的逆序, 加上 k, 再加上根到右侧最深节点的路径; (2) 左侧最长路径; (3) 右侧最长路径。

```

maxPath Empty = []
maxPath (Node Empty k Empty) = [k]
maxPath (Node l k r) = longest [(reverse depthPath l) # k : depthPath r,
                                maxPath l, maxPath r] where
  longest = maximumBy (compare `on` length)
  depthPath = foldt id (\xs k ys → k : longest [xs, ys]) []

```

这一方法在计算深度时遍历一次, 计算左右分枝时又各遍历了一次。为了避免重复, 我们可以自底向上在节点中填写深度 *d* 和最大距离 *m*。通过树映射:  $Tree\ A \mapsto Tree\ (Int, Int)$  遍历一次获得答案:

```

maxDist = extract ◦ mapTr where
  extract Empty = 0
  extract (Node _ (_, m) _) = m
  mapTr Empty = Empty
  mapTr (Node l _ r) = f (mapTr l) (mapTr r)
  f l r = Node l (1 + max d1 d2, maximum [d1 + d2, m1, m2]) r where
    (d1, m1) = pairOf l
    (d2, m2) = pairOf r
    pairOf Empty = (0, 0)
    pairOf (Node _ k _) = k

```

我们也可以利用叠加来简化:

```

maxDist = snd ◦ pair ◦ foldt id g Empty where
  g l _ r = Node l (1 + max d1 d2, maximum [d1 + d2, m1, m2]) r where
    (d1, m1) = pair l
    (d2, m2) = pair r
  pair = (maybe (0, 0) id) ◦ key

```

## 答案 3.1

### 3.1.1. 实现从左向右处理的插入操作。

```

Void insert([T] xs, T x) {
  Int i = 0, n = length(xs)
  append(xs, x)
  while i < n and xs[i] < x { i++ }
  while i < n {
    xs[n] = xs[n - 1]

```

```

        n = n - 1
    }
    xs[i] = x
}

```

3.1.2. 定义单独的插入函数以实现插入排序。

```

Void insert([T] xs, T x) {
    append(xs, x)
    Int i = length(xs) - 1
    while i > 0 and xs[i] < xs[i-1] {
        swap(xs[i], xs[i-1])
        i--
    }
}

[T] sort([T] xs) {
    [T] ys = []
    for x in xs {
        insert(ys, x)
    }
    return ys
}

```

### 答案 3.2

3.2.1. 使用索引数组, 排序结果是一个重新排列的索引。给出一个方法, 根据新的索引 *Next*, 重新排列数组 *A*。

```

[K] reorder([K] xs, [Int] next) {
    Int i = -1
    [Int] ys = []
    while next[i] ≠ -1 {
        append(ys, xs[next[i]])
        i = next[i]
    }
    return ys
}

```

### 答案 4.1

4.1.1. 对于较大的通讯录, 为了加快构建速度, 可以使用两个并发的任务: 一个从头向后, 另外一个从后向前读取。当两个任务相遇时结束。这样构建出的二叉搜索树是什么样子的? 如果把通讯录分成更多片断, 使用多任务会得到什么结果? 参考图 4.2, 找出更多的不平衡情况。

由于通讯录按照字典顺序排列, 头尾两个任务分别产生两棵不平衡树。头:  $T_h = ((\dots(\emptyset, k_1, \emptyset), \dots), k_m, \emptyset)$ , 尾:  $T_t = (\emptyset, k_{m+1}, (\emptyset, k_{m+2}, \dots(\emptyset, k_n, \emptyset))\dots)$ 。

形如图 4.2(c)。如果分成多个片段,每个任务产生一个如  $T_h$  形状的子树。这些子树组合在一起成为一棵非平衡树。图 4.2(b) 在构建时,元素大小交错,每个节点都有一个空分枝。

### 答案 4.2

4.2.1. 实现右旋 RIGHT-ROTATE 操作。

```

1: function RIGHT-ROTATE( $T, y$ )
2:    $p \leftarrow \text{PARENT}(y)$ 
3:    $x \leftarrow \text{LEFT}(y)$                                 ▷ 设  $x \neq \text{NIL}$ 
4:    $a \leftarrow \text{LEFT}(x)$ 
5:    $b \leftarrow \text{RIGHT}(x)$ 
6:    $c \leftarrow \text{RIGHT}(y)$ 
7:   REPLACE( $y, x$ )                                       ▷ 用  $x$  替换  $y$ 
8:   SET-SUBTREES( $y, b, c$ )                                ▷ 令  $b, c$  为  $y$  的子树
9:   SET-SUBTREES( $x, a, y$ )                                ▷ 令  $a, y$  为  $x$  的子树
10:  if  $p = \text{NIL}$  then                                    ▷ 此前  $y$  是根节点
11:     $T \leftarrow x$ 
12:  return  $T$ 

```

### 答案 4.3

4.3.1. 证明含有  $n$  个节点的红黑树,其高度  $h$  不会超过  $2\lg(n+1)$ 。

我们首先引入“黑色高度”的定义。节点  $x$  的黑色高度  $bh(x)$  是从  $x$  (不含  $x$ ) 到某一叶子节点路径上的黑色节点数目。根据红黑树的性质 5, 所有路径包含相同数量的黑色节点。所以这一定义是有效的。特别地,称根结点的黑色高度为红黑树的黑色高度。

**证明.** 我们首先证明以任何节点  $x$  为根的子树至少包含  $2^{bh(x)} - 1$  个节点。我们对  $x$  的高度使用数学归纳法:若高度为 0, 则  $x = \text{NIL}$ , 子树至少含有  $2^0 - 1 = 0$  个节点。现在考虑分枝节点  $x$ , 它的左右子树的黑色高度是  $bh(x)$  (子树的根为黑色) 或  $bh(x) - 1$  (子树的根为红色)。由于  $x$  的子树高度一定小于  $x$  的高度, 根据归纳假设, 每棵子树至少含有  $2^{bh(x)-1} - 1$  个节点。所以子树  $x$  至少含有:  $2(2^{bh(x)-1} - 1) + 1 = 2^{bh(x)} - 1$  个节点。

令红黑树的高度为  $h$ , 根据红黑树性质 4, 不存在连续的红色节点。所以任一路径上, 黑色节点的数目超过一半。因此红黑树的黑色高度至少为  $h/2$ 。因此:

$$n \geq 2^{h/2} - 1 \Rightarrow 2^{h/2} \leq n + 1$$

两边取对数得到:  $h/2 \leq \lg(n+1)$ , 即:  $h \leq 2\lg(n+1)$ 。 □

## 答案 4.4

4.4.1. 不使用模式匹配,分别检查四种情况实现 *insert* 算法。

```

Node<T> insert(Node<T> t, T x) = makeBlack(ins(t, x))

Node<T> makeBlack(Node<T> t) {
    t.color = Color.BLACK
    return t
}

Node<T> ins(Node<T> t, T x) {
    if t == null then return Node(null, x, null, Color.BLACK)
    return if x < t.key
        then balance(t.color, ins(t.left, x), t.key, t.right)
        else balance(t.color, t.left, t.key, ins(t.right, x))
}

Node<T> balance(Color c, Node<T> l, T k, Node<T> r) {
    return if c == Color.BLACK {
        if isRed(l) and isRed(l.left) {
            Node(Node(l.left.left, l.left.key, l.left.right, Color.BLACK),
                l.key,
                Node(l.right, k, r, Color.BLACK),
                Color.RED)
        } else if isRed(l) and isRed(l.right) {
            Node(Node(l.left, l.key, l.right.left, Color.BLACK),
                l.right.key,
                Node(l.right.right, k, r, Color.BLACK),
                Color.RED)
        } else if isRed(r) and isRed(r.right) {
            Node(Node(l, k, r.left, Color.BLACK),
                r.key,
                Node(r.left.right, r.right.key, r.right.right, Color.BLACK),
                Color.RED)
        } else if isRed(r) and isRed(r.left) {
            Node(Node(l, k, r.left.left, Color.BLACK),
                r.left.key,
                Node(r.left.right, r.key, r.right, Color.BLACK),
                Color.RED)
        } else {
            Node(l, k, r, c)
        }
    } else {
        Node(l, k, r, c)
    }
}

Bool isRed(Node<T> t) = (t != null and t.color == Color.RED)

```

## 答案 4.5

4.5.1. 实现“标记——重建”删除算法：标记被删除的节点，但不进行真正的移除。当被标记的节点数目超过 50% 时重建树。

我们把每个值  $x$  和一个标记  $a$  存入节点中，树的类型为： $Tree(K, Bool)$ ，插入  $x$  时，使用  $insert(x, True)$  将节点标记为存在。删除时将标记  $a$  变为  $False$ 。然后统计已删除节点个数，并启动重建。

$$delete\ x = rebuild \circ del\ x$$

其中：

$$del\ x\ \emptyset = \emptyset$$

$$del\ x\ (c, l, (k, a), r) = \begin{cases} x < k : & (c, del\ x\ l, (k, a), r) \\ x > k : & (c, l, (k, a), del\ x\ r) \\ x = k : & (c, l, (k, False), r) \end{cases}$$

如果树中被标记为删除的节点超过一半，我们将树转换成列表，筛选出剩余节点，然后重建。

$$rebuild\ t = \begin{cases} size\ t < \frac{1}{2}(cap\ t) : & (fromList \circ toList)\ t \\ 否则 : & t \end{cases}$$

其中  $toList$  在中序遍历时，忽略标记为删除的节点：

$$toList\ \emptyset = []$$

$$toList\ (c, l, (k, a), r) = \begin{cases} a : & toList\ l \# [k] \# toList\ r \\ 否则 : & toList\ l \# toList\ r \end{cases}$$

为了避免每次遍历整棵树统计节点，需要在节点中缓存树的大小 ( $cap$ ) 和未删除节点个数 ( $size$ )。将树的类型扩充为： $Tree(K, Bool, Int, Int)$ 。并定义  $node$  函数填充节点数目：

$$node\ \emptyset = \emptyset$$

$$node\ c\ l\ (k, a, \_, \_) r = (c, l, (k, a, sz, ca), r)$$

其中：

$$\begin{cases} sz = size\ l + size\ r + (\text{if } a \text{ then } 1 \text{ else } 0) \\ ca = 1 + cap\ l + cap\ r \end{cases}$$

函数  $size$  和  $cap$  从节点中取出缓存的内容：

$$size\ \emptyset = 0 \qquad cap\ \emptyset = 0$$

$$size\ (\_, (\_, \_, sz, \_), \_) = sz \qquad cap\ (\_, (\_, \_, \_, ca), \_) = ca$$

最后，我们用  $node$  函数替换红黑树插入和删除中的  $(c, l, k, r)$ ，如下面的例子程序：

```

data Elem a = Elem a Bool Int Int deriving (Eq)

active (Elem _ a _ _) = a
getElem (Elem x _ _ _) = x

instance Ord a => Ord (Elem a) where
  compare = compare `on` getElem

insert x = makeBlack ◦ ins (Elem x True 1 1) where
  ins e Empty = Node R Empty e Empty
  ins e (Node color l k r)
    | e < k    = balance color (ins e l) k r
    | otherwise = balance color l k (ins e r)
  makeBlack (Node _ l k r) = Node B l k r

balance B (Node R (Node R a x b) y c) z d = node R (node B a x b) y (node B c z d)
balance B (Node R a x (Node R b y c)) z d = node R (node B a x b) y (node B c z d)
balance B a x (Node R b y (Node R c z d)) = node R (node B a x b) y (node B c z d)
balance B a x (Node R (Node R b y c) z d) = node R (node B a x b) y (node B c z d)
balance color l k r = node color l k r

node c l (Elem k a _ _) r = Node c l (Elem k a sz ca) r where
  sz = size l + size r + if a then 1 else 0
  ca = cap l + cap r + 1

size Empty = 0
size (Node _ _ (Elem _ _ sz _) _) = sz

cap Empty = 0
cap (Node _ _ (Elem _ _ _ ca) _) = ca

delete x = rebuild ◦ del x where
  del _ Empty = Empty
  del x (Node c l e@(Elem k a sz ca) r)
    | x < k = node c (del x l) e r
    | x > k = node c l e (del x r)
    | x == k = node c l (Elem k False 0 0) r

rebuild t | 2 * size t < cap t = (fromList ◦ toList) t
          | otherwise = t

fromList :: (Ord a) => [a] → RBTREE (Elem a)
fromList = foldr insert Empty

toList Empty = []
toList (Node _ l e r) | active e = toList l # [getElem e] # toList r
                      | otherwise = toList l # toList r

```

### 答案 5.1

5.1.1. 我们只验证了 AVL 树的高度性质, 完成验证程序检查一棵二叉树是否是 AVL



树。

除了验证高度性质,我们可以利用二叉搜索树的 *toList* 函数,检查中序遍历结果是否有序:

```
verifyAVL t = isAVL t && sorted (toList t) where
  sorted [] = True
  sorted xs = and (zipWith (<=) xs (tail xs))
```

### 答案 6.1

6.1.1. 是否可以将定义 `Branch (IntTrie a) (Maybe a) (IntTrie a)` 变为 `Branch (IntTrie a) a (IntTrie a)`, 如果值不存在返回 *Nothing*, 否则返回 *Just v*?

不仅 *lookup*, 我们需要在 *insert* 中也正确处理没有值的中间节点, 如图 6.3 中的空白圆圈节点。如果不使用 `Maybe` 数据类型, 我们还需要额外增加一个代数数据类型的定义:

```
data IntTrie a = Empty
              | Branch (IntTrie a) a (IntTrie a)
              | EmptyBranch (IntTrie a) (IntTrie a)
```

### 答案 6.2

6.2.1. 编写程序实现整数前缀树的 *lookup* 算法。

```
import Data.Bits

type Key = Int
type Prefix = Int
type Mask = Int

data IntTree a = Empty
              | Leaf Key a
              | Branch Prefix Mask (IntTree a) (IntTree a)

lookup :: Key → IntTree a → Maybe a
lookup _ Empty = Nothing
lookup k (Leaf k' v) = if k == k' then Just v else Nothing
lookup k (Branch p m l r) | match k p m = if zero k m then lookup k l
                          | otherwise = lookup k r

match :: Key → Prefix → Mask → Bool
match k p m = (mask k m) == p

mask :: Int → Mask → Int
mask x m = (x .&. complement (m - 1))
```

```
zero :: Int → Mask → Bool
zero x m = x .&. (shiftR m 1) == 0
```

- 6.2.2. 实现整数 trie 和整数树的前序遍历,仅输出值不为空的节点键。结果有何规律?我们先考虑如何通过前序遍历将整数 trie 转换成关联列表。前序遍历的顺序是递归的“中-左-右”。空树的遍历结果是 []。对分枝节点  $(l, m, r)$ ,令左右子树的递归前序遍历结果为  $as, bs$ ,中间的  $m$  类型是 Maybe,如果是 Nothing,则递归结果是  $as + bs$ ;如果是 Just  $v$ ,则结果是  $(k, v):as + bs$ ,其中  $k$  是节点对应的二进制整数编码(0 向左、1 向右的小端整数)。

```
toList = go 0 1 where
  go _ _ Empty = []
  go k n (Branch l m r) = case m of
    Nothing → as + bs
    (Just v) → (k, v) : as + bs
  where
    as = go k      (2 * n) l
    bs = go (n + k) (2 * n) r
```

我们从根结点开始,令  $k = 0$ ,深度  $d = 0$ 。向左则  $k' = 0$ ,向右则  $k' = 1 = 2^d + k = 1 + 0$ ;进入下一层  $d = 1$ 。4 个节点对应的  $k$  为  $(00)_2 = 0, (10)_2 = 2^1 + 0, (01)_2 = 1, (11)_2 = 2^1 + 1$ 。一般地,对于第  $d$  层节点,令  $k = (a_d \dots a_2 a_1 a_0)_2$ ,向左则  $k' = k$ ,向右则  $k' = a_d 2^d + k$ 。因此在上述实现中,我们用  $k = 0, n = 2^0 = 1$  启动函数  $go\ k\ n$ 。向左递归调用  $go\ k\ 2n$ 、向右递归调用  $go\ (n + k)\ 2n$ 。这样通过  $keys = fst \circ unzip \circ toList$  就可以获得前序遍历的键列表。上述实现中  $as + bs$  需要线性时间。我们可以利用尾递归优化:

```
toList = go 0 1 [] where
  go _ _ z Empty = z
  go k n z (Branch l m r) = case m of
    Nothing → xs
    (Just v) → (k, v) : xs
  where xs = go k (2 * n) (go (n + k) (2 * n) z r) l
```

进一步,我们可以抽象出整数 trie 的前序叠加操作。和第二章中的  $fold$  相比,树中的 key 是计算出来的。

```
foldpre f z = go 0 1 z where
  go _ _ z Empty = z
  go k n z (Branch l m r) = f k m (go k (2 * n) (go (n + k) (2 * n) z r) l)
```

使用叠加操作,我们可以重新定义  $toList$  为:

```
toList = foldpre f [] where
  f _ Nothing xs = xs
  f k (Just v) xs = (k, v) : xs
```

整数前缀树的前序遍历更加简单,无需计算 key。我们可以定义抽象的前序叠加:

```
foldpre _ z Empty = z
foldpre f z (Leaf k v) = f k v z
foldpre f z (Branch p m l r) = foldpre f (foldpre f z r) l
```

然后实现从树转换为键-值列表:

```
toList = foldpre (\k v xs → (k, v):xs) []
keys = fst ◦ unzip ◦ toList
```

如果前序遍历输出一棵树中的所有键,无论是小端整数 trie 还是大端整数前缀树,它们都是按照表示键的二进制位顺序升序排列的。为了验证,我们定义一个转换函数 *bitsLE*,将小端整数转换为对应的二进制位列表。然后可以验证前序遍历的顺序如下:

```
verify kvs = sorted $ map bitsLE $ keys $ fromList kvs where
  sorted [] = True
  sorted xs = and $ zipWith (<=) xs (tail xs)
  bitsLE 0 = []
  bitsLE n = (n `mod` 2) : bitsLE (n `div` 2)
```

其中 *kvs* 是随机输入的键-值列表。对应大端整数前缀树的验证程序如下:

```
verify kvs = sorted $ keys $ fromList kvs
```

## 答案 6.3

6.3.1. 消除 *lookup* 算法中的递归,用循环实现前缀树的查找。

```
Optional<V> lookup(PrefixTree<K, V> t, K key) {
  if t == null then return Optional.Nothing
  Bool match
  repeat {
    match = False
    for k, tr in t.subtrees {
      if k == key then return Optional.of(tr.value)
      (K prefix, K k1, K k2) = lcp(key, k)
      if prefix ≠ [] and k2 == [] {
        match = True
        key = k1
        t = tr
        break
      }
    }
  } until not match
  return Optional.Nothing
}
```

## 答案 6.4

## 6.4.1. 使用 trie 实现自动补齐和预测式输入。

对于输入的前缀  $ks$ , 我们沿着 trie 前进至子节点  $t$ , 然后将  $t$  下面的子树全部展开。最后截断前  $n$  个候选项:

```
import Data.Map (Map)
import qualified Data.Map as Map

startsWith :: Ord k => [k] -> Trie k v -> [[(k), v]]
startsWith [] (Trie Nothing ts) = enum ts
startsWith [] (Trie (Just v) ts) = ([], v) : enum ts
startsWith (k:ks) (Trie _ ts) = case Map.lookup k ts of
  Nothing -> []
  Just t -> map (first (k:)) (startsWith ks t)

enum :: Ord k => Map k (Trie k v) -> [[(k), v]]
enum = (concatMap (\(k, t) ->
  map (first (k:)) (startsWith [] t))) o Map.assocs

get n k t = take n $ startsWith k t
```

其中  $first f (a, b) = (f a, b)$ , 将函数  $f$  应用到一对值的前一个上。实现预测式输入时, 针对每个数字, 我们都在  $M_{T9}$  中找到所有可能的字符, 然后到 trie 中递归查找对应的子树:

```
findT9 [] _ = [[]]
findT9 (d:ds) (Trie _ ts) = concatMap find cts where
  cts = case Map.lookup d mapT9 of
    Nothing -> []
    Just cs -> Map.assocs $ Map.filterWithKey (\c _ -> c `elem` cs) ts
  find (c, t) = map (c:) (findT9 ds t)
```

## 6.4.2. 对于返回多个候选结果的前缀树查找算法, 如何保证输出的结果按照字典顺序排序? 这会对性能产生怎样的影响?

由练习 6.2.2, 对于二叉前缀树, 按照前序遍历可以获得字典顺序的结果。对于多叉前缀树, 我们需要对子树按字典顺序遍历。如果用自平衡二叉搜索树管理子树(如红黑树或 AVL 树), 可以在线性时间内按序遍历子树(见练习 2.2.2)。如果用散列表或者关联列表管理子树, 则需要  $O(n \lg n)$  时间对子树排序。

## 答案 7.1

7.1.1. 我们是否可以用  $\leq$  使得 B 树含有重复元素?

可以用  $\leq$  使 B 树包含重复元素: 所有左侧元素小于等于  $x$ , 而  $x$  小于等于所有右侧元素。但在查找、删除时需要额外处理重复元素。一般我们限制 B 树中的键唯一, 而一个键可以对应多个值( $k \mapsto [v_1, v_2, \dots]$ ), 称为多值映射(multi-map)。

## 7.1.2. 使用循环消除“先分拆再插入”算法中的递归。

```
BTree<K, deg> insertNonfull(BTree<K, deg> tr, K key) {
```

```

var root = t
while not is_leaf(t) {
  Int i = length(t.keys)
  while i > 0 and key < t.keys[i-1] {
    i = i - 1
  }
  if full(d, t.subTrees[i]) {
    split(d, t, i)
    if key > t.keys[i] then i = i + 1
  }
  t = t.subTrees[i]
}
orderedInsert(t.keys, key)
return root

```

7.1.3. 我们使用线性查找获得元素插入的位置。请使用二分查找对命令式实现进行改进。算法复杂度会提升么？

```

void orderedInsert([K] xs, K x) {
  append(xs, x)
  Int p = binarySearch(xs, x)
  for Int i = length(lst), i > p, i = i - 1 {
    xs[i] = xs[i-1]
  }
  xs[p] = x
}

Int binarySearch([K] xs, K x) {
  Int l = 0, u = length(xs)
  while l < u {
    Int m = (l + u) / 2
    if xs[m] == x {
      return m
    } else if xs[m] < x {
      l = m + 1
    } else {
      u = m
    }
  }
  return l
}

```

复杂度仍然是线性时间的。尽管二分查找需要  $O(\lg n)$  时间，但是插入需要  $O(n)$  时间移动数组以获取空位。

## 答案 7.2

7.2.1. 使用二分查找改进命令式查找算法。

```

Optional<(BTree<K, deg>, Int)> lookup(BTree<K, deg> tr, K key) {
  loop {
    Int l = 0, u = length(tr.keys)

```

```

while l < u {
  Int m = (l + u) / 2
  if key == tr.keys[m] {
    return Optional.of((tr, m))
  } else if t.keys[m] < key {
    l = m + 1
  } else {
    u = m
  }
}
if isLeaf(tr) {
  return Optional.Nothing
} else {
  tr = tr.subTrees[l]
}
}
}

```

### 答案 7.3

7.3.1. 我们在本节中使用了前驱子分枝中的最大元素  $k' = \max(t')$  替换要删除的元素  $k$ , 然后递归地在  $t'$  中删除  $k'$ 。还有一种对称的处理方法: 用后继分枝中的最小元素来替换  $k$ 。请实现这一方法

我们首先要定义 *min* 函数, 获取 B 树中的最小元素:

```

min' (BTree ks []) = head ks
min' (BTree _ ts) = min' $ head ts

```

在用  $x$  划分树时, 需要用  $\leq$  替换  $<$  使得  $x$  可能出现在左侧的尾部。为此我们将划分条件抽出作为参数传入:

```

partitionWith p (BTree ks ts) = (l, t, r) where
  l = (ks1, ts1)
  r = (ks2, ts2)
  (ks1, ks2) = L.span p ks
  (ts1, (t:ts2)) = L.splitAt (length ks1) ts

```

接下来就可以对称地实现删除了:

```

delete' x (d, t) = fixRoot (d, del x t) where
  del x (BTree ks []) = BTree (L.delete x ks) []
  del x t = if (Just x) == (listToMaybe $ reverse ks') then
    let k' = min' t' in
      balance d ((init ks') + [k'], ts') (del k' t') r
    else balance d l (del x t') r
  where
    (l@(ks', ts'), t', r) = partitionWith (<= x) t

```

7.3.2. 实现列表对 B 树的删除算法。

```

delete x (d, t) = fixRoot (d, del x t) where
  del _ Empty = Empty
  del x t = if (Just x) == fmap fst (listToMaybe r) then
    case t' of
      Empty → balance d l Empty (tail r)
      _     → let k' = max' t' in
              balance d l (del k' t') ((k', snd $ head r):(tail r))
    else balance d l (del x t') r
  where
    (l, t', r) = partition (< x) t

max' t@(BTree _ _ []) = max' (stepL t)
max' (BTree _ _ [(k, Empty)]) = k
max' (BTree _ _ [(k, t)]) = max' t
max' t = max' (stepR t)

```

此外,我们需要对 *balance* 函数增加额外的处理,修复删除后不平衡的情形:

```

balance :: Int → [(a, BTree a)] → BTree a → [(a, BTree a)] → BTree a
balance d l t r | full d t = fixFull
                | low d t = fixLow l t r
                | otherwise = BTree l t r
  where
    fixFull = let (t1, k, t2) = split d t in BTree l t1 ((k, t2):r)
    fixLow ((k', t'):l) t r = balance d l (unsplit t' k' t) r
    fixLow l t ((k', t'):r) = balance d l (unsplit t k' t') r
    fixLow l t r = t      — l == r == []

```

其中 *unsplit* 是 *split* 的逆运算,它的定义为:

```

unsplit t1 k t2@(BTree (_:_ ) _ ) = unsplit t1 k (stepL t2)
unsplit t1@(BTree _ _ (_:_ )) k t2 = unsplit (stepR t1) k t2

```

## 答案 8.1

- 8.1.1. 不正确。 $[a_2, a_3, \dots, a_n]$  的结构不再是一个堆,仅对第一个元素使用 HEAPIFY 是不够的,需要用 BUILD-HEAP 重建堆。
- 8.1.2. 不可以,原因同上。

## 答案 8.2

- 8.2.1. 用命令式的方式实现左偏堆、斜堆。

我们先实现左偏堆,为了方便回溯,我们在节点中增加一个父引用:

```

data Node<T> {
  T value
  Int rank = 1
  Node<T> left = null, right = null, parent = null
}

```

合并两个左偏堆时,我们先自顶向下沿着右子树合并,然后在沿着父引用更新秩。如果左侧的秩较小就交换左右子树。为了方便处理空树,我们引入一个空节点指向根。

```

Node<T> merge(Node<T> a, Node<T> b) {
    var h = Node(null) // the sentinel node
    while a ≠ null and b ≠ null {
        if b.value < a.value then swap(a, b)
        var c = Node(a.value, parent = h, left = a.left)
        h.right = c
        h = c
        a = a.right
    }
    h.right = if a ≠ null then a else b
    while h.parent ≠ null {
        if rank(h.left) < rank(h.right) then swap(h.left, h.right)
        h.rank = 1 + rank(h.right)
        h = h.parent
    }
    h = h.right
    if h ≠ null then h.parent = null
    return h
}

Int rank(Node<T> x) = if x ≠ null then x.rank else 0

Node<T> insert(Node<T> h, T x) = merge(Node(x), h)

T top(Node<T> h) = h.value

Node<T> pop(Node<T> h) = merge(h.left, h.right)

```

斜堆的实现更简单,我们无需维护子树的秩,省去了回溯的过程,无需父节点引用:

```

data Node<T> {
    T value
    Node<T> left = null, right = null
}

```

我们仅实现合并函数,其它函数和左偏堆相同:

```

Node<T> merge(Node<T> a, Node<T> b) {
    var h = Node(None)
    var root = h
    while a ≠ null and b ≠ null {
        if b.value < a.value then swap(a, b)
        var c = Node(a.value, left = null, right = a.left)
        h.left = c
        h = c
        a = a.right
    }
    h.left = if a ≠ null then a else b
}

```



```

    root = root.left
    return root
}

```

### 8.2.2. 定义堆的叠加操作

$$\begin{aligned}
 \text{fold } f \ z \ \emptyset &= z \\
 \text{fold } f \ z \ H &= \text{fold } f \ (f \ (\text{top } H) \ z) \ (\text{pop } H)
 \end{aligned}$$

### 答案 9.1

9.1.1. 这里需要用链接而不是追加。追加操作的复杂度是线性的,和序列的长度成正比,而链接是常数时间的。

9.1.2. 实现原地选择排序程序。

```

Void sort([K] xs) {
    var n = length(xs)
    for var i = 0 to n - 1 {
        var m = i
        for Int j = i + 1 to n - 1 {
            if xs[j] < xs[m] then m = j
        }
        swap(xs[i], xs[m])
    }
}

```

### 答案 9.2

9.2.1. 将递归的锦标赛排序实现为升序。

把  $max$  和  $-\infty$  替换为  $min$  和  $\infty$  就可以实现升序排序。我们可以进一步把它们抽象成参数:

```

minBy p a b = if p a b then a else b

merge p t1 t2 = Br t1 (minBy p (key t1) (key t2)) t2

fromListWith p xs = build $ map wrap xs where
    build [] = Empty
    build [t] = t
    build ts = build $ pair ts
    pair (t1:t2:ts) = (merge p t1 t2) : pair ts
    pair ts = ts

popWith p inf = delMin where
    delMin (Br Empty _ Empty) = Br Empty inf Empty
    delMin (Br l k r) | k == key l = let l' = delMin l in
        Br l' (minBy p (key l') (key r)) r
        | k == key r = let r' = delMin r in

```

```

      Br l (minBy p (key l) (key r')) r'

toListWith p inf = flat where
  flat Empty = []
  flat t | inf == key t = []
         | otherwise = (top t) : (flat $ popWith p inf t)

sortBy p inf xs = toListWith p inf $ fromListWith p xs where

```

`sortBy (<) ∞` 定义了升序排序, 而 `sortBy (>) - ∞` 定义了降序排序。

### 9.2.2. 锦标赛树排序可以处理相等元素么? 它是稳定排序么?

可以处理相等元素。使用练习 9.2.1 的结果, `sortBy (≤) ∞` 即可实现含有相等元素的升序排序。它不是稳定排序。

### 9.2.3. 比较锦标赛树排序和二叉搜索树排序, 它们的时间和空间效率如何。

它们的时间复杂度都是  $O(n \lg n)$ , 空间复杂度都是  $O(n)$ 。区别在于, 二叉搜索树建立后不再改变(除非插入、删除)。而锦标赛树在排序后变成含有  $n$  个无穷节点的空树。

### 9.2.4. 比较堆排序和锦标赛树排序, 它们的时间和空间效率如何。

它们的时间复杂度都是  $O(n \lg n)$ , 空间复杂度都是  $O(n)$ 。区别在于, 堆排序结束后堆变空, 而锦标赛树在排序后仍然占据  $O(n)$  的空间。

## 答案 10.1

### 10.1.1. 编程产生帕斯卡三角形

```

pascal = gen [1] where
  gen cs (x:y:xs) = gen ((x + y) : cs) (y:xs)
  gen cs _ = 1 : cs

```

任给帕斯卡三角形中的一行, 此函数产生下一行。可以这样产生前  $n$  行:

```
take n (iterate pascal [1])
```

### 10.1.2. 证明二项式树 $B_n$ 中第 $i$ 行的节点数为 $\binom{n}{i}$ 。

证明. 使用数学归纳法。  $B_0$  只有一个根节点。设  $B_n$  中每行满足二项式系数。树  $B_{n+1}$  由两棵  $B_n$  组成。第 0 行为根节点:  $1 = \binom{n+1}{0}$ 。第  $i$  行的节点包含两部分, 一部分是最左侧子树  $B_n$  的第  $i-1$  行, 一部分是另一棵  $B_n$  的第  $i$  行, 总共有:

$$\begin{aligned}
 \binom{n}{i-1} + \binom{n}{i} &= \frac{n!}{(i-1)!(n-i+1)!} + \frac{n!}{i!(n-i)!} \\
 &= \frac{n![i + (n-i+1)]}{i!(n-i+1)!} \\
 &= \frac{(n+1)!}{i!(n-i+1)!} \\
 &= \binom{n+1}{i}
 \end{aligned}$$

□

10.1.3. 证明二项式树  $B_n$  中含有  $2^n$  个节点。

证明. 根据上一练习中证明的结果,  $B_n$  中各行节点相加为:

$$\begin{aligned} & \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} \quad \text{各行相加} \\ = & (1+1)^n \quad \text{令二项式}(a+b)^n \text{中} a=b=1 \\ = & 2^n \end{aligned}$$

□

10.1.4. 用容器保存子树, 实现二项式树的链接程序。这种方式有何问题, 怎样解决?

如果用数组保存子树, 在所有元素前插入需要线性时间:

```

1: function LINK'(T1, T2)
2:   if KEY(T2) < KEY(T1) then
3:     Exchange T1 ↔ T2
4:   PARENT(T2) ← T1
5:   INSERT(SUB-TREES(T1), 1, T2)
6:   RANK(T1) ← RANK(T2) + 1
7:   return T1

```

为此, 我们可以将子树逆序保存。这样将新树附加到末尾只需要常数时间。

### 答案 10.2

试说明 DECREASE 的分摊复杂度为常数时间  $O(1)$ 。

定义势能函数:

$$\Phi(H) = t(H) + 2m(H)$$

其中  $t(H)$  为堆中树的棵数,  $m(H)$  是带有标记的节点。每个标记成对设置、切下并清除, 其系数为 2。DECREASE 用  $O(1)$  时间将  $x$  切下, 然后递归调用 CASCADE-CUT。设递归  $c$  次。每次用  $O(1)$  时间调用 CUT, 然后递归。因此 DECREASE 的总消耗为  $O(c)$ 。

接下来计算势能的变化。令  $H$  是调用 DECREASE 前的堆。每次递归的 CASCADE-CUT 切下一个带标记的节点, 并清除其标记 (最后一次除外)。结束后, 堆中共有  $t(H) + c$  棵树。包括原来的  $t(H)$  棵树, 切下后放入堆中的  $c - 1$  棵树, 以及以  $x$  为根的子树。最多有  $m(H) - c + 2$  个带标记的节点。即原有的  $m(H)$  个节点减去  $c - 1$  个在 CASCADE-CUT 中清除标记的节点。最后一次递归调用中可能标记一个节点。这样势能的变化最多为:

$$t(H) + c + 2(m(H) - c + 2) - [t(H) + 2m(H)] = 4 - c$$

这样分摊成本最多为  $O(c) + 4 - c = O(1)$ 。

### 答案 10.3

- 10.3.1. 如果连续向配对堆中插入  $n$  个元素, 然后执行弹出。当  $n$  很大时, 弹出的单次性能变得很差(尽管分摊性能是  $O(\lg n)$ )。如何避免这一最差情况?

设置子树棵树的阈值  $m$ 。在插入时检查, 如果超过阈值就执行一次弹出操作, 然后把弹出的元素再加入回堆。

```
MAX_SUBTREES = 16

Node<K> insert(Node<K> h, K x) {
    if h ≠ null and length(h.subTrees) > MAX_SUBTREES {
        h = insert(pop(h), top(h))
    }
    return merge(h, Node(x))
}
```

- 10.3.2. 实现配对堆的删除。

为了方便删除, 我们给每个节点增加一个父引用:

```
data Node<K> {
    K key
    Node<K> parent = null
    [Node<K>] subTrees = []

    Node<K>(K k) { key = k }
}
```

考虑删除值为  $x$  的元素。我们先从堆  $h$  中找到以  $x$  为根的子树  $t$ 。如果  $t$  是  $h$  的根, 我们只要执行弹出操作  $pop$  即可, 否则我们通过  $t$  的父节点  $p$ , 从子树中把  $t$  切下。然后对  $t$  执行弹出操作, 最后将  $pop(t)$  合并回  $h$ 。

```
Node<K> delete(Node<K> h, K x) {
    var tr = lookuptr(h, x)
    if tr == null then return h
    if tr == h then return pop(h)
    tr.parent.subtrees.remove(tr)
    tr.parent = null
    return merge(pop(tr), h)
}

Node<K> lookuptr(Node<K> h, K x) {
    if h.key == x then return h
    for var t in h.subtrees {
        var tr = lookuptr(t, x)
        if tr ≠ null then return tr
    }
    return null
}
```

```
}
}
```

递归查找的时间复杂度为  $O(n)$ , 其中  $n$  是堆中元素的数目, 删除的复杂度为  $O(m)$ , 其中  $m$  是所在节点包含子树的棵树。总体复杂度为  $O(n)$ 。

### 10.3.3. 实现配对堆的优先级调整 DECREASE-KEY

如果调整根节点  $h$  的值, 我们直接用新值  $x$  覆盖。否则, 我们通过  $tr$  的父节点从子树中将  $tr$  切下, 更新  $tr$  的值为  $x$  后, 将  $tr$  合并回  $h$ :

```
Node<K> decreaseKey(Node<K> h, Node<K> tr, K x) {
    if tr == null or tr.key < x then return h
    tr.key = x
    if tr == h then return h
    tr.parent.subtrees.remove(tr) // O(m), where m = length(subtrees)
    tr.parent = null
    return merge(tr, h)
}
```

使用时, 我们可以先调用 `lookuptr(h, y)` 获得要调整的节点, 然后把  $y$  的值更新为  $x$ , 即: `decreaseKey(h, lookuptr(h, y), x)`。其复杂度和删除相同, 为  $O(n)$ 。

## 答案 11.1

循环缓冲区在初始化时规定了最大的容量, 如果使用头、尾两个指针, 而不用 `count`, 如何检测队列是否为空? 是否已满?

尽管可以分别考虑头部在尾部前面、后面的情况, 如图 11.4, 我们思考更简单一致的方法。考虑向左右两侧无限伸展的区域, 记头部索引为  $h$ 、尾部为  $t$ 、缓冲区容量为  $s$ 、含有内容的左闭右开区间为  $[h, t)$ 。空、满判断条件如下:

$$\begin{cases} \text{empty}(h, t) : & h = t \\ \text{full}(h, t) : & t - h = s \end{cases}$$

循环缓冲区相当于在此基础上引入了模运算(时钟运算), 记  $[n]_s = n \bmod s$ 。对上述判满条件取模就会发现:  $[t]_s - [h]_s = [s]_s = 0$ , 即  $[t]_s = [h]_s$ 。这恰恰是判空条件取模的结果。因此仅仅分析头、尾取模后的结果是无法判断队列空、满的。我们要么引入一个标记(标记  $h$ 、 $t$  的先后顺序), 要么在空、满判定时不取模, 而仅仅在索引元素时取模。考虑实现时的整数字长限制, 可以选择一个与  $s$  互素的大数  $p(p > s)$  取模进行判定, 即:

$$\begin{cases} \text{empty}(h, t) : & [h]_p = [t]_p \\ \text{full}(h, t) : & [t - h]_p = s \end{cases}$$

例子程序:

```

Int P_LIMIT = 104743 // the 10000th prime
Bool empty(Queue<K> q) = (q.h == q.t)
Bool full(Queue<K> q) = (q.s == (q.t - q.h) mod P_LIMIT)

void enqueue(Queue<K> q, K x) {
    if not full(q) {
        q.t = (q.t + 1) mod P_LIMIT
        q.buf[q.t mod q.s] = x
    }
}

Optional<K> dequeue(Queue<K> q) {
    Optional<K> x = Optional.Nothing
    if not empty(q) {
        x = Optional.of(q.buf[q.h mod q.s])
        q.h = (q.h + 1) mod P_LIMIT
    }
    return x
}

```

## 答案 11.2

11.2.1. 为什么要在 `push` 时也要进行平衡检查和调整?  
考虑这样的情况:先 `push a ([ ], [ ])`, 然后再 `pop`。

11.2.2. 试分析双列表队列的分摊复杂度。

使用记账法分析。尾部列表  $r$  中的每个元素记 1 分。向尾部列表 `push`, 执行一次加入操作, 并增加 1 分。分摊消耗为  $O(2)$ 。每次 `pop`, 如果未造成列表反转, 则执行一次取出操作, 但没有增减分数。分摊消耗为  $O(1)$ 。如果造成列表反转, 则执行  $m$  次反转和 1 次取出, 其中  $m$  是尾部列表  $r$  的长度, 并且使用了  $r$  列表中的  $m$  分。所以分摊消耗为  $O(m + 1 - m) = O(1)$ 。

11.2.3. 实现双数组队列。

```

1: function PUSH(Q, x)
2:   APPEND(FRONT(Q), x)

3: function POP(Q)
4:   if REAR(Q) = [ ] then
5:     REAR(Q) ← REVERSE(FRONT(Q))
6:     FRONT(Q) ← [ ]
7:   n ← LENGTH(REAR(Q))
8:   x ← REAR(Q)[n]
9:   LENGTH(REAR(Q)) ← n - 1
10:  return x

```

## 答案 11.3

11.3.1. 在 *abort* 函数中, 当  $n = 0$  时, 我们实际上撤销了上一个链接元素的操作, 去掉了  $x$  而返回  $a$  作为结果。为什么需要回滚一个元素?

只有弹出操作 *pop* 会调用 *abort* 函数。  $n = 0$  时, 轮转操作刚完成, 状态即将从  $(S_c, 0, (x:a), f')$  转换成  $(S_f, a)$ 。而上次链接好的  $x$  正是要弹出的元素, 因此需要去掉  $x$  返回  $a$  作为结果。

11.3.2. 使用双数组实现实时队列。注意: 当开始递进反转时, 不能一次性复制数组, 否则就会将性能降低到线性时间。请实现一个惰性复制, 使得每步反转时仅复制一个元素。

设可以在常数时间内获得数组的长度。我们在数组  $f$  的尾部加入元素 (*push*), 在数组  $r$  的尾部弹出元素 (*pop*)。双数组不平衡时, 启动一个状态机分步计算  $acc = reverse(f) ++ r$ 。如果  $f \neq []$ , 我们取出尾部元素, 添加到  $acc$  的末尾。当  $f$  反转结束后, 我们从  $r$  左侧逐一将元素加到  $acc$  末尾, 即  $append(acc, r[i])$ , 其中  $i = 0, 1, \dots, |r| - 1$ 。分步计算时仍然可能从  $r$  的末尾弹出元素, 当  $i$  超出  $|r|$  时计算结束。

```

data State<K> {
  [K] acc, front, rear
  Int idx

  State([K] f, [K] r) {
    acc = [], front = f, rear = r
    idx = 0
  }

  // compute reverse(f) ++ r step by step
  Self step() {
    if front  $\neq$  [] then acc.append(front.popLast()) // reversing
    if s.front == [] and idx < length(rear) { // concatenating
      acc.append(rear[idx])
      idx = idx + 1
    }
  }

  Bool done() = (front == [] and length(rear) < idx)
}

data RealtimeQueue<K> {
  [K] front = []
  [K] rear = []
  State<K> state = null

  Bool isEmpty() = (front == [] and rear == [])

  Self push(K x) {
    front.append(x)
    balance()
  }
}

```

```

}

K pop() {
  x = rear.popLast()
  balance()
  return x
}

Void balance() {
  if state == null and length(rear) < length(front) {
    state = State(front, rear).step()
    front = []
  }
  if state != null and state.step().done() {
    rear = state.acc
    state = null
  }
}
}

```

### 答案 11.4

11.4.1. 如何实现双向队列,在头部尾部都支持常数时间的元素添加和删除。

双列表或双数组都可以实现双向队列。以双列表为例,定义两对操作: $push_l/pop_l$ 和 $push_r/pop_r$ 。在 $pop_l/pop_r$ 时按需进行反转。以下实现每次按需反转一半元素。

```

empty = ([], [])

isEmpty ([], []) = True
isEmpty _ = False

pushL x (f, r) = (x:f, r)

pushR (f, r) x = (f, x:r)

popL ([], []) = (Nothing, empty)
popL ([], r) = let (as, bs) = splitAt (length r `div` 2) r in
  popL (reverse bs, as)
popL (x:f, r) = (Just x, (f, r))

popR ([], []) = (empty, Nothing)
popR (f, []) = let (as, bs) = splitAt (length f `div` 2) f in
  popR (as, reverse bs)
popR (f, x:r) = ((f, r), Just x)

```

### 答案 12.1

12.1.1. 如何处理索引越界情况?



我们可以引入 Maybe 类型来处理越界情况。如果索引  $i < 0$ , 返回 Nothing, 如果索引过大, 则最终递归转换成森林为空的情形, 返回 Nothing。

```

getAt [] _ = Nothing
getAt (t:ts) i | i < 0 = Nothing
                | i < size t = lookupTree i t
                | otherwise = getAt ts (i - size t)

where
  lookupTree 0 (Leaf x) = Just x
  lookupTree i (Node sz t1 t2) = if i < sz `div` 2 then lookupTree i t1
                                else lookupTree (i - sz `div` 2) t2

```

## 答案 12.2

12.2.1. 实现数值表示序列的随机访问  $S[i], 1 \leq i \leq n$ 。其中  $n$  是序列长度。

我们跳过索引越界的情况:

```

getAt (Zero:ts) i = getAt ts i
getAt (One t:ts) i = if i < size t then lookupTree i t
                    else getAt ts (i - size t)

where
  lookupTree 0 (Leaf x) = x
  lookupTree i (Node sz t1 t2) = if i < sz `div` 2 then lookupTree i t1
                                else lookupTree (i - sz `div` 2) t2

```

12.2.2. 分析数值表示序列删除操作的分摊复杂度。

考虑插入的逆过程, 从含有  $n = 2^m$  的序列依次删除元素, 直到序列变空。相应的分析和插入对称, 删除的分摊复杂度是常数时间。

12.2.3. 可以用长度为  $2^m$  的数组表示完全二叉树 ( $m$  是非负整数)。请用数组实现二叉树森林的插入和随机访问。

```

data List<K> {
  Int size = 0
  [[K]] trees = [[]]
}

Int nbits(Int n) {
  Int i = 0
  while n ≠ 0 {
    i = i + 1
    n = n / 2
  }
  return i
}

List<K> insert(List<K> xs, K x) {
  var i = nbits(xs.size xor (1 + xs.size)) // locate the bit flip to 1
  if i ≥ length(xs.trees) then xs.trees.append([])
  xs.trees[i] = [x]
  for Int j = 0, j < i, j++ {

```

```

        xs.trees[i] = xs.trees[i] ++ xs.trees[j]
        xs.trees[j] = []
    }
    xs.size = xs.size + 1
    return xs
}

Optional<K> get(List<K> xs, Int i) {
    for t in xs.trees {
        Int size = length(t)
        if i < size then return Optional.of(t[i])
        i = i - size
    }
    return Optional.Nothing
}

```

### 答案 12.3

12.3.1. 试分析双数组序列删除的分摊复杂度。

定义双数组序列的势能为数组长度差： $\Phi(s) = |r| - |f| = n - m$ ，其中  $m = |f|$ 、 $n = |r|$ 。从头部删除时，若  $f \neq []$ ，则直接用  $O(1)$  时间去掉  $f$  的最后一个元素。如果  $f = []$ ，用  $O(n)$  时间将  $r$  折半、反转替换成  $f'$ 。再用  $O(1)$  去掉  $f'$  的最后一个元素。这样分摊代价为：

$$\begin{aligned}
 c &= n + 1 + \Phi(s') - \Phi(s) \\
 &= n + 1 + (|r'| - |f'|) - (|r| - |f|) \\
 &= n + 1 + (n - \lceil \frac{n}{2} \rceil) - (\lceil \frac{n}{2} \rceil - 1) - (n - 0) \\
 &= 1
 \end{aligned}$$

故从头部删除的分摊复杂度为  $O(1)$ ，对称地，从尾部删除的分摊复杂度也是  $O(1)$ 。

### 答案 12.4

12.4.1. 消除递归，用循环的方式实现手指树插入。对于手指树  $T = (f, t, r)$ ，令  $\text{MID}(T) = t$  以获取中间部分。

```

1: function INSERT(x, T)
2:   n = (x)
3:    $\perp \leftarrow p \leftarrow ([], T, [])$ 
4:   while |FRONT(T)| ≥ 3 do
5:     f ← FRONT(T)
6:     n ← (f[2], f[3], ...)
7:     FRONT(T) ← [n, f[1]]

```

```

8:      $p \leftarrow T$ 
9:      $T \leftarrow \text{MID}(T)$ 
10:  if  $T = \text{NIL}$  then
11:      $T \leftarrow ([n], \text{NIL}, [ ])$ 
12:  else if  $|\text{FRONT}(T)| = 1$  and  $\text{REAR}(T) = [ ]$  then
13:      $\text{REAR}(T) \leftarrow \text{FRONT}(T)$ 
14:      $\text{FRONT}(T) \leftarrow [n]$ 
15:  else
16:      $\text{INSERT}(\text{FRONT}(T), n)$ 
17:   $\text{MID}(p) \leftarrow T$ 
18:   $T \leftarrow \text{MID}(\perp), \text{MID}(\perp) \leftarrow \text{NIL}$ 
19:  return  $T$ 

```

我们将待插入元素  $x$  装入一个单元素叶子 ( $x$ )。如果  $f$  手指超出 3 个元素, 就沿着中间子树, 进行一次自顶向下的遍历。我们将  $f$  中除第一个元素外的剩余部分抽出, 放入一个新节点  $n$  (深度增加 1), 然后继续将  $n$  插入到中间子树中。将待插入内容和  $f$  中剩下的一个元素组成新的  $f$  手指。遍历结束后, 我们要么到达了一个空树, 要么到达了一棵子树, 这棵子树的  $f$  手指仍可容纳更多元素。对于第一种情况, 我们创建一个叶子节点, 对于后一种情况, 我们将  $n$  插入到  $f$  最前面。最后, 我们返回树的根  $T$ 。为了简化处理, 我们创建了一个特殊的  $\perp$  节点, 它是根节点的父亲节点。

### 答案 12.5

#### 12.5.1. 消除递归, 用循环实现删除。

如果删除后 front 手指变空, 就从中部分子树中“借”节点。但是树的形式有可能是不规则的, 例如 front 手指和中间子树都为空。这种情况通常是由于分拆操作造成的。

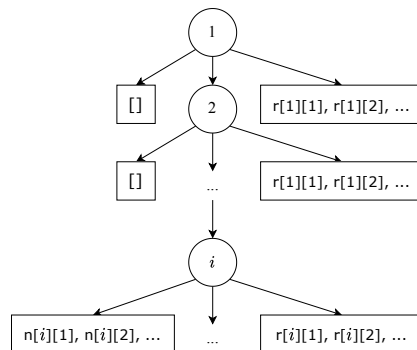
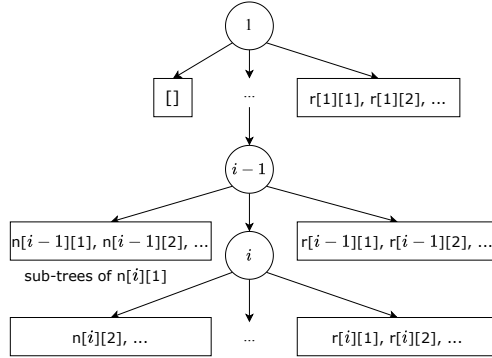


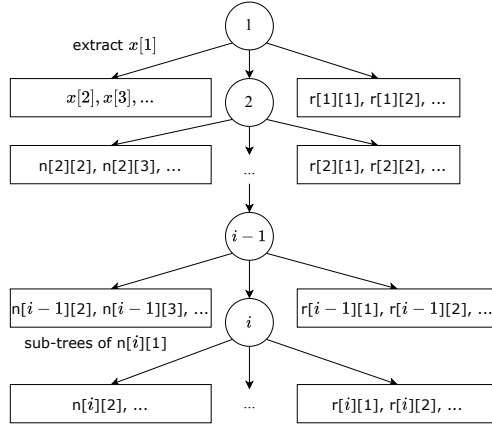
图 12.9: 不规则树, 第  $i$  层子树的  $f$  不空

我们要从不规则的手指树中删除第一个元素。首先进行一轮自顶向下的遍历,

找到一棵子树,或者它的  $f$  不为空,或者  $f$  和中间子树都为空,如图 12.9。对于前者,我们可以从  $f$  中提取出第一个元素(是一个节点)。对于后者,由于只有  $r$  不为空,我们交换  $f, r$ , 转换成前一种情况。此后,我们检查从  $f$  中取出的节点是否为叶子。如果不是,我们需要继续提取。我们沿着父节点一直向上回溯,直到我们提取到一个叶子节点。此时我们将到达树的根节点。图 12.10 描述了这一过程。



提取第一个元素  $n[i][1]$ , 然后将它的子节点放到上一级树的  $f$  手指中。



重复  $i$  次,最终提取到  $x[1]$ 。

图 12.10: 自底向上遍历,直到提取出一个叶子节点

根据这一思路,下面的算法实现了列表头部的删除操作(假设树不为空)。

```

1: function EXTRACT( $T$ )
2:    $\perp \leftarrow ([], T, [])$ 
3:   while FRONT( $T$ ) = [ ] and MID( $T$ )  $\neq$  NIL do
4:      $T \leftarrow$  MID( $T$ )
5:   if FRONT( $T$ ) = [ ] and REAR( $T$ )  $\neq$  [ ] then
6:     EXCHANGE FRONT( $T$ )  $\leftrightarrow$  REAR( $T$ )
7:    $f \leftarrow$  FRONT( $T$ ),  $r \leftarrow$  REAR( $T$ )
8:    $n \leftarrow (f[1], f[2], \dots)$ 

```

$\triangleright n$  是 2-3 树

```

9:   repeat
10:     FRONT( $T$ )  $\leftarrow [n_2, n_3, \dots]$ 
11:      $n \leftarrow n_1$ 
12:      $T \leftarrow \text{PARENT}(T)$ 
13:     if MID( $T$ ) becomes empty then
14:       MID( $T$ )  $\leftarrow \text{NIL}$ 
15:   until  $n$  is leaf
16:   return (ELEM( $n$ ), MID( $\perp$ ))

```

这里函数 ELEM( $n$ ) 返回叶子节点  $n$  中保存的唯一元素。由于不规则的树存在，需要调整从手指树中获取第一个和最后一个元素的定义。我们不再仅返回手指的第一个或者最后一个子节点。如果手指为空，而中间子树不为空，我们就沿着中间部分向下遍历，直到发现手指不为空，或者所有的节点都存储在另一侧的手指中。

```

1: function FIRST-LEAF( $T$ )
2:   while FRONT( $T$ ) = [] and MID( $T$ )  $\neq$  NIL do
3:      $T \leftarrow \text{MID}(T)$ 
4:   if FRONT( $T$ ) = [] and REAR( $T$ )  $\neq$  [] then
5:      $n \leftarrow \text{REAR}(T)[1]$ 
6:   else
7:      $n \leftarrow \text{FRONT}(T)[1]$ 
8:   while  $n$  is NOT leaf do
9:      $n \leftarrow n_1$ 
10:  return  $n$ 

11: function FIRST( $T$ )
12:  return ELEM(FIRST-LEAF( $T$ ))

```

其中第二个循环中，如果节点不是叶子，就不断沿着第一个子节点遍历。获取最后一个元素与此类似。

### 答案 12.6

12.6.1. 在随机访问时，如何处理空树  $\emptyset$  和索引越界的情况？

我们可以在索引时进行越界检查，例如：

$$\begin{aligned}
 \emptyset[i] &= \text{Nothing} \\
 T[i] &= \begin{cases} i < 0 \text{ 或 } i \geq \text{size } T : \text{Nothing} \\ \text{否则} : & \dots \end{cases}
 \end{aligned}$$



```

cutNode :: (Sized a) => Int -> Node a -> ([a], Maybe (Place a), [a])
cutNode i (Tr2 _ a b) | i < sa = ([], Just (Place i a), [b])
                    | otherwise = ([a], Just (Place (i - sa) b), [])
    where sa = size a
cutNode i (Tr3 _ a b c) | i < sa = ([], Just (Place i a), [b, c])
                    | i < sab = ([a], Just (Place (i - sa) b), [c])
                    | otherwise = ([a, b], Just (Place (i - sab) c), [])
    where sa = size a
          sab = sa + size b

```

我们可以用分割修改指定位置的元素、删除指定位置的元素、以及 MTF 操作，它们的复杂度都是  $O(\lg n)$ 。

```

setAt s i x = case cut i s of
  (_, Nothing, _) -> s
  (xs, Just y, ys) -> xs ++ (x <| ys)

extractAt s i = case cut i s of (xs, Just y, ys) -> (y, xs ++ ys)

moveToFront i s = if i < 0 || i >= size s then s
                 else let (a, s') = extractAt s i in a <| s'

```

### 答案 13.1

13.1.1. 改进基本快速排序的定义，优化对单元素列表的处理。增加一条处理

```
sort [x] = [x]
```

### 答案 13.2

13.2.1. 两两合并更快。这个问题本质上是合并  $k$  个已序序列。简化问题，设每个序列的平均长度为  $n$ 。使用叠加相当于先合并  $s_1$  和空序列， $merge([], s_1) = [] \oplus s_1$ ，然后合并  $s_2$ ， $[\ ] \oplus s_1 \oplus s_2$ ，然后再合并  $s_3$ ， $[\ ] \oplus s_1 \oplus s_2 \oplus s_3 \cdots$  所以总复杂度为： $O(n + 2n + 3n + 4n + \dots + kn) = O(n \frac{k(k+1)}{2}) = O(nk^2)$ 。而两两合并，第一轮是  $s_1 \oplus s_2, s_3 \oplus s_4 \cdots$  用时  $O(kn)$ ，接下来一轮： $(s_1 \oplus s_2) \oplus (s_3 \oplus s_4) \cdots$  又是  $O(kn)$ 。总共进行  $\lg k$  轮，总复杂度为： $O(nk \lg k)$ 。故两两合并的速度更快。我们也可以利用一个大小为  $k$  的小顶堆，保存  $k$  个序列各自的最小元素进行合并，复杂度也是  $O(nk \lg k)$ 。

### 答案 13.3

13.3.1. 将两两组合的操作抽象成叠加  $foldp$ ，并使用它定义自底向上的归并排序。我们要求二元组合函数  $f$  满足结合性条件： $f(f(x, y), z) = f(x, f(y, z))$ 。令二元运算  $f$  的单位元为  $z$ ，定义两两叠加为：

```

foldp f z [] = z
foldp f z [x] = f x z
foldp f z xs = foldp f z (pairs xs) where
  pairs (x:y:ys) = (f x y) : pairs ys
  pairs ys = ys

```

这样累加可定义为  $sum = foldp (+) 0$ , 而自底向上的归并排序定义为<sup>6</sup>:

```

sort = foldp merge [] ◦ map (:[])

```

### 答案 13.4

#### 13.4.1. 使用归并排序的策略从一个序列产生二叉搜索树。

可以自顶向下或自底向上构建二叉搜索树。自顶向下: 均分序列, 递归构建两棵二叉搜索树再合并; 自底向上: 将每个元素封装成单叶子节点, 然后不断两两合并成一棵树。两种方法都依赖二叉树合并<sup>8</sup>。若任一棵树为空, 合并结果为另外一棵。否则令两棵树为  $(A, x, B)$  和  $(C, y, D)$ 。若  $x < y$ , 用  $y$  划分  $B$  为两棵树:  $B_y, B^y$ , 其中  $B_y$  包含全部  $\{b < y\}$  的节点,  $B^y$  包含  $\{b \geq y\}$  的节点; 用  $x$  划分  $C$  为  $C_x, C^x$ 。然后构建合并结果:

$$(A, x, B) \oplus (C, y, D) = ((A \oplus C_x, x, B_y \oplus C^x), y, D \oplus B^y)$$

其中 ‘ $\oplus$ ’ 表示树合并。对称地, 若  $x \geq y$ , 我们划分子树  $A, D$ 。下面的例子程序实现了树划分、合并。它使用了练习 13.3 定义的两两叠加  $foldp$ 。

```

toTree :: (Ord a) => [a] -> Tree a
toTree = foldp merge Empty ◦ map leaf

partition x Empty = (Empty, Empty)
partition x (Node a y b)
  | x < y = let (a1, a2) = partition x a in (a1, Node a2 y b)
  | otherwise = let (b1, b2) = partition x b in (Node a y b1, b2)

merge Empty t = t
merge t Empty = t
merge (Node a x b) (Node c y d)
  | x < y = let
    (b1, b2) = partition y b
    (c1, c2) = partition x c
  in Node (Node (merge a c1) x (merge b1 c2)) y (merge d b2)
  | otherwise = let
    (a1, a2) = partition y a
    (d1, d2) = partition x d
  in Node (merge a1 c) y (Node (merge a2 d1) x (merge b d2))

```

<sup>6</sup>  $(:[])$  相当于  $x xs \mapsto [x]:xs$ 。

<sup>8</sup> 如果是二叉搜索树合并的独立题目, 可以将每棵树转换成数组或列表(中序遍历), 然后利用式 (13.31) 合并, 最后用中间位置元素作为根结点重建二叉树。



## 答案 14.1

14.1.1. 证明  $k$  选择的平均复杂度为  $O(n)$ 。

参考快速排序的复杂度分析 13.1.3。

14.1.2. 为了查找  $A$  中的前  $k$  小元素,我们可以获取  $x = \max(\text{take } k A), y = \min(\text{drop } k A)$ 。

如果  $x < y$ , 则  $A$  的前  $k$  个元素就是答案; 否则我们用  $x$  划分前  $k$  个元素, 用  $y$  划分剩余元素, 然后在子序列  $[a \leftarrow A, x < a < y]$  中递归查找前  $k'$  个元素, 其中  $k' = k - |[a \leftarrow A, a \leq x]|$ 。请实现这一算法, 并分析复杂度。

```

1: procedure TOPS( $k, A$ )
2:    $l \leftarrow 1$ 
3:    $u \leftarrow |A|$ 
4:   loop
5:      $i \leftarrow \text{MAX-AT}(A[l..k])$ 
6:      $j \leftarrow \text{MIN-AT}(A[k+1..u])$ 
7:     if  $A[i] < A[j]$  then
8:       break
9:     EXCHANGE  $A[l] \leftrightarrow A[j]$ 
10:    EXCHANGE  $A[k+1] \leftrightarrow A[i]$ 
11:     $l \leftarrow \text{PARTITION}(A, l, k)$ 
12:     $u \leftarrow \text{PARTITION}(A, k+1, u)$ 

```

平均情况下的复杂度是  $O(n)$ 。分析思路大致为: 每轮循环用线性时间找到最大、最小值  $i, j$ , 然后执行两轮线性时间的划分。如果划分平衡, 分摊下来每次平均丢弃一半元素。这样总时间为  $O(n + n/2 + n/4 \dots) = O(n)$ 。

14.1.3. 查找两个已序数组  $A$  和  $B$  的“简化中值”, 满足时间复杂度  $O(\lg(m+n))$ , 其中  $m = |A|, n = |B|$ , 分别为两个数组的长度。数组下标从 0 开始。简化中值定义为:  $\text{median}(A, B) = C[\lfloor \frac{m+n}{2} \rfloor]$ , 其中  $C = \text{merge}(A, B)$  为合并后的已序数组<sup>5</sup>。

我们给出两个方法。第一个方法在两个数组中分别进行二分查找。令  $l = 0, u = m$ , 我们猜测中值在  $A$  中的索引为:  $i = \lfloor \frac{l+u}{2} \rfloor$ 。根据简化中值的定义, 共有  $h = \lfloor \frac{m+n}{2} \rfloor$  个元素在它之前。其中,  $A$  中有  $i$  个元素在  $A[i]$  之前, 如果猜测正确, 则在  $B$  中应有  $j = h - i$  个元素在  $A[i]$  之前。检查不等式  $B[j] \leq A[i] \leq B[j+1]$ 。如果成立, 说明猜测的  $A[i]$  就是中值。否则根据偏大或偏小调整  $l, u$  重新二分查找。下面的例子程序实现了这一方法:

<sup>5</sup>简化定义与常见的统计学定义不同。后者对于含有  $n$  个元素升序序列  $x$  的中值定义为:

$$\text{median}(x) = \begin{cases} \text{odd}(n) : & x[\frac{n+1}{2}] \\ \text{even}(n) : & \frac{1}{2}(x[\frac{n}{2}] + x[\frac{n}{2} + 1]) \end{cases}$$

```

K median([K] a, [K] b) {
    if a == [] then return b[length(b) / 2]
    if b == [] then return a[length(a) / 2]
    Int i = medianOf(a, b)
    return if i == -1 then return median(b, a) else a[i]
}

Int medianOf([K] a, [K] b) {
    Int l = 0, u = length(a)
    while l < u {
        var i = (l + u) / 2
        var j = (length(a) + length(b)) / 2 - i
        if j < 1 or j ≥ len(b) {
            if (j == 0 and a[i] ≤ b[0]) or
                (j == len(b) and b[j - 1] ≤ a[i]) then return i
            if j ≥ len(b) then l = i + 1 else u = i
        } else {
            if b[j - 1] ≤ a[i] and a[i] ≤ b[j] then return i
            if a[i] < b[j - 1] then l = i + 1 else u = i
        }
    }
    return -1
}

```

第二种方法是利用一个通用的查找第  $k$  个元素的函数。假设  $m \geq n$  (否则交换  $A$ 、 $B$ )，如果任一数组为空，则返回另外一个的第  $k$  个元素。如果  $k = 1$ ，返回  $A[0]$ 、 $B[0]$  中较小的一个。否则猜测  $j = \min(k/2, n)$ ， $i = k - j$ ，然后比较  $A[i]$  和  $B[j]$ 。如果  $A[i] < B[j]$ ，我们丢弃所有  $A[i]$  前和  $B[j]$  后的元素，然后递归在剩余元素中寻找第  $k - i$  大的；否则丢弃所有  $B[j]$  前和  $A[i]$  后的，然后递归。

```

K median([K] xs, [K] ys) {
    Int n = length(xs), m = length(ys)
    return kth(xs, 0, n, ys, 0, m, (m + n) / 2 + 1)
}

K kth([K] xs, Int x0, Int x1, [K] ys, Int y0, Int y1, Int k) {
    if x1 - x0 < y1 - y0 then return kth(ys, y0, y1, xs, x0, x1, k)
    if x1 ≤ x0 then return ys[y0 + k - 1]
    if y1 ≤ y0 then return xs[x0 + k - 1]
    if k == 1 then return min(xs[x0], ys[y0])
    var j = min(k / 2, y1 - y0), i = k - j
    i = x0 + i, j = y0 + j
    if xs[i - 1] < ys[j - 1] then
        return kth(xs, i, x1, ys, y0, j, k - i + x0)
    else
        return kth(xs, x0, i, ys, j, y1, k - j + y0)
}

```

另外，我们不能这样定义简化中值： $m \in A \# B$ ，使得：

$$|[y \leftarrow A \# B, y < m]| - |[y \leftarrow A \# B, y > m]| = 0, \pm 1$$

即大于  $m$  的元素和小于  $m$  的元素同样多或差 1。考虑这样的反例:  $[0, 1, 2, 3, 3, 3, 3, 3, 5]$ , 即使用小于等于、大于等于也会有问题。

14.1.4. 消除递归, 通过更新搜索边界实现马鞍查找。

```

1: function SOLVE( $f, z$ )
2:    $p \leftarrow 0, q \leftarrow z$ 
3:    $S \leftarrow \phi$ 
4:   while  $p \leq z$  且  $q \geq 0$  do
5:      $z' \leftarrow f(p, q)$ 
6:     if  $z' < z$  then
7:        $p \leftarrow p + 1$ 
8:     else if  $z' > z$  then
9:        $q \leftarrow q - 1$ 
10:    else
11:       $S \leftarrow S \cup \{(p, q)\}$ 
12:       $p \leftarrow p + 1, q \leftarrow q - 1$ 
13:  return  $S$ 

```

14.1.5. 实现二维搜索算法并分析复杂度: 矩形搜索区域的左下角是最小值, 右上角是最大值。若搜索值  $z$  小于最小值或者大于最大值无解; 否则从中心划一个十字, 分割成 4 个小矩形, 然后递归搜索。

```

1: procedure SEARCH( $f, z, a, b, c, d$ )           ▷ ( $a, b$ ): 左下角 ( $c, d$ ): 右上角
2:   if  $z \leq f(a, b)$  或  $f(c, d) \geq z$  then
3:     if  $z = f(a, b)$  then
4:       record ( $a, b$ ) as a solution
5:     if  $z = f(c, d)$  then
6:       record ( $c, d$ ) as a solution
7:   return
8:    $p \leftarrow \lfloor \frac{a+c}{2} \rfloor$ 
9:    $q \leftarrow \lfloor \frac{b+d}{2} \rfloor$ 
10:  SEARCH( $f, z, a, q, p, d$ )
11:  SEARCH( $f, z, p, q, c, d$ )
12:  SEARCH( $f, z, a, b, p, q$ )
13:  SEARCH( $f, z, p, b, c, q$ )

```

我们可以利用主定理进行递归复杂度分析: 设在面积为  $A$  的矩形内搜索的时间为  $T(A)$ 。我们用  $O(1)$  时间检查  $z \leq f(a, b)$  或  $f(c, d) \geq z$  是否成立, 然后分割为 4 个面积相等的小矩形递归:  $T(A) = 4T(A/4) + O(c)$ 。根据主定理复杂度为  $O(A) = O(mn)$ , 和矩形的面积, 即边长的乘积成正比。这和 在矩形内逐点检查

效果相同。

### 答案 14.2

14.2.1. 扩展众数算法, 在  $A$  中寻找出现次数超过  $\lfloor n/k \rfloor$  的  $k$  个众数, 其中  $n = |A|$ 。提示: 每次删掉  $k$  个不同元素, 直到最后剩下的元素种类不足  $k$  个。如果某个元素是  $k$ -众数(多于  $\lfloor n/k \rfloor$  个), 则一定会剩下来。

我们建立一个字典  $Map : T \mapsto Int$ , 其中  $T$  是  $A$  中的元素类型。这个字典记录了候选者  $a$  的净胜票数。最开始字典为空  $\emptyset$ 。我们一边扫描  $A$  一边更新字典:  $foldr\ maj\ \emptyset\ A$ , 其中  $maj$  定义为:

$$maj\ a\ m = \begin{cases} a \in m : & m[a] \leftarrow m[a] + 1 \\ |m| < k : & m[a] \leftarrow 1 \\ \text{否则} : & filter\ (b \mapsto m[b] \neq 0)\ \{b \mapsto m[b] - 1 \mid b \in m\} \end{cases} \quad (14.21)$$

对于  $A$  中每个元素  $a$ , 如果  $a \notin m$  不在字典中, 并且字典中的候选者不足  $k$  个, 我们将  $a$  加入字典, 并记录为 1 票  $m[a] \leftarrow 1$ ; 如果  $a \in m$ , 我们将票数加一  $m[a] \leftarrow m[a] + 1$ ; 否则如果字典中已有  $k$  个候选, 我们把每个候选者的得票减 1, 如果票数为 0 则剔除。

最后, 我们还需把  $m$  中最后剩余的候选者验证一下, 看看它们的票数是否超过了  $n/k$ , 令  $m' = \{(a, 0) \mid a \in m\}$ , 然后再遍历一次  $A$ :  $foldr\ cnt\ m'\ A$ , 其中  $cnt$  定义为:

$$cnt\ a\ m' = \text{if } a \in m' \text{ then } m'[a] \leftarrow m'[a] + 1 \text{ else } m' \quad (14.22)$$

这样  $m'$  中记录了这些候选者的票数, 我们留下超过  $n/k$  的:  $keys\ (filter\ (> n/k)\ m')$ 。下面的例子程序实现了这一方法:

```
majorities k xs = verify $ foldr maj Map.empty xs where
  maj x m | x `Map.member` m = Map.adjust (1+) x m
          | Map.size m < k = Map.insert x 1 m
          | otherwise = Map.filter (≠0) $ Map.map (-1 +) m
  verify m = Map.keys $ Map.filter (> th) $ foldr cnt m' xs where
    m' = Map.map (const 0) m
    cnt x m = if x `Map.member` m then Map.adjust (1+) x m else m
    th = (length xs) `div` k
```

对应的迭代实现如下:

- 1: **function** MAJ( $k, A$ )
- 2:      $m \leftarrow \{\}$
- 3:     **for** each  $a$  in  $A$  **do**
- 4:         **if**  $a \in m$  **then**

```

5:         m[a] ← m[a] + 1
6:     else if |m| < k then
7:         m[a] ← 1
8:     else
9:         for each c in m do
10:            m[c] ← m[c] - 1
11:            if m[c] = 0 then
12:                REMOVE(c, m)
13: for each c in m do
14:     m[c] ← 0
15: for each a in A do
16:     if a ∈ m then
17:         m[a] ← m[a] + 1
18: r = [], n ← |A|
19: for each c in m do
20:     if m[c] >  $\frac{n}{k}$  then
21:         ADD(c, r)
22: return r

```

▷ 验证

### 答案 14.3

14.3.1. 修改最大子序列和的实现, 返回对应最大和的子序列。

如果除了最大子序列和, 还希望返回子序列, 我们可以在 fold 过程中使用两对值  $P_m$  和  $P$ , 每对值都包括子序列的和与子序列本身  $(S, L)$ 。

$$\text{max}_s = \text{fst} \cdot \text{foldr } f \ ((0, []), (0, []))$$

$$\text{其中: } f \ x \ (P_m, (S, L)) = (P'_m, P')$$

$$\text{在 } f \ \text{中: } P' = \text{max}((0, []), (x + S, x:L)), P'_m = \text{max}(P_m, P')$$

14.3.2. 本特利<sup>[2]</sup>给出了一个分而治之的方法求子数组最大和。复杂度为  $O(n \log n)$ 。思路是将列表在中点分成两份。我们可以递归地找出前半部分的最大和, 和后半部分的最大和, 和跨越中点部分的最大和。实现这一方法。

```

1: function MAX-SUM(A)
2:   if A = ϕ then
3:     return 0
4:   else if |A| = 1 then
5:     return MAX(0, A[1])
6:   else
7:     m ←  $\lfloor \frac{|A|}{2} \rfloor$ 

```

```

8:      a ← MAX-FROM(REVERSE(A[1...m]))
9:      b ← MAX-FROM(A[m + 1...|A|])
10:     c ← MAX-SUM(A[1...m])
11:     d ← MAX-SUM(A[m + 1...|A|])
12:     return MAX(a + b, c, d)

```

```

13: function MAX-FROM(A)
14:   sum ← 0, m ← 0
15:   for i ← 1 to |A| do
16:     sum ← sum + A[i]
17:     m ← MAX(m, sum)
18:   return m

```

复杂度存在递归关系  $T(n) = 2T(n/2) + O(n)$ 。利用主定理可知复杂度为  $O(n)$ 。

14.3.3. 在  $m \times n$  的二维整数矩阵中寻找子矩阵, 使得各元素相加后的和最大。我们从矩阵中的第一行开始, 每轮增加一行  $[M[1, *], M[2, *], \dots, M[i, *]]$ 。然后沿列把各行中的值加起来转换成一维向量:

$$V = \left[ \sum_{j=1}^i M[1, j], \sum_{j=1}^i M[2, j], \dots, \sum_{j=1}^i M[n, j] \right]$$

再使用本节中介绍的 *maxsum* 求出向量  $V$  中的最大和, 并记录全局最大和。

```

maxSum = maximum ◦ (map maxS) ◦ acc ◦ rows where
  rows = init ◦ tails           — exclude the empty row
  acc = concatMap (scanl1 (zipWith (+))) — accumulated sum along columns
  maxS = snd ◦ (foldl f (0, 0))      — max sum in a vector
  f (m, s) x = let m' = max (m + x) 0
                s' = max m' s in (m', s')

```

其中 *tails* 的定义见练习 1.12.2, *zipWith* 的定义见 1.7, *concatMap* 的定义见 6.5.1。 *scanl* 和 *foldl* 类似, 但它每次把结果都保存到一个列表中。 *scanl1* 是 *scanl* 的一个特殊情况, 初始值是列表的第一个元素。

$$\begin{aligned} \text{scanl1 } f \ [] &= [] \\ \text{scanl1 } f \ (x:xs) &= \text{scanl } f \ x \ xs \end{aligned}$$

其中:

$$\begin{aligned} \text{scanl } f \ q \ [] &= [q] \\ \text{scanl } f \ (x:xs) &= q : \text{scanl } f \ (f \ q \ x) \ xs \end{aligned}$$

对应的命令式例子程序如下:

```

K maxsum2([[K]] m) {
  Int n = length(m), k = length(m[0]) // number of row, col
  K maxs = 0                          // max so far

```

```

for i = 0 to n - 1 {
  xs = [0] * k
  for j = i to n - 1 {
    xs = [x + y for (x, y) in zip(xs, m[j])]
    maxs = max(maxs, maxsum1(xs))
  }
}
return maxs
}

K maxsum1([K] xs) {
  K s = 0 // max so far
  K m = 0 // max end here
  for x in xs {
    m = max(m + x, 0)
    s = max(m, s)
  }
  return s
}

```

### 答案 15.1

15.1.1. 使用栈来找出迷宫问题的所有解。

```

dfsSolveAll m from to = map reverse $ solve [[from]] [] where
  solve [] ss = ss
  solve (c@(p:path):cs) ss
    | p == to = solve cs (c:ss) — find a solution, go on search
    | otherwise = let os = filter (`notElem` path) (adjacent p) in
      if os == [] then solve cs ss
      else solve ((map (:c) os) ++ cs) ss
  adjacent (x, y) = [(x', y') |
    (x', y') ← [(x-1, y), (x+1, y), (x, y-1), (x, y+1)],
    inRange (bounds m) (x', y'), m ! (x', y') == 0]

```

对应的命令式程序

```

[[Int, Int]] solve([[Int]] m, (Int, Int) src, (Int, Int) dst) {
  [[Int, Int]] stack = [[src]]
  [[Int, Int]] s = []
  while stack ≠ [] {
    path = stack.pop()
    if last(path) == dst {
      s.append(path)
    } else {
      for p in adjacent(m, last(path)) {
        if not p in path then stack.append(path + [p])
      }
    }
  }
  return s
}

```

```

[(Int, Int)] adjacent([[Int, Int]] m, (Int x, Int y)) {
  [(Int, Int)] ps = []
  for (dx, dy) in [(0, 1), (0, -1), (1, 0), (-1, 0)] {
    Int x1 = x + dx, y1 = y + dy
    if 0 ≤ x1 < len(m[0]) and 0 ≤ y1 < len(m)
      and m[y][x] == 0 then ps.append((x1, y1))
  }
  return ps
}

```

### 答案 15.2

15.2.1. 改进八皇后的算法,使得它可以解决  $n$  皇后问题。

所谓  $n$  皇后问题,是在  $n \times n$  棋盘上摆放  $n$  个皇后,我们可以将例子程序中的 8 换成  $n$  传入即可。下表列出了前几个皇后数目和解:

| $n$ | 解                                                |
|-----|--------------------------------------------------|
| 1   | [1]                                              |
| 2   | 无解                                               |
| 3   | 无解                                               |
| 4   | [2,4,1,3],[3,1,4,2]                              |
| 5   | [2,4,1,3,5],[3,1,4,2,5],[1,3,5,2,4], ... 共 10 个解 |

15.2.2. 八皇后问题存在 92 个不同的解。对于任何一个解,将其旋转  $\pm 90^\circ$  也都是八皇后问题的解。并且翻转也是解。真正不同的解只有 12 个。修改八皇后的程序,找出这 12 个不同的解。

八皇后解的对称性,本质上是正方形棋盘的对称性。群是解决对称的有力武器。正方形的对称性被二面体群  $D_8$  定义,包括如下 8 个置换:恒等置换  $id$ ,围绕中心旋转  $90^\circ$ 、 $180^\circ$ 、 $270^\circ$ 。水平、垂直翻转,沿两个对角线翻转。位于位置  $(i, j)$  的皇后在这 8 个置换下变为:

| 置换                                     | 位置                                             |
|----------------------------------------|------------------------------------------------|
| $id$                                   | $(i, j)$                                       |
| Y 翻转, X 翻转                             | $(9 - i, j)$ 、 $(i, 9 - j)$                    |
| 2 个对角线翻转                               | $(j, i)$ 、 $(9 - j, 9 - i)$                    |
| $90^\circ$ 、 $180^\circ$ 、 $270^\circ$ | $(9 - j, i)$ 、 $(9 - i, 9 - j)$ 、 $(j, 9 - i)$ |

每找到一个解,我们就用这 8 种置换产生 8 个解,并认为它们是等价的。我们用一个集合来维护 12 个本质不同的解:

```
import Data.List ((\), sortOn)
```



```

import Data.Set (Set, empty, insert, notMember, size)
import Data.Tuple (swap)

d8 = [id,
      reverse, map (9 - ),           — reflect Y, X
      trans swap, trans (λ(i, j) → (9 - j, 9 - i)), — reflect AC, BD
      trans (λ(i, j) → (9 - j, i)), — 90
      trans (λ(i, j) → (9 - i, 9 - j)), — 180
      trans (λ(i, j) → (j, 9 - i))] — 270

where
  trans f xs = snd $ unzip $ sortOn fst $ map f $ zip [1..8] xs

uniqueSolve = dfs [[]] (empty :: Set [Int]) where
  dfs [] s = s
  dfs (c:cs) s
    | length c == 8 = dfs cs (uniqueAdd c s)
    | otherwise = dfs (([x:c | x ← [1..8] \\ c,
                        not (attack x c)] ++ cs) s
  uniqueAdd c s = if all (`notMember` s) [f c | f ← d8]
                 then insert c s else s
  attack x cs = let y = 1 + length cs in
                any (λ(c, r) → abs(x - c) == abs(y - r)) $ zip (reverse cs) [1..]

```

这 12 个解为:

```

[3,6,4,1,8,5,7,2], [3,6,8,1,4,7,5,2], [4,1,5,8,6,3,7,2], [4,2,7,3,6,8,5,1]
[4,6,8,3,1,7,5,2], [4,7,1,8,5,2,6,3], [5,2,4,7,3,8,6,1], [5,3,8,4,7,1,6,2]
[5,7,1,3,8,6,4,2], [5,7,4,1,3,8,6,2], [6,2,7,1,4,8,5,3], [6,4,7,1,8,2,5,3]

```

### 答案 15.3

- 15.3.1. 修改跳跃青蛙问题的函数式解法,使得它可以解决每侧  $n$  只青蛙的情况。我们只要利用  $n$  产生起始、结束状态即可:

```

solve n = dfs [[start]] [] where
  dfs [] s = s
  dfs (c:cs) s
    | head c == end = dfs cs (reverse c:s)
    | otherwise = dfs ((map (:c) $ moves $ head c) ++ cs) s
  start = replicate n (-1) ++ [0] ++ replicate n 1
  end = reverse start

```

### 答案 15.4

- 15.4.1. 改进扩展欧几里得算法,寻找线性组合使得  $|x| + |y|$  最小,以解决倒水问题。

```

import Data.List
import Data.Function (on)

— Extended Euclidean Algorithm
gcmex a 0 = (a, 1, 0)

```

```

gcmex a b = (g, y', x' - y' * (a `div` b)) where
  (g, x', y') = gcmex b (a `mod` b)

— Solve the linear Diophantine equation ax + by = c
solve a b c | c `mod` g ≠ 0 = (0, 0, 0, 0) — no solution
             | otherwise = (x1, u, y1, v)

where
  (g, x0, y0) = gcmex a b
  (x1, y1) = (x0 * c `div` g, y0 * c `div` g)
  (u, v) = (b `div` g, a `div` g)

— Minimize |x| + |y|
jars a b c = (x, y) where
  (x1, u, y1, v) = solve a b c
  x = x1 - k * u
  y = y1 + k * v
  k = minimumBy (compare `on` (\i → abs (x1 - i * u) +
                                abs (y1 + i * v))) [-m..m]

  m = max (abs x1 `div` u) (abs y1 `div` v)

— Populate the steps
water a b c = if x > 0 then pour a x b y
              else map swap $ pour b y a x

where
  (x, y) = jars a b c

— Pour from a to b, fill a for x times, and empty b for y times.
pour a x b y = steps x y [(0, 0)]

where
  steps 0 0 ps = reverse ps
  steps x y ps@((a', b'):_ )
    | a' == 0 = steps (x - 1) y ((a, b'):ps) — fill a
    | b' == b = steps x (y + 1) ((a', 0):ps) — empty b
    | otherwise = steps x y ((max (a' + b' - b) 0,
                                min (a' + b') b):ps) — a to b

```

更多内容,请参见《同构——编程中的数学》第二章,2.2.3节。

## 答案 15.5

### 15.5.1. 按层遍历一棵二叉树。

按广度优先顺序访问各个节点,使用 12.5 中的序列作为队列:

```

data Tr a = E | Br (Tr a) a (Tr a)

toList t = bfs (Q.singleton t) where
  bfs Q.Empty = []
  bfs (E :<| q) = bfs q
  bfs ((Br l k r) :<| q) = k : bfs (q |> l |> r)

```

### 15.5.2. 康威滑动趣题。

我们用 0 到 7 这 8 个数字的排列代表棋子,其中 0 代表空位。空位前、后的棋子可以滑入空位。我们将其表示为 0 在排列中左右移动。向右移动到尾部时,继续向右则回到头部,向左到头部时,继续向左回到尾部。特别地,如果空位 0 在排列头部,它可以和第 5 位置上的棋子交换,反之亦可。康威滑动游戏的解很长,需要 12948 步移动才能反转棋子。其中最后几步如下:

```
start = [0..7]
end = 0:[7,6..1]

solve1 = dfs [[start]] where
  dfs [] = []
  dfs (c:cs)
    | head c == end = reverse c
    | otherwise = dfs ((map (:c) $ moves c) ++ cs)

moves (s:visited) = filter (`notElem` visited) [fwd s, bk s, cut s]
  where
    fwd xs = case break (0 ==) xs of
      (as, 0:b:bs) → as ++ (b:0:bs)
      (a:as, [0]) → 0:as ++ [a]
    bk xs = case break (0 ==) xs of
      ([], 0:bs) → bs ++ [0]
      (as, 0:bs) → (init as) ++ (0 : last as : bs)
    cut xs = case splitAt 4 xs of
      ((0:as), (x:bs)) → (x:as) ++ (0:bs)
      ((x:as), (0:bs)) → (0:as) ++ (x:bs)
      _ → xs
```

```
...
[1,0,7,6,5,4,3,2],[1,7,0,6,5,4,3,2],[1,7,6,0,5,4,3,2],[1,7,6,5,0,4,3,2],
[1,7,6,5,4,0,3,2],[1,7,6,5,4,3,0,2],[1,7,6,5,4,3,2,0],[0,7,6,5,4,3,2,1]
```

## 答案 15.6

15.6.1. 实现命令式的哈夫曼码表生成算法。

```
data Node<T> {
  Optional<T> c = Nothing
  Int w
  Node<T> left = null, right = null

  Bool isLeaf() = (left == null and right == null)
}

Node<T> merge(Node<T> a, Node<T> b) = Node(Nothing, a.w + b.w, a, b)

Bool (<)(Node<T> a, Node<T> b) = (a.w < b.w)

Node<T> huffman([Node<T>] ts) {
  while length(ts) > 1 {
    Int n = length(ts)
```

```

    for Int i = n - 3 down to i 0 {
        if ts[i] < max(ts[n-1], ts[n-2]) {
            Int j = if ts[n-1] < ts[n-2] then n - 2 else n - 1
            swap(ts[i], ts[j])
        }
    }
    ts[n-2] = merge(ts[n-1], ts[n-2])
    ts.popLast()
}
return ts[0]
}

Map<T, [T]> codeTab(Node<T> t, [T] bits = [], Map<T, [T]> codes = {}) {
    if t.isLeaf() {
        codes[t.c] = bits
    } else {
        codeTab(t.left, bits + [0], codes)
        codeTab(t.right, bits + [1], codes)
    }
    return codes
}
}

```

### 答案 15.7

15.7.1. 使用堆来构造哈夫曼树: 不断从堆顶取出两颗子树, 合并后放回堆中。

$$\text{Huffman } H = \begin{cases} H = \emptyset: & \emptyset \\ |H| = 1: & \text{pop } H \\ \text{否则}: & \text{Huffman } (\text{push } (\text{merge } t_a \ t_b) \ H'') \end{cases}$$

其中:  $(t_a, H') = \text{pop } H$ ,  $(t_b, H'') = \text{pop } H'$

```

1: function HUFFMAN(H)
2:   while |H| > 1 do
3:     ta ← POP(H)
4:     tb ← POP(H)
5:     PUSH(H, MERGE(ta, tb))
6:   return POP(H)

```

15.7.2. 如果字符已按权重排序成列表  $A$ , 存在一个线性时间的构造哈夫曼树的方法: 用队列  $Q$  保存合并结果。不断从  $Q$  和  $A$  头部取出较小的树, 合并后入队。处理完列表中所有树后, 队列中将只剩下一棵树。即最终的哈夫曼树。请实现这一方法。

$\text{Huffman } (t:ts) = \text{build } (t, (ts, \emptyset))$ , 其中:

$$\begin{aligned}
 \text{build } (t, ([], \emptyset)) &= t \\
 \text{build } (t, h) &= \text{build } (\text{extract } (ts, \text{push } (\text{merge } t \ t') \ q))
 \end{aligned}$$

其中:  $(t', (ts, q)) = extract\ h$

$$extract\ (t:ts, \emptyset) = (t, (ts, \emptyset))$$

$$extract\ ([], q) = (t, ([], q'), \text{其中: } (t, q') = pop\ q$$

$$extract\ (t:ts, q) = \begin{cases} t' < t : (t', (t:ts, q')), \text{其中: } (t', q') = pop\ q \\ t < t' : (t, (ts, q)) \end{cases}$$

15.7.3. 给定哈夫曼树  $T$ , 用左侧叠加实现哈夫曼解码。

$decode = snd \circ (foldl\ lookup\ (T, []))$ , 其中:

$$lookup\ ((w, c), cs)\ b = (T, c:cs)$$

$$lookup\ ((w, l, r), cs)\ b = \text{if } b = 0 \text{ then } (l, cs) \text{ else } (r, cs)$$

### 答案 15.8

15.8.1. 使用叠加操作构造最长公共子序列的最优解表格

```
import Data.Sequence (Seq, singleton, fromList, index, (>))

lcs xs ys = construct $ foldl f (singleton $ fromList $ replicate (n+1) 0)
              (zip [1..] xs) where
  (m, n) = (length xs, length ys)
  f tab (i, x) = tab |> (foldl longer (singleton 0) (zip [1..] ys)) where
    longer r (j, y) = r |> if x == y
      then 1 + (tab `index` (i-1) `index` (j-1))
      else max (tab `index` (i-1) `index` j) (r `index` (j-1))
  construct tab = get (reverse xs, m) (reverse ys, n) where
    get ([], 0) ([], 0) = []
    get ((x:xs), i) ((y:ys), j)
      | x == y = get (xs, i-1) (ys, j-1) # [x]
      | (tab `index` (i-1) `index` j) > (tab `index` i `index` (j-1)) =
        get (xs, i-1) ((y:ys), j)
      | otherwise = get ((x:xs), i) (ys, j-1)
```

### 答案 15.9

15.9.1. 对最长公共子序列问题, 另一种自底向上的解法是子表格中记录子序列的长度和“方向”。“N”代表向北, “W”代表向西, “NW”代表向西北。这些方向指示我们如何构建最终的结果。我们从表格的右下角开始, 如果值为“NW”, 我们就沿着对角线移动到左上方的格子; 如果值为“N”, 就垂直移动到上方的格子; 如果为“W”, 就水平移动到左侧的格子。编程实现这一算法。

```
data DIR = N | W | NW

[K] lcs([K] xs, [K] ys) {
  Int m = length(x), n = length(ys)
```

```

[[Int, DIR]] c = [[(0, null)] * (n + 1)] * (m + 1)
for i = 1 to m {
  for j = 1 to n {
    if xs[i-1] == ys[j-1] {
      c[i][j] = (fst(c[i-1][j-1]) + 1, DIR.NW)
    } else {
      c[i][j] = if fst(c[i-1][j]) > fst(c[i][j-1])
        then (fst(c[i-1][j]), DIR.N)
        else (fst(c[i][j-1]), DIR.W)
    }
  }
}
return rebuild(c, xs, ys)
}

[K] rebuild([[Int, DIR]] c, [K] xs, [K] ys) {
  [K] r = []
  Int m = length(xs), n = length(ys)
  while m > 0 and n > 0 {
    DIR d = snd(c[m][n])
    if d == DIR.NW {
      r.append(xs[m - 1]) // or ys[n - 1]
      m = m - 1, n = n - 1
    } else if d == DIR.N {
      m = m - 1
    } else if d == DIR.W {
      n = n - 1
    }
  }
  return reverse(r)
}

```

15.9.2. 对于子集和的上下限, 一定有  $l \leq 0 \leq u$  成立么? 上下限能进一步缩小么?

若要求非空子集(空集的和为 0), 则不等式  $l \leq 0 \leq u$  不一定成立。考虑全部是正数的集合  $X$ , 下限不可能小于 0, 并且  $l = \min(X)$ 。反之全部是负数的集合上限不可能大于 0, 且  $u = \max(X)$ 。

15.9.3. 求两个字符串间的编辑距离。

```

Int lev([K] s, [K] t) {
  [[Int]] d = [[0]*n]*m //d[i][j]: distance between s[:i] and t[:j]
  for Int i = 0 to length(s) {
    d[i][0] = i //drop all chars of source prefix gives []
  }
  for Int j = 0 to length(t) {
    d[0][j] = j //insert all chars of target prefix to []
  }
  for Int j = 1 to length(t) {
    for i = 1 to length(m) {
      c = if s[i-1] == t[j-1] then 0 else 1
      d[i][j] = min([d[i-1][j] + 1, //deletion
                    d[i][j-1] + 1, //insertion

```

```
        d[i-1][j-1] + c) //substitution
    }
}
return d[length(s)][length(t)]
}
```





## 参考文献

- [1] Richard Bird. “Pearls of functional algorithm design”. Cambridge University Press; 1 edition (November 1, 2010). ISBN-10: 0521513383. pp1 - pp6.
- [2] Jon Bentley. “Programming Pearls(2nd Edition)”. Addison-Wesley Professional; 2 edition (October 7, 1999). ISBN-13: 978-0201657883 (中文版:《编程珠玑》)
- [3] Chris Okasaki. “Purely Functional Data Structures”. Cambridge university press, (July 1, 1999), ISBN-13: 978-0521663502
- [4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein. “Introduction to Algorithms, Second Edition”. The MIT Press, 2001. ISBN: 0262032937. (中文版:《算法导论》)
- [5] Chris Okasaki. “Ten Years of Purely Functional Data Structures”. <http://okasaki.blogspot.com/2008/02/ten-years-of-purely-functional-data.html>
- [6] SGI. “Standard Template Library Programmer’s Guide”. <http://www.sgi.com/tech/stl/>
- [7] Wikipedia. “Fold(high-order function)”. [https://en.wikipedia.org/wiki/Fold\\_\(higher-order\\_function\)](https://en.wikipedia.org/wiki/Fold_(higher-order_function))
- [8] Wikipedia. “Function Composition”. [https://en.wikipedia.org/wiki/Function\\_composition](https://en.wikipedia.org/wiki/Function_composition)
- [9] Wikipedia. “Partial application”. [https://en.wikipedia.org/wiki/Partial\\_application](https://en.wikipedia.org/wiki/Partial_application)
- [10] Miran Lipovaca. “Learn You a Haskell for Great Good! A Beginner’s Guide”. No Starch Press; 1 edition April 2011, 400 pp. ISBN: 978-1-59327-283-8
- [11] Wikipedia. “Bubble sort”. [https://en.wikipedia.org/wiki/Bubble\\_sort](https://en.wikipedia.org/wiki/Bubble_sort)
- [12] Donald E. Knuth. “The Art of Computer Programming, Volume 3: Sorting and Searching (2nd Edition)”. Addison-Wesley Professional; 2 edition (May 4, 1998) ISBN-10: 0201896850 ISBN-13: 978-0201896855

- [13] Chris Okasaki. “FUNCTIONAL PEARLS Red-Black Trees in a Functional Setting”. J. Functional Programming. 1998
- [14] Wikipedia. “Red-black tree”. [https://en.wikipedia.org/wiki/Red-black\\_tree](https://en.wikipedia.org/wiki/Red-black_tree)
- [15] Lyn Turbak. “Red-Black Trees”. <http://cs.wellesley.edu/~cs231/fall01/red-black.pdf> Nov. 2, 2001.
- [16] Rosetta Code. “Pattern matching”. [http://rosettacode.org/wiki/Pattern\\_matching](http://rosettacode.org/wiki/Pattern_matching)
- [17] Hackage. “Data.Tree.AVL”. <http://hackage.haskell.org/packages/archive/AvlTree/4.2/doc/html/Data-Tree-AVL.html>
- [18] Wikipedia. “AVL tree”. [https://en.wikipedia.org/wiki/AVL\\_tree](https://en.wikipedia.org/wiki/AVL_tree)
- [19] Guy Cousinear, Michel Mauny. “The Functional Approach to Programming”. Cambridge University Press; English Ed edition (October 29, 1998). ISBN-13: 978-0521576819
- [20] Pavel Grafov. “Implementation of an AVL tree in Python”. <http://github.com/pgrafov/python-avl-tree>
- [21] Chris Okasaki and Andrew Gill. “Fast Mergeable Integer Maps”. Workshop on ML, September 1998, pages 77-86.
- [22] D.R. Morrison, “PATRICIA – Practical Algorithm To Retrieve Information Coded In Alphanumeric”, Journal of the ACM, 15(4), October 1968, pages 514-534.
- [23] Wikipedia. “Suffix Tree”. [https://en.wikipedia.org/wiki/Suffix\\_tree](https://en.wikipedia.org/wiki/Suffix_tree)
- [24] Wikipedia. “Trie”. <https://en.wikipedia.org/wiki/Trie>
- [25] Wikipedia. “T9 (predictive text)”. [https://en.wikipedia.org/wiki/T9\\_\(predictive\\_text\)](https://en.wikipedia.org/wiki/T9_(predictive_text))
- [26] Wikipedia. “Predictive text”. [https://en.wikipedia.org/wiki/Predictive\\_text](https://en.wikipedia.org/wiki/Predictive_text)
- [27] Esko Ukkonen. “On-line construction of suffix trees”. Algorithmica 14 (3): 249-260. doi:10.1007/BF01206331. <http://www.cs.helsinki.fi/u/ukkonen/SuffixT1withFigs.pdf>
- [28] Weiner, P. “Linear pattern matching algorithms”, 14th Annual IEEE Symposium on Switching and Automata Theory, pp. 1-11, doi:10.1109/SWAT.1973.13

- [29] Esko Ukkonen. “Suffix tree and suffix array techniques for pattern analysis in strings”. <http://www.cs.helsinki.fi/u/ukkonen/Erice2005.ppt>
- [30] Suffix Tree (Java). [http://en.literateprograms.org/Suffix\\_tree\\_\(Java\)](http://en.literateprograms.org/Suffix_tree_(Java))
- [31] Robert Giegerich and Stefan Kurtz. “From Ukkonen to McCreight and Weiner: A Unifying View of Linear-Time Suffix Tree Construction”. *Science of Computer Programming* 25(2-3):187-218, 1995. <http://citeseer.ist.psu.edu/giegerich95comparison.html>
- [32] Robert Giegerich and Stefan Kurtz. “A Comparison of Imperative and Purely Functional Suffix Tree Constructions”. *Algorithmica* 19 (3): 331–353. doi:10.1007/PL00009177. <http://www.zbh.uni-hamburg.de/pubs/pdf/GieKur1997.pdf>
- [33] Bryan O’Sullivan. “suffixtree: Efficient, lazy suffix tree implementation”. <http://hackage.haskell.org/package/suffixtree>
- [34] Danny. <http://hkn.eecs.berkeley.edu/~dyoo/plt/suffixtree/>
- [35] Dan Gusfield. “Algorithms on Strings, Trees and Sequences Computer Science and Computational Biology”. Cambridge University Press; 1 edition (May 28, 1997) ISBN: 9780521585194
- [36] Lloyd Allison. “Suffix Trees”. <http://www.allisons.org/ll/AlgDS/Tree/Suffix/>
- [37] Esko Ukkonen. “Suffix tree and suffix array techniques for pattern analysis in strings”. <http://www.cs.helsinki.fi/u/ukkonen/Erice2005.ppt>
- [38] Esko Ukkonen “Approximate string-matching over suffix trees”. *Proc. CPM 93. Lecture Notes in Computer Science* 684, pp. 228-242, Springer 1993. <http://www.cs.helsinki.fi/u/ukkonen/cpm931.ps>
- [39] Wikipeda. “B-tree”. <https://en.wikipedia.org/wiki/B-tree>
- [40] Wikipedia. “Heap (data structure)”. [https://en.wikipedia.org/wiki/Heap\\_\(data\\_structure\)](https://en.wikipedia.org/wiki/Heap_(data_structure))
- [41] Wikipedia. “Heapsort”. <https://en.wikipedia.org/wiki/Heapsort>
- [42] Rosetta Code. “Sorting algorithms/Heapsort”. [http://rosettacode.org/wiki/Sorting\\_algorithms/Heapsort](http://rosettacode.org/wiki/Sorting_algorithms/Heapsort)
- [43] Wikipedia. “Leftist Tree”. [https://en.wikipedia.org/wiki/Leftist\\_tree](https://en.wikipedia.org/wiki/Leftist_tree)
- [44] Bruno R. Preiss. *Data Structures and Algorithms with Object-Oriented Design Patterns in Java*. <http://www.brpreiss.com/books/opus5/index.html>

- [45] Donald E. Knuth. “The Art of Computer Programming. Volume 3: Sorting and Searching.”. Addison-Wesley Professional; 2nd Edition (October 15, 1998). ISBN-13: 978-0201485417. Section 5.2.3 and 6.2.3
- [46] Wikipedia. “Skew heap”. [https://en.wikipedia.org/wiki/Skew\\_heap](https://en.wikipedia.org/wiki/Skew_heap)
- [47] Sleator, Daniel Dominic; Jarjan, Robert Endre. “Self-adjusting heaps” SIAM Journal on Computing 15(1):52-69. doi:10.1137/0215004 ISSN 00975397 (1986)
- [48] Wikipedia. “Splay tree”. [https://en.wikipedia.org/wiki/Splay\\_tree](https://en.wikipedia.org/wiki/Splay_tree)
- [49] Sleator, Daniel D.; Tarjan, Robert E. (1985), “Self-Adjusting Binary Search Trees”, Journal of the ACM 32(3):652 - 686, doi: 10.1145/3828.3835
- [50] NIST, “binary heap”. <http://xw2k.nist.gov/dads//HTML/binaryheap.html>
- [51] Donald E. Knuth. “The Art of Computer Programming, Volume 3: Sorting and Searching (2nd Edition)”. Addison-Wesley Professional; 2 edition (May 4, 1998) ISBN-10: 0201896850 ISBN-13: 978-0201896855
- [52] Wikipedia. “Strict weak order”. [https://en.wikipedia.org/wiki/Strict\\_weak\\_order](https://en.wikipedia.org/wiki/Strict_weak_order)
- [53] Wikipedia. “FIFA world cup”. [https://en.wikipedia.org/wiki/FIFA\\_World\\_Cup](https://en.wikipedia.org/wiki/FIFA_World_Cup)
- [54] Wikipedia. “K-ary tree”. [https://en.wikipedia.org/wiki/K-ary\\_tree](https://en.wikipedia.org/wiki/K-ary_tree)
- [55] Wikipedia, “Pascal’s triangle”. [https://en.wikipedia.org/wiki/Pascal's\\_triangle](https://en.wikipedia.org/wiki/Pascal's_triangle)
- [56] Hackage. “An alternate implementation of a priority queue based on a Fibonacci heap.”, <http://hackage.haskell.org/packages/archive/pqueue-mtl/1.0.7/doc/html/src/Data-Queue-FibQueue.html>
- [57] Chris Okasaki. “Fibonacci Heaps.” <http://darcs.haskell.org/nofib/gc/fibheaps/orig>
- [58] Michael L. Fredman, Robert Sedgwick, Daniel D. Sleator, and Robert E. Tarjan. “The Pairing Heap: A New Form of Self-Adjusting Heap” Algorithmica (1986) 1: 111-129.
- [59] Maged M. Michael and Michael L. Scott. “Simple, Fast, and Practical Non-Blocking and Blocking Concurrent Queue Algorithms”. <http://www.cs.rochester.edu/research/synchronization/pseudocode/queues.html>
- [60] Herb Sutter. “Writing a Generalized Concurrent Queue”. Dr. Dobb’s Oct 29, 2008. <http://drdobbs.com/cpp/211601363?pgno=1>

- [61] Wikipedia. “Tail-call”. [https://en.wikipedia.org/wiki/Tail\\_call](https://en.wikipedia.org/wiki/Tail_call)
- [62] Wikipedia. “Recursion (computer science)”. [https://en.wikipedia.org/wiki/Recursion\\_\(computer\\_science\)#Tail-recursive\\_functions](https://en.wikipedia.org/wiki/Recursion_(computer_science)#Tail-recursive_functions)
- [63] Harold Abelson, Gerald Jay Sussman, Julie Sussman. “Structure and Interpretation of Computer Programs, 2nd Edition”. MIT Press, 1996, ISBN 0-262-51087-1 (中文版: 裘宗燕 译《计算机程序的构造和解释》)
- [64] Chris Okasaki. “Purely Functional Random-Access Lists”. Functional Programming Languages and Computer Architecture, June 1995, pages 86-95.
- [65] Ralf Hinze and Ross Paterson. “Finger Trees: A Simple General-purpose Data Structure,” in Journal of Functional Programming 16:2 (2006), pages 197-217. <http://www.soi.city.ac.uk/~ross/papers/FingerTree.html>
- [66] Guibas, L. J., McCreight, E. M., Plass, M. F., Roberts, J. R. (1977), “A new representation for linear lists”. Conference Record of the Ninth Annual ACM Symposium on Theory of Computing, pp. 49-60.
- [67] Generic finger-tree structure. <http://hackage.haskell.org/packages/archive/fingertree/0.0/doc/html/Data-FingerTree.html>
- [68] Wikipedia. “Move-to-front transform”. [https://en.wikipedia.org/wiki/Move-to-front\\_transform](https://en.wikipedia.org/wiki/Move-to-front_transform)
- [69] Robert Sedgwick. “Implementing quick sort programs”. Communication of ACM. Volume 21, Number 10. 1978. pp.847 - 857.
- [70] Jon Bentley, Douglas McIlroy. “Engineering a sort function”. Software Practice and experience VOL. 23(11), 1249-1265 1993.
- [71] Robert Sedgwick, Jon Bentley. “Quicksort is optimal”. <http://www.cs.princeton.edu/~rs/talks/QuicksortIsOptimal.pdf>
- [72] Fethi Rabhi, Guy Lapalme. “Algorithms: a functional programming approach”. Second edition. Addison-Wesley, 1999. ISBN: 0201-59604-0
- [73] Simon Peyton Jones. “The Implementation of functional programming languages”. Prentice-Hall International, 1987. ISBN: 0-13-453333-X
- [74] Jyrki Katajainen, Tomi Pasanen, Jukka Teuhola. “Practical in-place mergesort”. Nordic Journal of Computing, 1996.
- [75] José Bacelar Almeida and Jorge Sousa Pinto. “Deriving Sorting Algorithms”. Technical report, Data structures and Algorithms. 2008.

- [76] Cole, Richard (August 1988). "Parallel merge sort". *SIAM J. Comput.* 17 (4): 770-785. doi:10.1137/0217049. (August 1988)
- [77] Powers, David M. W. "Parallelized Quicksort and Radixsort with Optimal Speedup", Proceedings of International Conference on Parallel Computing Technologies. Novosibirsk. 1991.
- [78] Wikipedia. "Quicksort". <https://en.wikipedia.org/wiki/Quicksort>
- [79] Wikipedia. "Total order". [http://en.wikipedia.org/wiki/Total\\_order](http://en.wikipedia.org/wiki/Total_order)
- [80] Wikipedia. "Harmonic series (mathematics)". [https://en.wikipedia.org/wiki/Harmonic\\_series\\_\(mathematics\)](https://en.wikipedia.org/wiki/Harmonic_series_(mathematics))
- [81] M. Blum, R.W. Floyd, V. Pratt, R. Rivest and R. Tarjan, "Time bounds for selection," *J. Comput. System Sci.* 7 (1973) 448-461.
- [82] Edsger W. Dijkstra. "The saddleback search". EWD-934. 1985. <https://www.cs.utexas.edu/users/EWD/ewd09xx/EWD934.PDF>.
- [83] Robert Boyer, and Strother Moore. "MJRTY - A Fast Majority Vote Algorithm". *Automated Reasoning: Essays in Honor of Woody Bledsoe*, Automated Reasoning Series, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991, pp. 105-117.
- [84] Cormode, Graham; S. Muthukrishnan (2004). "An Improved Data Stream Summary: The Count-Min Sketch and its Applications". *J. Algorithms* 55: 29-38.
- [85] Knuth Donald, Morris James H., jr, Pratt Vaughan. "Fast pattern matching in strings". *SIAM Journal on Computing* 6 (2): 323-350. 1977.
- [86] Robert Boyer, Strother Moore. "A Fast String Searching Algorithm". *Comm. ACM* (New York, NY, USA: Association for Computing Machinery) 20 (10): 762-772. 1977
- [87] R. N. Horspool. "Practical fast searching in strings". *Software - Practice & Experience* 10 (6): 501-506. 1980.
- [88] Wikipedia. "Boyer-Moore string search algorithm". [https://en.wikipedia.org/wiki/Boyer-Moore\\_string\\_search\\_algorithm](https://en.wikipedia.org/wiki/Boyer-Moore_string_search_algorithm)
- [89] Wikipedia. "Eight queens puzzle". [https://en.wikipedia.org/wiki/Eight\\_queens\\_puzzle](https://en.wikipedia.org/wiki/Eight_queens_puzzle)
- [90] George Pólya. "How to solve it: A new aspect of mathematical method". Princeton University Press(April 25, 2004). ISBN-13: 978-0691119663

- [91] Wikipedia. “David A. Huffman”. [https://en.wikipedia.org/wiki/David\\_A.\\_Huffman](https://en.wikipedia.org/wiki/David_A._Huffman)
- [92] Andrei Alexandrescu. “Modern C++ design: Generic Programming and Design Patterns Applied”. Addison Wesley February 01, 2001, ISBN 0-201-70431-5
- [93] Benjamin C. Pierce. “Types and Programming Languages”. The MIT Press, 2002. ISBN:0262162091
- [94] Joe Armstrong. “Programming Erlang: Software for a Concurrent World”. Pragmatic Bookshelf; 1 edition (July 18, 2007). ISBN-13: 978-1934356005
- [95] SGI. “transform”. <http://www.sgi.com/tech/stl/transform.html>
- [96] ACM/ICPC. “The drunk jailer.” Peking University judge online for ACM/ICPC. <http://poj.org/problem?id=1218>.
- [97] Haskell wiki. “Haskell programming tips”. 4.4 Choose the appropriate fold. [http://www.haskell.org/haskellwiki/Haskell\\_programming\\_tips](http://www.haskell.org/haskellwiki/Haskell_programming_tips)
- [98] Wikipedia. “Dot product”. [https://en.wikipedia.org/wiki/Dot\\_product](https://en.wikipedia.org/wiki/Dot_product)
- [99] Xinyu LIU. “Isomorphism - mathematics of programming”. <https://github.com/liuxinyu95/unplugged>





# 索引

- [k 选择, 259](#)
- [AVL 树, 69](#)
  - [命令式插入, 74](#)
  - [定义, 69](#)
  - [平衡调整, 72](#)
  - [插入, 71](#)
  - [检验, 74](#)
- [BFS, 299](#)
- [Boyer-Moore 众数问题, 270](#)
- [B 树, 107](#)
  - [删除, 119](#)
  - [插入, 109](#)
  - [查找, 118](#)
- [DFS, 279](#)
- [KMP, 274](#)
- [Knuth-Morris-Pratt 算法, 274](#)
- [LCS, 308](#)
- [List](#)
  - [split at, 20](#)
- [MTF, 221](#)
- [Patricia, 91](#)
- [reduce, 24](#)
- [T9, 98](#)
- [trie, 88](#)
  - [插入, 89](#)
  - [查找, 90](#)
- [严格弱序, 154](#)
- [中序遍历, 36](#)
- [二分查找, 260](#)
- [二叉堆, 131](#)
  - [Heapify, 132](#)
  - [pop, 134](#)
  - [top-k, 134](#)
  - [弹出, 134](#)
  - [提升优先级, 136](#)
  - [插入\(push\), 136](#)
  - [构造堆, 133](#)
  - [获取顶部元素, 134](#)
- [二叉搜索树, 33](#)
  - [删除, 40](#)
  - [前驱/后继, 39](#)
  - [插入, 34](#)
  - [最小元素/最大元素, 39](#)
  - [查找, 38](#)
  - [随机构建, 43](#)
- [二叉树, 33](#)
  - [遍历, 36](#)
- [二叉随机访问列表](#)
  - [从头部删除, 206](#)
  - [插入, 206](#)
  - [随机访问, 208](#)
- [二项式堆, 165](#)
  - [push, 169](#)
  - [定义, 166](#)
  - [弹出, 171](#)
  - [插入, 169](#)
  - [链接, 168](#)

- 二项式树, 165
  - 合并, 170
- 伸展堆, 141
  - pop, 145
  - splay, 142
  - top, 145
  - 合并, 146
  - 弹出, 145
  - 插入, 145
- 倒水问题, 291
- 八皇后问题, 282
- 列表
  - break, 20
  - cons, 2
  - foldl, 24
  - foldr, 23
  - for each, 16
  - get at, 2
  - init, 3
  - rindex, 4
  - span, 20
  - unzip, 29
  - ZF 表达式, 15
  - zip, 29
  - 丢弃, 19
  - 中缀, 28
  - 串联, 25
  - 修改, 6
  - 分割, 19
  - 分组, 20
  - 切分, 20
  - 列表解析, 15
  - 删除, 8
  - 判空, 1
  - 前缀, 28
  - 匹配, 28
  - 反向索引, 4
  - 反转, 18
  - 变换, 14
  - 右侧叠加, 23
  - 后缀, 28
  - 和, 10
  - 头, 2
  - 存在检查, 26
  - 定义, 1
  - 尾, 2
  - 属于, 26
  - 左侧叠加, 24
  - 截取, 19
  - 提取子列表, 19
  - 插入, 6
  - 映射, 15
  - 更改, 5
  - 最大值, 13
  - 最小值, 13
  - 末尾元素, 3
  - 条件丢弃, 20
  - 条件截取, 20
  - 构造, 2
  - 查找(find), 27
  - 查询(lookup), 26
  - 添加, 5
  - 积, 10
  - 空, 1
  - 索引, 2
  - 过滤, 27
  - 连接, 9
  - 逐一映射, 14
  - 长度, 2
- 前序遍历, 36
- 前缀树, 91
  - 插入, 91
  - 查找, 95
- 动态规划, 306
- 区间遍历, 40
- 华容道, 295

- 双数组列表
  - 删除和平衡, 211
  - 插入和添加, 211
  - 随机访问, 211
- 叠加, 23
- 后序遍历, 36
- 哈夫曼编码, 301
- 基数树, 79
  - 整数 trie, 79
- 堆排序, 137
- 子集和问题, 310
- 完全二叉树, 131
- 尾调用, 10
- 尾递归, 10
- 尾递归调用, 10
- 左侧孩子, 右侧兄弟, 168
- 左偏堆, 138
  - pop, 140
  - S-值, 138
  - top, 140
  - 合并, 139
  - 弹出, 140
  - 秩, 138
- 左偏树
  - 堆排序, 140
  - 插入, 140
- 并行归并排序, 254
- 并行快速排序, 254
- 广度优先搜索, 299
- 序列
  - 二叉随机访问列表, 205
  - 二叉随机访问列表的数字表示, 209
  - 双数组序列, 211
  - 可链接列表, 212
  - 手指树, 215
- 归并排序, 242
  - 分配工作区, 245
  - 原地工作区, 246
  - 原地归并排序, 246
  - 定义, 242
  - 归并, 243
  - 性能分析, 244
  - 自底向上归并排序, 253
  - 自然归并排序, 249
- 快速排序, 227
  - 三分划分, 234
  - 三路划分, 236
  - 划分(partition), 228
  - 双向扫描, 235
  - 平均情况分析, 232
  - 性能分析, 231
  - 改进, 234
- 手指树
  - 头部删除, 217
  - 头部插入, 215
  - 尾部删除, 219
  - 尾部添加, 217
  - 连接, 219
  - 随机访问, 220
- 换零钱问题, 305
- 插入排序, 47
  - 二分查找, 49
  - 二叉搜索树, 51
  - 列表插入排序, 50
  - 插入, 48
- 整数 Patricia, 83
- 整数 trie
  - 插入, 80
  - 查找, 82
- 整数前缀树, 83
  - 插入, 83
  - 查找, 88
- 斐波那契堆, 173
  - 删除最小元素, 175
  - 合并, 174
  - 弹出, 175

- 提升优先级, 180
- 插入, 174
- 斜堆, 141
  - pop, 141
  - top, 141
  - 合并, 141
  - 弹出, 141
  - 插入, 141
- 最大和问题, 272
- 最小可用数, iii
- 最长公共子序列, 308
- 柯里化, 10
- 柯里化形式, 10
- 树旋转, 54
- 深度优先搜索, 279
- 狼、羊、白菜趣题, 288
- 等价关系, 22
- 红黑树, 56
  - 删除, 59
  - 双重黑色, 323
  - 命令式删除, 323
  - 命令式插入, 64
  - 插入, 57
  - 红黑性质, 56
- 统计单词, 33
- 自动补齐, 95
- 贪心算法, 301
- 跳棋趣题, 284
- 迷宫问题, 279
- 选择排序, 151
  - 查找最小元素, 152
  - 递归查找最小元素, 152
- 配对堆, 183
  - pop, 184
  - top, 183
  - 删除, 187
  - 定义, 183
  - 弹出, 184
  - 提升优先级, 184
  - 插入, 183
- 重建树, 37
- 锦标赛淘汰法, 157
- 队列
  - 单向链表实现, 193
  - 双列表队列, 196
  - 双数组队列, 197
  - 实时队列, 198
  - 平衡队列, 197
  - 循环缓冲区, 194
  - 惰性实时队列, 201
- 马鞍搜索, 262
- 鸡尾酒排序, 154

# GNU Free Documentation License

Version 1.3, 3 November 2008

Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc.

[<http://fsf.org/>](http://fsf.org/)

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

## Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document “free” in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of “copyleft”, which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

## 1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The “**Document**”,

below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as “**you**”. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A “**Modified Version**” of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A “**Secondary Section**” is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document’s overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The “**Invariant Sections**” are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The “**Cover Texts**” are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A “**Transparent**” copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not “Transparent” is called “**Opaque**”.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include

PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The “**Title Page**” means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, “Title Page” means the text near the most prominent appearance of the work’s title, preceding the beginning of the body of the text.

The “**publisher**” means any person or entity that distributes copies of the Document to the public.

A section “**Entitled XYZ**” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “**Acknowledgements**”, “**Dedications**”, “**Endorsements**”, or “**History**”.) To “**Preserve the Title**” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

## 2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

## 3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

## 4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.



- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

- M. Delete any section Entitled “Endorsements”. Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled “Endorsements” or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version’s license notice. These titles must be distinct from any other section titles.

You may add a section Entitled “Endorsements”, provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

## 5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original

author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

## 6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

## 7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

## 8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections

with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

## 9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

## 10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <http://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License “or any later version” applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy’s public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

## 11. RELICENSING

“Massive Multiauthor Collaboration Site” (or “MMC Site”) means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A “Massive Multiauthor Collaboration” (or “MMC”) contained in the site means any set of copyrightable works thus published on the MMC site.

“CC-BY-SA” means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

“Incorporate” means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is “eligible for relicensing” if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

## **ADDENDUM: How to use this License for your documents**

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled “GNU Free Documentation License”.

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the “with ... Texts.” line with this:

with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.