第5章 树 结 构

5.1 树

5.1.1 树的基本概念

树是数据元素之间具有层次关系的非线性结构，是由 n 个节点构成的有限集合，节点数为 0 的树叫空树。树必须满足以下条件。

(1) 有且仅有一个被称为根的节点。

(2) 其余节点可分为 m 个互不相交的有限集合，每个集合又构成一棵树，叫根节点的子树。

与线性结构不同，树中的数据元素具有一对多的逻辑关系，除根节点以外，每个数据元素可以有多个后继但有且仅有一个前驱，反映了数据元素之间的层次关系。

树是递归定义的。节点是树的基本单位，若干个节点组成一棵子树，若干棵互不相交的子树组成一棵树。

人们在生活中所见的家谱、Windows 的文件系统等，虽然表现形式各异，但在本质上是树结构。图 5.1 给出了树的逻辑结构示意图。

树的表示方法有多种，如树形表示法、文氏图表示法、凹入图表示法和广义表表示法。图 5.1 所示为树形表示法，图 5.2 给出了用其他 3 种表示法对树的表示。

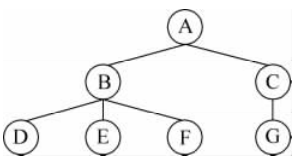

图 5.1 树的逻辑结构示意图

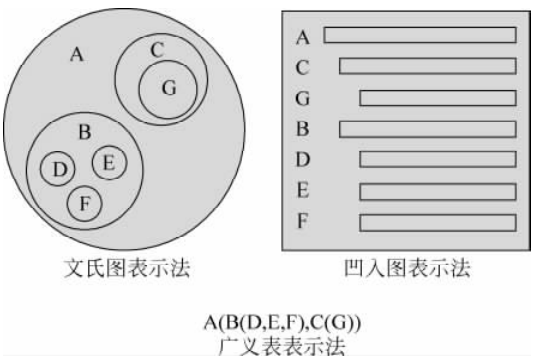

图 5.2 树的 3 种表示方法

5.1.2 树的术语

1. 节点

树的节点就是构成树的数据元素，就是其他数据结构中存储的数据项，在树形表示法中

用圆圈表示。

2. 节点的路径

节点的路径是指从根节点到该节点所经过节点的顺序排列。

3. 路径的长度

路径的长度指的是路径中包含的分支数。

4. 节点的度

节点的度指的是节点拥有的子树的数目。

5. 树的度

树的度指的是树中所有节点的度的最大值。

6. 叶节点

叶节点是树中度为 0 的节点,也叫终端节点。

7. 分支节点

分支节点是树中度不为 0 的节点,也叫非终端节点。

8. 子节点

子节点是指节点的子树的根节点,也叫孩子节点。

9. 父节点

具有子节点的节点叫该子节点的父节点,也叫双亲节点。

10. 子孙节点

子孙节点是指节点的子树中的任意节点。

11. 祖先节点

祖先节点是指节点的路径中除自身之外的所有节点。

12. 兄弟节点

兄弟节点是指和节点具有同一父节点的节点。

13. 节点的层次

树中根节点的层次为 0,其他节点的层次是父节点的层次加 1。

14. 树的深度

树的深度是指树中所有节点的层次数的最大值加 1。

15. 有序树

有序树是指树的各节点的所有子树具有次序关系,不可以改变位置。

16. 无序树

无序树是指树的各节点的所有子树之间无次序关系,可以改变位置。

17. 森林

森林是由多个互不相交的树构成的集合。给森林加上一个根节点就变成一棵树,将树的根节点删除就变成森林。

5.2 二 叉 树

5.2.1 二叉树的基本概念

1. 普通二叉树

二叉树是特殊的有序树,它也是由 n 个节点构成的有限集合。当 $n=0$ 时称为空二叉

树。二叉树的每个节点最多只有两棵子树,子树也为二叉树,互不相交且有左右之分,分别称为左二叉树和右二叉树。

二叉树也是递归定义的,在树中定义的度、层次等术语同样也适用于二叉树。

2. 满二叉树

满二叉树是特殊的二叉树,它要求除叶节点外的其他节点都具有两棵子树,并且所有的叶节点都在同一层上,如图 5.3 所示。

3. 完全二叉树

完全二叉树是特殊的二叉树,若完全二叉树具有 n 个节点,它要求 n 个节点与满二叉树的前 n 个节点具有完全相同的逻辑结构,如图 5.4 所示。

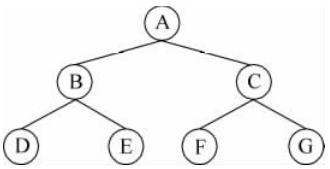

图 5.3　满二叉树

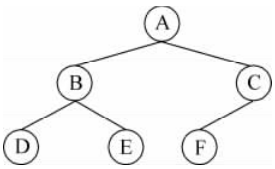

图 5.4　完全二叉树

5.2.2　二叉树的性质

性质 1:二叉树中第 i 层的节点数最多为 $2i$。

证明:当 $i=0$ 时只有一个根节点,成立;假设对所有的 $k(0\leqslant k<i)$ 成立,即第 $i-1$ 层上最多有 2^{i-1} 个节点,那么由于每个节点最多有两棵子树,在第 i 层上节点数最多为 $2^{i-1}\times 2=2^i$ 个,得证。

性质 2:深度为 h 的二叉树最多有 2^h-1 个节点。

证明:由性质 1 得,深度为 h 的二叉树的节点个数最多为 $2^0+2^1+\cdots+2^{h-1}=2^h-1$,得证。

性质 3:若二叉树的叶节点的个数为 n,度为 2 的节点个数为 m,有 $n=m+1$。

证明:设二叉树中度为 1 的节点个数为 k,二叉树的节点总数为 s,有 $s=k+n+m$。又因为除根节点外每个节点都有一个进入它的分支,所以 $s-1=k+2\times m$。整理后得到 $n=m+1$,得证。

性质 4:具有 n 个节点的完全二叉树,其深度为 $\lfloor\log_2 n\rfloor+1$ 或者 $\lceil\log_2(n+1)\rceil$。

证明:设此二叉树的深度为 h,由性质 2 可得 $2^{h-1}\leqslant n<2^h$,两边取对数,可得 $h-1\leqslant\log_2 n<h$,因为 h 为整数,所以 $h=\lfloor\log_2 n\rfloor+1$,得证。

性质 5:具有 n 个节点的完全二叉树,从根节点开始自上而下、从左向右对节点从 0 开始编号。对于任意一个编号为 i 的节点:

(1) 若 $i=0$,节点为根节点,没有父节点;若 $i>0$,则父节点的编号为 $\lfloor(i-1)/2\rfloor$。

(2) 若 $2i+1\geqslant n$,该节点无左孩子,否则左孩子节点的编号为 $2i+1$。

(3) 若 $2i+2\geqslant n$,该节点无右孩子,否则右孩子节点的编号为 $2i+2$。

【例 5.1】 对于任意一个满二叉树,其分支数 $B=2(n_0-1)$,其中 n_0 为终端节点数。

解:

设 n_2 为度为 2 的节点,因为在满二叉树中没有度为 1 的节点,所以有:

$$n = n_0 + n_2$$

设 B 为树中分支数，则：

$$n = B + 1$$

所以：

$$B = n_0 + n_2 - 1$$

再由二叉树的性质：

$$n_0 = n_2 + 1$$

代入上式有：

$$B = n_0 + n_0 - 1 - 1 = 2(n_0 - 1)$$

【例 5.2】 已知一棵度为 m 的树中有 n_1 个度为 1 的节点、n_2 个度为 2 的节点、…、n_m 个度为 m 的节点，问该树中共有多少个叶子节点？

解：设该树的总节点数为 n，则：

$$n = n_0 + n_1 + n_2 + \cdots + n_m$$

又

$$n = \text{分支数} + 1 = 0 \times n_0 + 1 \times n_1 + 2 \times n_2 + \cdots + m \times n_m + 1$$

由上述两式可得：

$$n_0 = n_2 + 2\,n_3 + \cdots + (m-1)\,n_m + 1$$

5.2.3 二叉树的存储结构

1. 二叉树的顺序存储结构

二叉树的顺序存储结构是指将二叉树的各个节点存放在一组地址连续的存储单元中，所有节点按节点序号进行顺序存储。因为二叉树为非线性结构，所以必须先将二叉树的节点排成线性序列再进行存储，实际上是对二叉树先进行一次层次遍历。二叉树的各节点间的逻辑关系由节点在线性序列中的相对位置确定。

可以利用 5.2.2 节中的性质 5 将二叉树的节点排成线性序列，将节点存放在下标为对应编号的数组元素中。为了存储非完全二叉树，需要在树中添加虚节点使其成为完全二叉树后再进行存储，这样会造成存储空间的浪费。

图 5.5 所示为二叉树的顺序存储结构示意图。

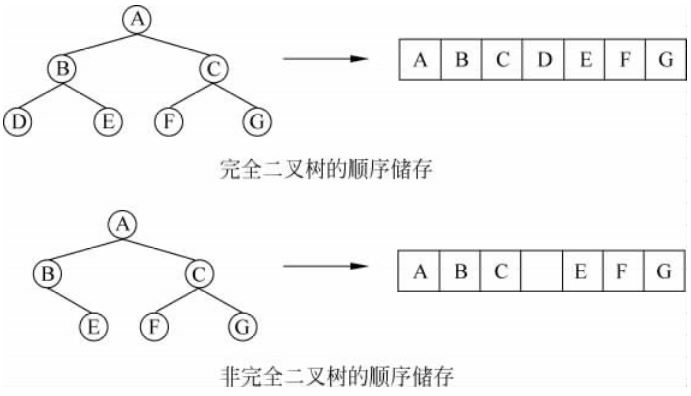

图 5.5 二叉树的顺序存储结构

2. 二叉树的链式存储结构

二叉树的链式存储结构是指将二叉树的各个节点随机存放在存储空间中,二叉树的各节点间的逻辑关系由指针确定。每个节点至少要有两条链分别连接左、右孩子节点才能表达二叉树的层次关系。

根据指针域个数的不同,二叉树的链式存储结构又分为以下两种。

1) 二叉链式存储结构

二叉树的每个节点设置两个指针域和一个数据域。数据域中存放节点的值,指针域中存放左、右孩子节点的存储地址。

采用二叉链表存储二叉树,每个节点只存储了到其孩子节点的单向关系,没有存储到其父节点的关系,因此要获得父节点将花费较多的时间,需要从根节点开始在二叉树中进行查找,所花费的时间是遍历部分二叉树的时间,且与查找节点所处的位置有关。

2) 三叉链式存储结构

二叉树的每个节点设置 3 个指针域和一个数据域。数据域中存放节点的值,指针域中存放左、右孩子节点和父节点的存储地址。

图 5.6 所示为二叉链式存储和三叉链式存储的节点结构。

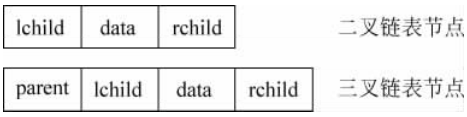

图 5.6 二叉和三叉链式存储的节点结构

两种链式存储结构各有优缺点,二叉链式存储结构空间利用率高,而三叉链式存储结构既便于查找孩子节点,又便于查找父节点。在实际应用中,二叉链式存储结构更加常用,因此本书中二叉树的相关算法都是基于二叉链式存储结构设计的。

3. 二叉链式存储结构的节点类的描述

```
class BiTreeNode(metaclass = ABCMeta):
    def __init__(self,data = None,lchild = None,rchild = None):
        self.data = data          # 数据域的值
        self.lchild = lchild      # 左孩子的指针
        self.rchild = rchild      # 右孩子的指针
```

4. 二叉树类的描述

此二叉树类基于二叉链式存储结构实现。

```
class BiTree(object):
    def __init__(self,root = None):
        self.root = root # 二叉树的根节点
```

二叉树的创建操作和遍历操作比较重要,将在下面的章节中进行详细介绍。

5.2.4 二叉树的遍历

二叉树的遍历是指沿着某条搜索路径访问二叉树的节点,每个节点被访问的次数有且仅有一次。

1. 二叉树的遍历方法

二叉树通常可划分为 3 个部分,即根节点、左子树和右子树。根据 3 个部分的访问顺序

不同，可将二叉树的遍历方法分为以下几种。

1）层次遍历

自上而下、从左到右依次访问每层的节点。

2）先序遍历

先访问根节点，再先序遍历左子树，最后先序遍历右子树。

3）中序遍历

先中序遍历左子树，再访问根节点，最后中序遍历右子树。

4）后序遍历

先后序遍历左子树，再后序遍历右子树，最后访问根节点。

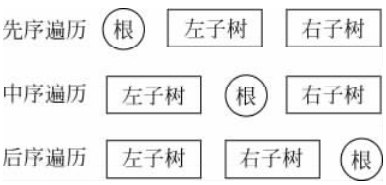

图 5.7　二叉树遍历序列的节点排列规律

图 5.7 描述了先序遍历、中序遍历和后序遍历序列的节点排列规律。

2. 二叉树遍历操作实现的递归算法

【算法 5.1】 先序遍历。

```
def preOrder(root):
    if root is not None:
        print(root.data,end = ' ')
        BiTree.preOrder(root.lchild)
        BiTree.preOrder(root.rchild)
```

【算法 5.2】 中序遍历。

```
def inOrder(root):
    if root is not None:
        BiTree.inOrder(root.lchild)
        print(root.data,end = ' ')
        BiTree.inOrder(root.rchild)
```

【算法 5.3】 后序遍历。

```
def postOrder(root):
    if root is not None:
        BiTree.postOrder(root.lchild)
        BiTree.postOrder(root.rchild)
        print(root.data,end = ' ')
```

3. 二叉树遍历操作实现的非递归算法

二叉树遍历操作的递归算法结构简洁，易于实现，但是在时间上开销较大，运行效率较低，为了解决这一问题，可以将递归算法转换为非递归算法，转换方式有以下两种。

（1）使用临时遍历保存中间结果，用循环结构代替递归过程。

（2）利用栈保存中间结果。

二叉树遍历操作实现的非递归算法利用栈结构通过回溯访问二叉树的每个节点。

1）先序遍历

先序遍历从二叉树的根节点出发，沿着该节点的左子树向下搜索，每遇到一个节点先访

问该节点,并将该节点的右子树入栈。先序遍历左子树完成后再从栈顶弹出右子树的根节点,然后采用相同的方法先序遍历右子树,直到二叉树的所有节点都被访问。其主要步骤如下。

(1) 将二叉树的根节点入栈。

(2) 若栈非空,将节点从栈中弹出并访问。

(3) 依次访问当前访问节点的左孩子节点,并将当前节点的右孩子节点入栈。

(4) 重复步骤(2)和(3),直到栈为空。

【算法 5.4】 先序遍历。

```
def preOrder2(root):
    p = root
    s = LinkStack()
    s.push(p)
    while not s.isEmpty():
        p = s.pop()
        print(p.data,end = ' ')
        while p is not None:
            if p.lchild is not None:
                print(p.lchild.data,end = ' ')
            if p.rchild is not None:
                s.push(p.rchild)
            p = p.lchild
```

2) 中序遍历

中序遍历从二叉树的根节点出发,沿着该节点的左子树向下搜索,每遇到一个节点就使其入栈,直到节点的左孩子节点为空。再从栈顶弹出节点并访问,然后采用相同的方法中序遍历节点的右子树,直到二叉树的所有节点都被访问。其主要步骤如下。

(1) 将二叉树的根节点入栈。

(2) 若栈非空,将栈顶节点的左孩子节点依次入栈,直到栈顶节点的左孩子节点为空。

(3) 将栈顶节点弹出并访问,并使栈顶节点的右孩子节点入栈。

(4) 重复步骤(2)和(3),直到栈为空。

【算法 5.5】 中序遍历。

```
def inOrder2(root):
    p = root
    s = LinkStack()
    s.push(p)
    while not s.isEmpty():
        while p.lchild is not None:
            p = p.lchild
            s.push(p)
        p = s.pop()
        print(p.data,end = ' ')
        if p.rchild is not None:
            s.push(p.rchild)
```

3）后序遍历

后序遍历从二叉树的根节点出发，沿着该节点的左子树向下搜索，每遇到一个节点需要判断其是否为第一次经过，若是则使节点入栈，后序遍历该节点的左子树，完成后再遍历该节点的右子树，最后从栈顶弹出该节点并访问。后序遍历算法的实现需要引入两个变量，一个为访问标记变量 flag，用于标记栈顶节点是否被访问，若 flag＝true，证明该节点已被访问，其左子树和右子树已经遍历完毕，可继续弹出栈顶节点，否则需要先遍历栈顶节点的右子树；一个为节点指针 t，指向最后一个被访问的节点，查看栈顶节点的右孩子节点，证明此节点的右子树已经遍历完毕，栈顶节点可出栈并访问。其主要步骤如下。

（1）将二叉树的根节点入栈，t 赋值为空。

（2）若栈非空，将栈顶节点的左孩子节点依次入栈，直到栈顶节点的左孩子节点为空。

（3）若栈非空，查看栈顶节点的右孩子节点，若右孩子节点为空或者与 p 相等，则弹出栈顶节点并访问，同时使 t 指向该节点，并置 flag 为 true；否则将栈顶节点的右孩子节点入栈，并置 flag 为 false。

（4）若 flag 为 true，重复步骤(3)；否则重复步骤(2)和(3)，直到栈为空。

【算法 5.6】 后序遍历。

```
def postOrder2(root):
    p = root
    t = None
    flag = True
    s = LinkStack()
    if p is not None:
        s.push(p)
        while p.child is None:
            p = p.lchild
            s.push(p)
        while not s.isEmpty() and flag:
            if p.rchild == t or p.rchild is None:
                print(p.data,end = ' ')
                flag = True
                t = p
                s.pop()
            else:
                s.push(p.rchild)
                flag = False
```

4）层次遍历

层次遍历操作是从根节点出发，自上而下、从左到右依次遍历每层的节点，可以利用队列先进先出的特性进行实现。先将根节点入队，然后将队首节点出队并访问，都将其孩子节点依次入队。其主要步骤如下。

（1）将根节点入队。

（2）若队非空，取出队首节点并访问，将队首节点的孩子节点入队。

（3）重复执行步骤(2)直到队为空。

【算法 5.7】 层次遍历。

```
def order(root):
    q = LinkQueue()
    q.offer(root)
    while not q.isEmpty():
        p = q.poll()
        print(p.data,end = ' ')
        if p.lchild is not None:
            q.offer(p.lchild)
        if p.rchild is not None:
            q.offer(p.rchild)
```

对于有 n 个节点的二叉树,因为每个节点都只访问一次,所以以上 4 种遍历算法的时间复杂度均为 $O(n)$。

4 种遍历算法的实现均利用了栈或队列,增加了额外的存储空间,存储空间的大小为遍历过程中栈或队列需要的最大容量。对于栈来说,其最大容量即为树的高度,在最坏情况下有 n 个节点的二叉树的高度为 n,所以其空间复杂度为 $O(n)$; 对于队列来说,其最大容量为二叉树相邻两层的最大节点总数,与 n 成线性关系,所以其空间复杂度也为 $O(n)$。

5.2.5 二叉树遍历算法的应用

二叉树的遍历操作是实现对二叉树其他操作的一个重要基础,本节介绍了二叉树遍历算法在许多应用问题中的运用。

1. 二叉树上的查找算法

二叉树上的查找是在二叉树中查找值为 x 的节点,若找到返回该节点,否则返回空值,可以在二叉树的先序遍历过程中进行查找,主要步骤如下。

(1) 若二叉树为空,则不存在值为 x 的节点,返回空值; 否则将根节点的值与 x 进行比较,若相等,返回该节点。

(2) 若根节点的值与 x 的值不等,则在左子树中进行查找,若找到,则返回该节点。

(3) 若没有找到,则在根节点的右子树中进行查找,若找到,返回该节点,否则返回空值。

【算法 5.8】 二叉树查找算法。

```
def searchNode(t,x):
    if t is None:
        return None
    if t.data == x:
        return t
    else:
        lresult = searchNode(t.lchild,x)
    if lresult == None:
        return searchNode(t.rchild,x)
    else:
        return lresult
```

2. 统计二叉树的节点个数的算法

二叉树的节点个数等于根节点加上左、右子树的节点的个数，可以利用二叉树的先序遍历序列，引入一个计数变量 count，count 的初值为 0，每访问根节点一次就将 count 的值加 1，其主要操作步骤如下。

（1）count 值初始化为 0。

（2）若二叉树为空，返回 count 值。

（3）若二叉树非空，则 count 值加 1，统计根节点的左子树的节点个数，并将其加到 count 中；统计根节点的右子树的节点个数，并将其加到 count 中。

【算法 5.9】 统计二叉树的节点个数。

```
def nodeCount(t):
    count = 0
    if t is not None:
        count += 1
        count += nodeCount(t.lchild)
        count += nodeCount(t.rchild)
    return count
```

3. 求二叉树的深度

二叉树的深度是所有节点的层次数的最大值加 1，也就是左子树和右子树的深度的最大值加 1，可以采用后序遍历的递归算法解决此问题，其主要步骤如下。

（1）若二叉树为空，返回 0。

（2）若二叉树非空，求左子树的深度，求右子树的深度。

（3）比较左、右子树的深度，取最大值加 1 即为二叉树的深度。

【算法 5.10】 求二叉树的深度。

```
def getDepth(t):
    if t is None:
        return 0
    ldepth = getDepth(t.lchild)
    rdepth = getDepth(t.rchild)
    if ldepth < rdepth:
        return rdepth + 1
    else:
        return ldepth + 1
```

5.2.6 二叉树的建立

二叉树遍历操作可使非线性结构的树转换成线性序列。先序遍历序列和后序遍历序列反映父节点和孩子节点间的层次关系，中序遍历序列反映兄弟节点间的左右次序关系。因为二叉树是具有层次关系的节点构成的非线性结构，并且每个节点的孩子节点具有左右次序，所以已知一种遍历序列无法唯一确定一棵二叉树，只有同时知道中序和先序遍历序列，或者同时知道中序和后序遍历序列，才能同时确定节点的层次关系和节点的左右次序，才能唯一确定一棵二叉树。

1. 由中序和先序遍历序列建立二叉树

其主要步骤为如下。

(1) 取先序遍历序列的第一个节点作为根节点,序列的节点个数为 n。

(2) 在中序遍历序列中寻找根节点,其位置为 i,可确定在中序遍历序列中根节点之前的 i 个节点构成的序列为根节点的左子树中序遍历序列,根节点之后的 $n-i-1$ 个节点构成的序列为根节点的右子树中序遍历序列。

(3) 在先序遍历序列中根节点之后的 i 个节点构成的序列为根节点的左子树先序遍历序列,先序遍历序列之后的 $n-i-1$ 个节点构成的序列为根节点的右子树先序遍历序列。

(4) 对左、右子树重复步骤(1)、(2)、(3),确定左、右子树的根节点和子树的左右、子树。

(5) 算法递归进行即可建立一棵二叉树。

假设二叉树的先序遍历序列为 ABECFG,中序遍历序列为 BEAFCG,由中序和先序遍历序列建立二叉树的过程如图 5.8 所示。

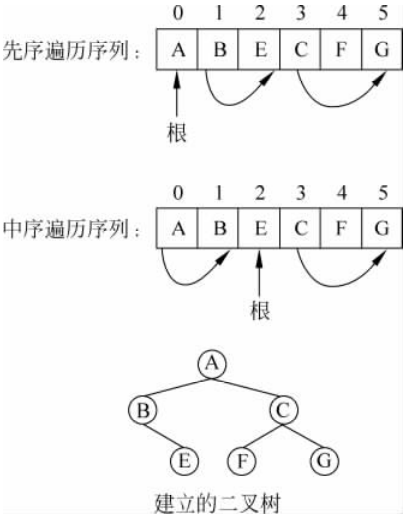

图 5.8 由中序和先序遍历序列建立二叉树

【算法 5.11】 由中序和先序遍历序列建立二叉树。

```
def createBiTree(preOrder,inOrder,preo,ino,n):
    if n>0:
        i = 0
        c = preOrder.charAt(preo) # c 为先序序列的根节点
        while i<n:
            if inOrder.charAt(i+ino)==c:
                break
            i += 1
        root = BiTreeNode(c)
        root.lchild = createBiTree(preOrder,inOrder,preo+1,ino,i).root
        # 递归寻找左子树的根节点
        root.rchild = createBiTree(preOrder,inOrder,preo+i+1,ino+i+1,n-i-1).root
        # 递归寻找右子树的根节点
```

2. 由标明空子树的先序遍历序列创建二叉树

其主要步骤如下。

(1) 从先序遍历序列中依次读取字符。

(2) 若字符为#,建立空子树。

(3) 建立左子树。

(4) 建立右子树。

【算法 5.12】 由标明空子树的先序遍历序列建立二叉树。

```
def createBiTree(preOrder,i): # i 为常数 0
    c = preOrder.charAt(i) # 取字符
    if c != '#':
        root = BiTreeNode(c)
        i += 1
        root.lchild = createBiTree(preOrder,i).root
        i += 1
        root.rchild = createBiTree(preOrder,i).root
    else:
        root = None
```

【例 5.3】 已知二叉树的中序和后序序列分别为 CBEDAFIGH 和 CEDBIFHGA,试构造该二叉树。

解:二叉树的构造过程如图 5.9 所示。图(c)即为构造出的二叉树。

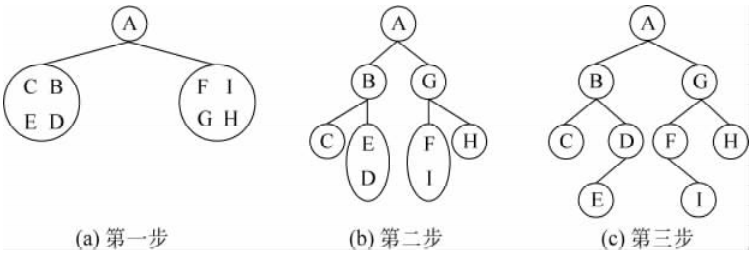

图 5.9 二叉树的构造过程

5.3 哈夫曼树及哈夫曼编码

目前常用的图像、音频、视频等多媒体信息数据量大,必须对它们采用数据压缩技术来存储和传输。数据压缩技术通过对数据进行重新编码来压缩存储,以便减少数据占用的存储空间,在使用时再进行解压缩,恢复数据的原有特性。

其压缩方法主要有有损压缩和无损压缩两种。有损压缩是指压缩过程中可能会丢失数据信息,如将 BMP 位图压缩成 JPEG 格式的图像,会有精度损失;无损压缩是指压缩存储数据的全部信息,确保解压后的数据不丢失。哈夫曼编码是数据压缩技术中的无损压缩技术。

5.3.1 哈夫曼树的基本概念

1. 节点间的路径

节点间的路径是指从一个节点到另一个节点所经过的节点序列。从根节点到 X 节点有且仅有一条路径。

2. 节点的路径长度

节点的路径长度是指从根节点到节点的路径上的边数。

3. 节点的权

节点的权是指人给节点赋予的一个具有某种实际意义的数值。

4. 节点的带权路径长度

节点的带权路径长度是指节点的权值和节点的路径长度的乘积。

5. 树的带权路径长度

树的带权路径长度是指树的叶节点的带权路径长度之和。

6. 最优二叉树

最优二叉树是指给定 n 个带有权值的节点作为叶节点构造出的具有最小带权路径长度的二叉树。最优二叉树也叫哈夫曼树。

5.3.2 哈夫曼树的构造

给定 n 个叶节点,它们的权值分别是$\{w_1,w_2,\cdots,w_n\}$,构造相应的哈夫曼树的主要步骤如下。

(1) 构造由 n 棵二叉树组成的森林,每棵二叉树只有一个根节点,根节点的权值分别为$\{w_1,w_2,\cdots,w_n\}$。

(2) 在森林中选取根节点权值最小和次小的两棵二叉树分别作为左子树和右子树去构造一棵新的二叉树,新二叉树的根节点权值为两棵子树的根节点权值之和。

(3) 将两棵二叉树从森林中删除,并将新的二叉树添加到森林中。

(4) 重复步骤(2)和(3),直到森林中只有一棵二叉树,此二叉树即为哈夫曼树。

假设给定的权值为{1,2,3,4,5},图 5.10 展示了哈夫曼树的构造过程。

【例 5.4】 对于给定的一组权值 $W=\{5,2,9,11,8,3,7\}$,试构造相应的哈夫曼树,并计算它的带权路径长度。

解:构造的哈夫曼树如图 5.11 所示。

树的带权路径长度如下:

$$\text{WPL}=2\times4+3\times4+5\times3+7\times3+8\times3+9\times2+11\times2=120$$

5.3.3 哈夫曼编码

在传送信息时需要将信息符号转化成二进制组成的符号串,一般每个字符由一个字节或两个字节表示,即 8 或 16 个位数。为了提高存储和传输效率,需要设计对字符集进行二进制编码的规则,使得利用这种规则对信息进行编码时编码位数最小,即需要传输的信息量最小。

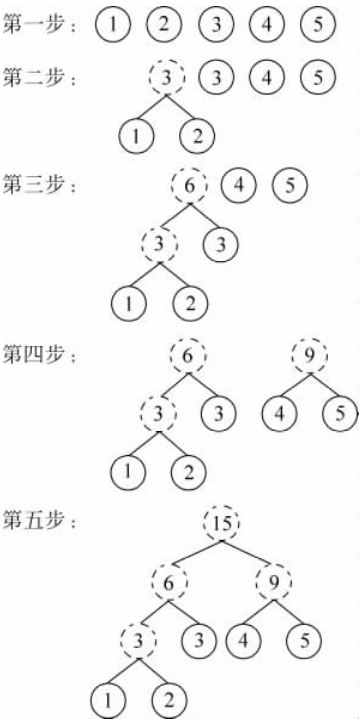

图 5.10　哈夫曼树的构造过程

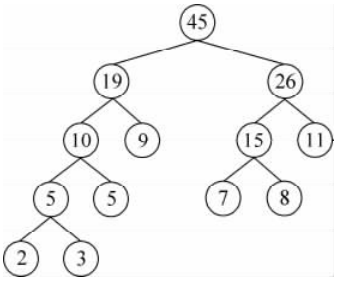

图 5.11　哈夫曼树

哈夫曼编码是一种不等长的编码方案，数据的编码因其使用频率的不同而长短不一，使用频率高的数据其编码较短，使用频率低的数据其编码较长，从而使所有数据的编码总长度最短。各数据的使用频率通过在全部数据中统计重复数据的出现次数获得。

又因为在编码序列中若使用前缀相同的编码来表示不同的字符会造成二义性，额外的分隔符号会造成传输信息量的增加，为了省去不必要的分隔符号，要求每一个字符的编码都不是另一个字符的前缀，即每个字符的编码都是前缀编码。

利用哈夫曼树构造出的哈夫曼编码是一种最优前缀编码，构造的主要步骤如下。

（1）对于具有 n 个字符的字符集，将字符的频度作为叶节点的权值，产生 n 个带权叶节点。

（2）根据 5.3.2 节中介绍的构造哈夫曼树的方法利用 n 个叶节点构造哈夫曼树。

（3）根据哈夫曼编码规则将哈夫曼树中的每一条左分支标记为 0，每一条右分支标记为 1，则可得到每个叶节点的哈夫曼编码。

哈夫曼编码的译码过程是构造过程的逆过程，从哈夫曼树的根节点开始对编码的每一位进行判别，如果为 0 进入左子树，如果为 1 进入右子树，直到到达叶节点，即译出了一个字符。

5.3.4　构造哈夫曼树和哈夫曼编码的类的描述

构造哈夫曼树需要从子节点到父节点的操作，译码时需要从父节点到子节点的操作，所以为了提高算法的效率将哈夫曼树的节点设计为三叉链式存储结构。一个数据域存储节点

的权值,一个标记域 flag 标记节点是否已经加入到哈夫曼树中,3 个指针域分别存储着指向父节点、孩子节点的地址。

节点类的描述如下:


```
class HuffmanNode(object):
    def __init__(self,data,weight):
        self.data = data          # 节点的值
        self.weight = weight      # 节点的权值
        self.parent = None        # 父节点
        self.lchild = None        # 左孩子
        self.rchild = None        # 右孩子
```

【算法 5.13】 构造哈夫曼树。

```
class HuffmanTree(object):
    def __init__(self,data):
        # data 是编码字符与出现次数的集合,例如 w = [('a',1),('b',2)]
        nodes = [ HuffmanNode(c,w) for c,w in data ]
        self.index = {} # 编码字符的索引
        while len(nodes)> 1:
            nodes = sorted(nodes,key = lambda x:x.weight)
            s = HuffmanNode(None,nodes[0].weight + nodes[1].weight)
            s.lchild = nodes[0]
            s.rchild = nodes[1]
            nodes[0].parent = nodes[1].parent = s
            nodes = nodes[2:]
            nodes.append(s)
        self.root = nodes[0]
        self.calIndex(self.root,'') # 递归计算每个字符的哈夫曼编码并保存

    def calIndex(self,root,str):
        if root.data is not None:
            # 保存字符的编码
            self.index[root.data] = str
        else:
            self.calIndex(root.lchild,str + '0')
            self.calIndex(root.rchild,str + '1')

    def queryHuffmanCode(self,c):
        if c not in self.index:
            raise Exception("未编码的字符")
        return self.index[c]
```

【算法 5.14】 若字符与出现频率对应关系如:[('a',5),('b',2),('c',9),('d',11),('e',8),('f',3),('g',7)]求哈夫曼编码。

```
data = [('a',5),('b',2),('c',9),('d',11),('e',8),('f',3),('g',7)]

t = HuffmanTree(data)

```

```
5   for c,w in data:
6       print('字符 %s 的哈夫曼编码为: %s'%(c,t.queryHuffmanCode(c)))
```

输出如下：

```
字符 a 的哈夫曼编码为: 010
字符 b 的哈夫曼编码为: 0110
字符 c 的哈夫曼编码为: 00
字符 d 的哈夫曼编码为: 10
字符 e 的哈夫曼编码为: 111
字符 f 的哈夫曼编码为: 0111
字符 g 的哈夫曼编码为: 110
```

【例 5.5】 已知某字符串 s 中共有 8 种字符，各种字符分别出现 2 次、1 次、4 次、5 次、7 次、3 次、4 次和 9 次，对该字符串用[0,1]进行前缀编码，问该字符串的编码至少有多少位？

解：以各字符出现的次数作为叶子节点的权值构造的哈夫曼编码树如图 5.12 所示。其带权路径长度＝2×5＋1×5＋3×4＋5×3＋9×2＋4×3＋4×3＋7×2＝98，所以该字符串的编码长度至少为 98 位。

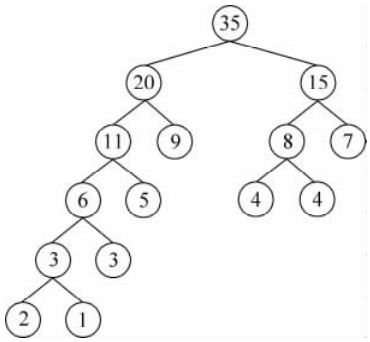

图 5.12　哈夫曼编码树

5.4　树 和 森 林

5.4.1　树的存储结构

一棵树包含各节点间的层次关系和兄弟关系，两种关系的存储结构不同。

树的层次关系必须采用链式存储结构存储，通过链连接父节点和孩子节点。

一个节点的多个孩子节点（互称兄弟节点）之间是线性关系，可以采用顺序存储结构或者链式存储结构。

1. 树的父母孩子链表

树的父母孩子链表采用顺序存储结构存储多个孩子节点，其中 children 数组存储多个孩子节点，各节点的 children 数组元素长度不同，为孩子个数。

2. 树的父母孩子兄弟链表

树的父母孩子兄弟链表采用链式存储结构存储多个孩子节点，节点的 child 链指向一个

孩子节点,sibling 链指向下一个兄弟节点。

森林也可以使用父母孩子兄弟链表进行存储,这种存储结构实际上是把一棵树转换成一棵二叉树存储。其存储规则如下。

(1) 每个节点采用 child 链指向其中一个孩子节点,多个孩子节点之间由 sibling 链连接起来,组成一条具有兄弟节点关系的单链表。

(2) 将每棵树采用树的父母孩子兄弟链表存储。

(3) 森林中的多棵树之间是兄弟关系,将这些树通过根的 sibling 链连接起来。

5.4.2 树的遍历规则

树的孩子优先遍历规则主要有两种,即先序遍历和后序遍历。树的遍历规则也是递归的。

(1) 树的先序遍历:访问根节点;按从左到右的次序遍历根的每一棵子树。

(2) 树的后序遍历:按从左到右的次序遍历根的每一棵子树;访问根节点。

树的层次遍历规则同二叉树。

小　　结

(1) 树是数据元素之间具有层次关系的非线性结构,是由 n 个节点构成的有限集合。与线性结构不同,树中的数据元素具有一对多的逻辑关系。

(2) 二叉树是特殊的有序树,它也是由 n 个节点构成的有限集合。当 $n=0$ 时称为空二叉树。二叉树的每个节点最多只有两棵子树,子树也为二叉树,互不相交且有左、右之分,分别称为左二叉树和右二叉树。

(3) 二叉树的存储结构分为两种,即顺序存储结构和链式存储结构。二叉树的顺序存储结构是指将二叉树的各个节点存放在一组地址连续的存储单元中,所有节点按节点序号进行顺序存储;二叉树的链式存储结构是指将二叉树的各个节点随机存放在存储空间中,二叉树的各节点间的逻辑关系由指针确定。

(4) 二叉树具有先序遍历、中序遍历、后序遍历和层次遍历 4 种遍历方式。

(5) 最优二叉树是指给定 n 个带有权值的节点作为叶节点构造出的具有最小带权路径长度的二叉树,也叫哈夫曼树。

(6) 哈夫曼编码是数据压缩技术中的无损压缩技术,是一种不等长的编码方案,使所有数据的编码总长度最短。

习　题　5

一、选择题

1. 如果节点 A 有 3 个兄弟,B 是 A 的双亲,则节点 B 的度是(　　)。

A. 1　　B. 2　　C. 3　　D. 4

2. 设二叉树有 n 个节点，则其深度为(　　)。

A. $n-1$　　B. n　　C. $\lfloor \log_2 n \rfloor + 1$　　D. 不能确定

3. 二叉树的前序序列和后序序列正好相反，则该二叉树一定是(　　)的二叉树。

A. 空或只有一个节点　　B. 高度等于其节点数

C. 任一节点无左孩子　　D. 任一节点无右孩子

4. 线索二叉树中某节点 R 没有左孩子的充要条件是(　　)。

A. R. lchild＝None　　B. R. ltag＝0

C. R. ltag＝1　　D. R. rchild＝None

5. 深度为 k 的完全二叉树最少有(　　)个节点、最多有(　　)个节点，具有 n 个节点的完全二叉树按层序从 1 开始编号，则编号最小的叶子节点的序号是(　　)。

A. $2^{k-2}+1$　　B. 2^k-1　　C. 2^{k-1}　　D. $2^{k-1}-1$

E. 2^{k+1}　　F. $2^{k+1}-1$　　G. $2^{k-1}+1$　　H. 2^k

6. 一个高度为 h 的满二叉树共有 n 个节点，其中有 m 个叶子节点，则(　　)成立。

A. $n=h+m$　　B. $h+m=2n$　　C. $m=h-1$　　D. $n=2m-1$

7. 任何一棵二叉树的叶子节点在前序、中序、后序遍历序列中的相对次序(　　)。

A. 肯定不发生改变　　B. 肯定发生改变

C. 不能确定　　D. 有时发生变化

8. 如果 T' 是由有序树 T 转换而来的二叉树，那么 T 中节点的前序序列就是 T' 中节点的(　　)序列，T 中节点的后序序列就是 T' 中节点的(　　)序列。

A. 前序　　B. 中序　　C. 后序　　D. 层序

9. 设森林中有 4 棵树，树中节点的个数依次为 n_1、n_2、n_3、n_4，则把森林转换成二叉树后其根节点的右子树上有(　　)个节点、根节点的左子树上有(　　)个节点。

A. n_1-1　　B. n_1　　C. $n_1+n_2+n_3$　　D. $n_2+n_3+n_4$

10. 讨论树、森林和二叉树的关系目的是为了(　　)。

A. 借助二叉树上的运算方法去实现对树的一些运算

B. 将树、森林按二叉树的存储方式进行存储并利用二叉树的算法解决树的有关问题

C. 将树、森林转换成二叉树

D. 体现一种技巧，没有什么实际意义

二、填空题

1. 树是 $n(n\geqslant 0)$ 个节点的有限集合，在一棵非空树中有_____个根节点，其余节点分成 $m(m>0)$ 个_____的集合，每个集合都是根节点的子树。

2. 树中某节点的子树的个数称为该节点的_____，子树的根节点称为该节点的_____，该节点称为其子树根节点的_____。

3. 一棵二叉树的第 $i(i\geqslant 1)$ 层最多有_____个节点；一棵有 $n(n>0)$ 个节点的满二叉树共有_____个叶子节点和_____个非终端节点。

4. 设高度为 h 的二叉树上只有度为 0 和度为 2 的节点，该二叉树的节点数可能达到的最大值是_____、最小值是_____。

5. 在深度为 k 的二叉树中所含叶子的个数最多为_____。

6. 具有 100 个节点的完全二叉树的叶子节点数为_____。

7. 已知一棵度为 3 的树有两个度为 1 的节点、3 个度为 2 的节点、4 个度为 3 的节点，则该树中有_____个叶子节点。

8. 某二叉树的前序遍历序列是 ABCDEFG，中序遍历序列是 CBDAFGE，则其后序遍历序列是_____。

9. 在具有 n 个节点的二叉链表中共有_____个指针域，其中_____个指针域用于指向其左、右孩子，剩下的_____个指针域则是空的。

10. 在有 n 个叶子的哈夫曼树中叶子节点总数为_____，分支节点总数为_____。

三、算法设计题

1. 设计算法求二叉树的节点个数；按前序次序打印二叉树中的叶子节点；求二叉树的深度。

2. 设计算法判断一棵二叉树是否为完全二叉树。

3. 使用栈将 Tree 类中的递归算法实现为非递归算法。

4. 编写一个非递归算法求出二叉搜索树中的所有节点的最大值，若树为空则返回空值。

5. 编写一个算法求出一棵二叉树中叶子节点的总数，参数初始指向二叉树的根节点。