**实验三 二叉树的应用**

PB17050941 李喆昊

1. 问题描述

题目：用huffman压缩技术实现对任意文件的压缩和解压缩处理

要求对所有的文件类型进行压缩，压缩之后的文件后缀名为huff。同时，可以对所有后缀名为huff的压缩文件进行解压缩。

题目具体要求：

1.      以1个字节(8bit)为单位进行huffman编码

2.      对任意文件进行压缩后可以输出一个后缀名为huff的单文件，并且可以对任意一个后缀名为huff的单文件进行解压还原出原文件。

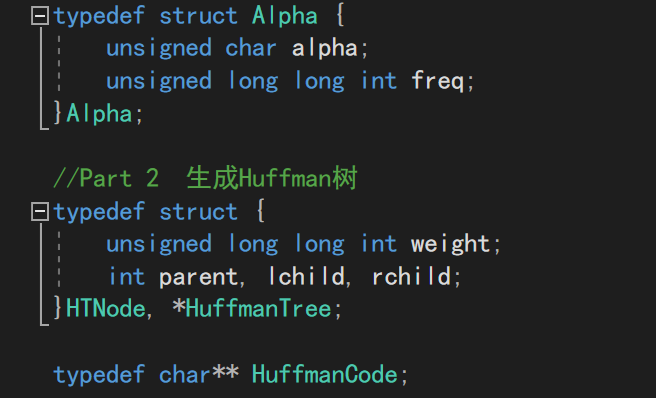
3.      群内将提供10不同种类的文件包括文档、图片、视频、可执行文件等进行压缩测试，要求可以完成压缩和解压的步骤，并且解压出来的文件没有任何损失。

1. 算法描述
2. 数据结构描述

采用二叉树的数组存储结构作为哈夫曼树的存储结构，逻辑结构为树形结构

采用数组作为哈夫曼编码的存储结构，使用一个指针数组存储指向各个字符Huffman编码的指针。

主要变量：

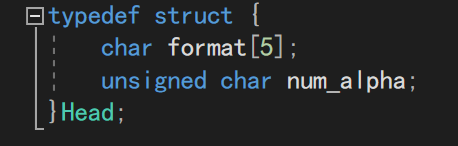


Alpha 结构体用于存储待压缩文件中的字符及该字符出现的频率

HTNode 是Huffman树的结点

HuffmanTree是指向Huffman树结点的指针

HuffmanCode是指向“指向哈夫曼编码的指针数组”的指针



Head结构体存储着写入Huff类型压缩文件头部的解压缩时候需要的信息。



压缩文件中的字符在计算机中都是用16位二进制编码表示的，所以最多有2^16=256种字符。freq就是存储每个字符的频率的数组。



alphabet的每个元素是Alpha结构体，由于存放压缩文件中字符的信息。

1. 程序结构描述



功能：这个函数用于在构建Huffman树时，从Huffman树结点数组HT的1-- num个字符中选取出现频次最小的两个字符，用s1和s2返回它们在HT数组中的序号。



功能：生成Huffman树并生成Huffman编码。HT是指向生成的Huffman树的根节点的指针，HC是指向“指向生成的哈夫曼编码的指针数组”的指针。Alphabet是待压缩文件中存储字符信息的数组，num\_alpha是待压缩文件中字符的数目。

三个重要的位操作函数：



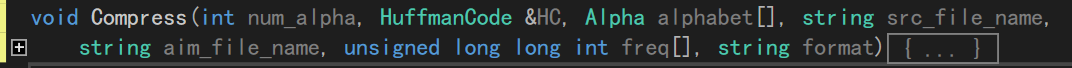
功能：将value的第index位变成1（二进制位），用于压缩时写入哈夫曼编码。



功能：将value的第index位变成0（二进制位），用于压缩时写入哈夫曼编码。



功能：获取value的第index位置（二进制位）的值，用于解压缩。



功能：对源文件进行压缩。



功能：对压缩后的.huff类型文件进行解压缩。

三．调试分析

（1）用于压缩与解压缩测试的文件是助教提供的多个不同类型的文件。

（2）程序调试中遇到的问题：

① 创建Huffman树时错误。

问题1：仔细检查后发现select函数有问题：在选取的时候不能够选出最小频次的。

问题2：在代码中有一个很容易错的地方是：我的存放Huffman结点的数组HT的0号位是没有使用的，因此第n个Huffman结点就在HT [n]位，而我一开始误以为是HT[n-1]位。

②压缩文件时：压缩后文件的尾部内容总是不对。导致调试时gif、mp4文件可以压缩与解压缩，但是docx、xls文件压缩之后无法成功解压缩。

解决方法：仔细思考压缩到尾部的情况，使用fwrite函数将最后没有满8位的字符也写到压缩文件中。

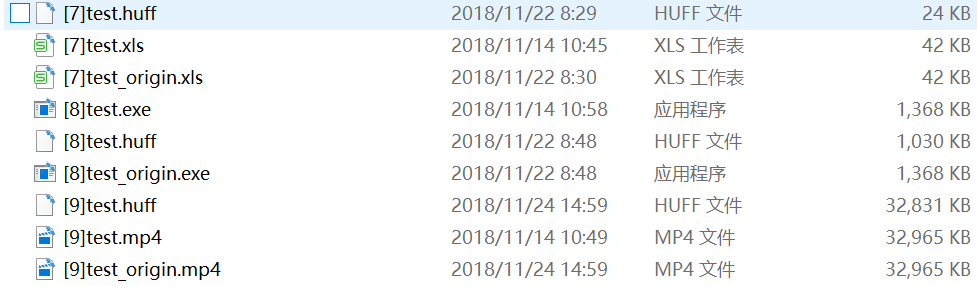
四．算法时空分析

1. 压缩文件：需要遍历所有字符进行频次统计，并生成Huffman树、Huffman编码，时间复杂度为O(n)。
2. 解压缩：需要遍历所有字符并在Huffman树中解码，时间复杂度为O(n)。

五．测试结果及分析

以压缩.bmp文件为例：压缩后的Huff文件相比于源文件大小明显降低：5626KB-->909KB。

但是以压缩mp4文件为例：.mp4文件压缩后大小反而有所增加。这种现象是由Huffman压缩算法本身的特点决定的，并不是程序错误。



总结：在各测试文件中，.mp4、.gif文件在压缩之后大小有可能会增加，.xls、bmp格式压缩后大小有明显的减少，其余格式压缩后大小都有一定的减少。

1. 实验体会和收获

本次实验很有意思，通过对不同格式的文件进行压缩，让我们对“在计算机内部各种文件都是采用二进制格式存储的”这句话有了更深刻的理解，这也是我们能够对各种格式的文件采用统一的压缩方式的根本原因。

同时本次实验也具有一定的挑战性，过程中我遇到了很多困难，例如文件尾部压缩失败的问题困扰了我很久。在摸索中，编程思路逐渐由最初的不知所措变得清晰起来，锻炼了我处理问题的能力与调试bug的耐心。